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

December 9, 2020

1 Introduction

(Frame the problem) 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 European Values Study (EVS), Longitudinal Internet Studies for the Social Sciences (LISS Panel), are easily available and initiatives have been undertaken to link and extend these datasets into a full systems of linked open data (LOD). Furthermore, there are many ways linking different sources of data can be advantageous (Jutte et al., 2011), from linking survey/administrative data with case-control studies to investigate the effects of socio-economic factors in shaping health outcomes (Kozyrskyj et al., 2009) [NEEDS BETTER CITATION]; to life-course and trans-generational studies monitoring social, psychological and medical features [NEEDS CITATION]

Making use of the full potential of these data sets requires dealing with the crucial problem of multivariate missing data. Missing data can occur on these types of data because of traditional reasons (e.g. attrition, unwillingness to answer sensitive questions), because of errors in the linkage, or because some individuals do not interact with a specific service/activity recorded in the considered data (Harron et al., 2017).

The tools researchers working large social surveys and linked data need to correct for the bias introduced by non-responses require special attention. In general, when performing Multiple Imputation (MI) (Rubin, 1987), one of the most widespread principled method to deal with missing cases, data handlers tend to prefer including more, rather than less, predictors in their imputation models. This practice increases the dimensionality of the imputation models but reduces the chances of specifying uncongenial imputation and analysis models (Meng, 1994) and of leaving out important predictors of missingness, which is important to meet the MAR assumption, a basic requirement for proper imputations. On top of this standard source of dimensionality, the large number of items included in survey and linked data, coupled with their longitudinal nature, and the necessity of preserving complex interactions and nonlinear relations, easily produces high-dimensional ($p > n$) imputation problems.

When data is sparse (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 imputation models in any parametric Multiple Imputation procedure. As a result, high dimensionality of the data matrix prevents a straightforward application of MI algorithms, such as MICE (van Buuren, 2012).

However, high-dimensional data imputation settings represent both an obstacle and an opportunity: an obstacle, as in the presence of high-dimensional data it is simply not possible to include all available variables in standard parametric imputation models; an

opportunity, because the large amount of features available has the potential to reduces the chances of leaving out of the imputation models important predictors of missingness.

(Discuss background literature)   Many solutions have been proposed to deal with missing values in high dimensional contexts. Some researchers have focused on single imputations 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 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 task. Its application to high dimensional data has been directly tackled by specific algorithms using either shrinkage or dimensionality reduction methods (Song and Belin, 2004; Zhao and Long, 2016; Deng et al., 2016). Furthermore, other methods, that could potentially suit well the purpose, are the use of dimensionality reduction within the imputation models (Howard et al., 2015), or the use of non-parametric prediction trees (Burgette and Reiter, 2010; Doove et al., 2014). However, most of these have been either proposed or tested exclusively in low-dimensional imputation settings.

(Focus/Reason to write paper)   With this article we set out to provide a comparison of these state-of-the-art imputation algorithms in high-dimensional scenarios. We compare imputation methods based on their ability to allow inferential statements that are as valid as if they were made on a dataset without missing data. Hence, in assessing the methods performances, the primary focus of this article is the *statistical validity* (Rubin, 1996) of the substantive analysis performed on data treated with different high-dimensional MI procedures. The comparison is developed both through simulation studies and a real survey data application.

The end goal is to provide recommendations for applied researchers in how do deal with missing values in a principled and achievable manner when faced with large social surveys and linked data.

(Content Summary)   This paper is organized as follows. Section 2 discusses the imputation methods compared. Section 3 presents two simulation studies, their design and the result of the comparison. Section 4 presents a 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, description of the limitations of the study, and future 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 $T < p$ variables with missing cases in at least one, and that they are also part of some substantive model of scientific interest. An imputation procedure targeting these T variables could be used to allow fitting a substantive model (e.g. some linear or logistic regression) without discarding data units (rows). The $p - T$ variables in the dataset constitute a pool of possible *auxiliary* variables that could be used to improve the imputation procedure.

Most of the methods described in this section iteratively impute each target variable with imputation models that use as predictors the other variables afflicted by missing values and the information contained in the auxiliary data.

2.1 Multiple Imputation Strategies

2.1.1 MICE with Bayesian Ridge (bridge)

The *bridge* imputation procedure closely follows a standard iterative MICE algorithm for imputation of multivariate missing data (van Buuren, 2012, p. 120, algorithm 4.3). At iteration m, for each target variable, plausible values of the imputation model parameters are drawn from its posterior distribution, and imputations are drawn from the posterior predictive distribution.

After initialization of the missing values, at each m-th iteration, the following sampling steps are performed for each target variable:

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

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

where $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ are draws from the parameters posterior distribution (1) and posterior predictive distribution (2), respectively, for the j-th target variable at the m-th iteration. The superscript (m) implies that the missing values in $\boldsymbol{Z}_{obs,j}^{(m)}$ and $\boldsymbol{Z}_{mis,j}^{(m)}$ are different at every iteration as they are filled in with the previous iteration draws.

The sampling of each $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ is done as in the standard *Bayesian imputation under normal linear model algorithm* described by (van Buuren, 2012, p. 68, algorithm 3.1) and implemented as in the *impute.mice.norm()* function of the *mice* R package. The algorithm uses a ridge penalty to avoid problems with the inversion of singular observed data matrix that can afflict the sampling in (1). By adding a biasing ridge penalty, singularity is circumvented and the sampling scheme described above is possible even on data that is sparse or afflicted by high collinearity.

2.1.2 MICE with Bayesian lasso (blasso)

A Bayesian Lasso linear model is a regular Bayesian multiple regression with a prior specification 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.

The Bayesian Lasso imputation algorithm (blasso) used here is a standard Multiple Imputation MCMC sampler that uses the shrinkage priors defined by Hans (2010) to compute the posterior distributions of the regression coefficients (which are used in (1)). Posterior parameters draws are then used to sample plausible values from the predictive distributions of the missing data. 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). 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 [LINK].

2.1.3 Direct Use of Regularized Regression (DURR)

As proposed by Zhao and Long (2016) and Deng et al. (2016), Regularized Regression can be directly used in a MICE algorithm to perform multiple imputation of high dimensional data. For a target variable $\boldsymbol{z}_j$, the DURR algorithm follows these directions:

- Generate a bootstrap sample $\boldsymbol{Z}^*$ by sampling with replacement rows of $\boldsymbol{Z}$. Denote $\boldsymbol{z}_{j,obs}^*$ and $\boldsymbol{Z}_{j,obs}^{*(m)}$ as the observed part of $\boldsymbol{z}_j^*$ and the corresponding values on the other variables in $\boldsymbol{Z}^*$, respectively. Suffix m is used to clarify that at each iteration $\boldsymbol{Z}_{j,obs}^{*(m)}$ is different as it includes values previously imputed on the other target variables.

- Use any regularized regression method (such as Lasso regression) to fit a linear model with $z_{j,obs}$ as outcome and $Z_{j,obs}^{*(m)}$ as set of predictors. This produces a set of parameter estimates (regression coefficients and error variance) $\hat{\theta}_j^{(m)}$ that can be considered as sampled from the parameters' posterior distribution conditioned on the observed part of the data (1).

- Predict $z_{j,mis}$, the missing values on target variable z_j, based on $Z_{j,mis}^{*(m)}$ and $\hat{\theta}_j^{(m)}$, to obtain draws from the posterior predictive distribution of the missing data (2).

At iteration m, these steps are repeated to for each j-th variable in the set of T target variables. After convergence, M different sets of imputations are kept to form M differently imputed data sets. Any substantive model can then be fit to each data, and estimates can be pooled appropriately.

2.1.4 Indirect Use of Regularized Regression (IURR)

While DURR performs simultaneously model trimming and parameter estimation, 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 $z_{j,obs}$ as dependent variable and $Z_{j,obs}^{(m)}$ as predictors (compared to DURR, there is no asterisk in the notation as the original data is used, not a bootstrap version). 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 $z_{j,obs}$ on the active set of predictors in $Z_{j,obs}^{(m)}$ and draw a new value of these coefficients by sampling from a multivariate normal distribution centered around these MLEs.

$$(\hat{\theta}_j^{(m)}, \hat{\sigma}_j^{(m)}) \sim N(\hat{\theta}_{MLE}^{(m)}, \hat{\Sigma}_{MLE}^{(m)}) \tag{3}$$

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

After convergence is reached, M differently imputed data sets are kept and used for the substantive analysis.

2.1.5 MICE with PCA (MICE-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 perform a standard MICE algorithm in a well-behaved low dimensional setting. The MICE-PCA imputation procedure can be summarized as follows:

- Extract Principal Components from all variables in Z that are not part of set T

- Create a new data matrix Z' by combining the target variables with the first principal components that cumulative explain at most 50% of the variance in the auxiliary variables.

- Use a standard MICE algorithm for imputation of multivariate missing data to obtain multiply imputed datasets from the low dimensional $\mathbf{Z}'$ and the set of target variables.

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 that implements and developed its ideas.

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

A variety of Multiple Imputation methods using regression and classification trees have been proposed (Reiter, 2005; Burgette and Reiter, 2010; Shah et al., 2014) They all share the following core steps:

- For a given variable z_j, target of imputation, a CART algorithm partitions $\mathbf{Z}_{j,obs}^{(m)}$ to identify a collection of leafs with homogeneous $z_{j,obs}$ values. Each leaf contains a subset of the observed z_j, called donors.

- Each unit with a missing value on the target variable is placed in one of the leafs based on its $\mathbf{Z}_{j,mis}^{(m)}$ values.

- Each missing value on z_j is sampled from the pool of corresponding leaf donors.

At iteration m, these steps are followed for all of the T target variables. After convergence, the last M datasets are kept as multiply imputed datasets that can be used for the analysis and pooling phases.

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) used in this paper is an adaptation of the one described for MI-CART. To impute z_j at iteration m, MI-RANF first draws K bootstrap samples from the rows of the data with observed z_j. One tree is fitted to every bootstrap sample, with random features selection, and donors are identified. Imputations are then drawn from a pool of donors combined from the K trees that have been fitted to Z_{obs}. Imputations are not sampled from donor values averaged across trees as this procedure would reduce the uncertainty incorporated in the imputation model.

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

We have also used an ideal standard MICE with Bayesian Linear Regression approach (MI-OP) that considered, for each target variable imputation model, the following groups of predictors:

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 recommended strategies to deal with a large number of possible predictors for the imputation models (van Buuren, 2012, p. 168). In this sense, it represents an *ideal* strategy that could be used to deal with high-dimensional data,

in the absence of alternatives. 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 applicable in practice, but it does offer an interesting benchmark for comparison.

2.2 Single data strategies

2.2.1 Single Imputation

We consider the MissForest imputation method proposed by Stekhoven and Bühlmann (2011). Being a non-parametric imputation approach it does not suffer from the problem of unidentified imputation models and it can accommodate for mixed data type of the missing variables. However, as a single imputation method we do not expect it will allow to perform statistically valid inference on the treated data.

2.2.2 Mean Imputation and Complete Case analysis

In the social sciences, and especially in the analysis of social surveys, imputing the mean of the observed values on a variable is still a quite popular choice in dealing with missing data. Therefore, we include this method to portray a picture of the possible improvements the different high-dimensional imputation algorithms can achieve.

Finally, for the sake of comparison, two additional approaches are considered that do not involve imputation: list-wise deletion (or CC, complete case analysis), which entails fitting the analysis models exclusively on the complete rows of the data; and a gold standard analysis (GS) which consists of fitting the substantive models on the underlying fully observed data and represents 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 ($R = 1000$) for each experiment:

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

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

3. Imputations - Each method described in section 2 to deal with missing values was used to impute NAs.

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 the available cores (between 20 and 30). 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 [INSERT GITHUB LINK].

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

3.1.1 Step 1: Data generations

Experiment 1 The $\boldsymbol{X}_{n \times p}$ data matrix in step 1 was generated by drawing from a multivariate normal model with a mean vector $\boldsymbol{\mu_0}$ of p 0s and 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 $\boldsymbol{X}_{n \times p}$ 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 $p = 5 \times l$ number of predictors in $\boldsymbol{X}$. Values on the observed items for the i-th observation were obtained with the following measurement model:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}\boldsymbol{\xi}_{i.} + \boldsymbol{\delta}_{i.} \qquad (4)$$

where $\boldsymbol{x}_i$ 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 [CITE Bollen1989].

The latent scores in $\boldsymbol{\xi}_{i.}$ are sampled from a multivariate normal distribution centered around an $n \times l$ vector of 0s, and 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 variancess 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).

3.1.2 Step 2: Missing data imposition

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

$$p(x_t = NA|X) = \boldsymbol{\Phi}(\tilde{X}\boldsymbol{\theta}) \qquad (5)$$

with x_t a variable target of missing data imposition, $\mathbf{\Phi}(.)$ being the logistic cumulative distribution function, $\tilde{X}$ the matrix of predictors participating in the missing data mechanism, and $\boldsymbol{\theta}$ a vector of non-trivial regression coefficients.

An offsetting constant was added to the linear combination $\tilde{X}\boldsymbol{\theta}$ to make observations with lower values of $\tilde{X}\boldsymbol{\theta}$ have higher chances of having a missing value on the target x_t. The offsetting constant was chosen to minimize 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 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 (5) with 4 variables included in $\tilde{X}$: two fully observed variables from the highly- 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 for the imputation methods. 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. The predictors in $\tilde{X}$ are always provided to the imputation algorithms so that the MAR assumption can be met.

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 (5).

3.1.3 Step 3: Imputation

Missing values are dealt with according to all the methods described in section 2.

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, an oracle Confirmatory Factor Analysis was also chosen to see how the factor scores are recovered after imputation.

3.2 Conditions

The simulation procedure just outlined was repeated with different experimental set ups. 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, and pm, the target proportion of *per* variable missing cases. 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: .1 or .3.

In experiment 2, we have also defined the latent structure 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 .5 and .6, or or .9 and .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

3.3 Comparison Criteria

After running the simulation procedure R times, the R Gold Standard estimates obtained by saving the item means, variances and covariances, 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) and Standardized Bias (SB) to quantify the bias introduced by the imputation procedures:

$$PRB = \frac{\bar{\hat{\theta}} - \theta}{\theta} \times 100 \tag{6}$$

$$SB = \frac{\bar{\hat{\theta}} - \theta}{SD_{\hat{\theta}}} \tag{7}$$

θ is the "true" value) of the focal parameter (e.g. mean of item 1, variance of item 2) and it is computed as $\frac{\sum_{r=1}^{R} \hat{\theta}_r^{GS}}{R}$, 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 $\frac{\sum_{r=1}^{R} \hat{\theta}_r^m}{R}$, with $\hat{\theta}_r^m$ being the estimate obtained after using the m imputation approach in the r-th repetition. $SD_{\hat{\theta}}$ represents the empirical standard deviation of θ.

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

$$CIC = \frac{\sum_{r=1}^{R} I(\hat{\theta} \in \widehat{CI}_r)}{R} \qquad (8)$$

where R is the total number of MCMC repetitions, θ and $\hat{CI}_r$ are, respectively, the estimate and the confidence interval for the focal estimates 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 generally 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.

In the present work, we followed Burton et al. (2006) and considered as problematic CI coverage rates outside of two SEs of the nominal coverage probability (p). The standard error of nominal coverage is defined as $SE(p) = \sqrt{p(1-p)/R}$, with $p = .95$.

An additional measure that can help understanding Confidence Interval Coverages is the width of the confidence intervals:

$$CIW = \frac{\sum_{r=1}^{R} [\widehat{CI}_{r,upper} - \widehat{CI}_{r,lower}]}{R} \qquad (9)$$

with $\widehat{CI}_{r,upper}$ and $\widehat{CI}_{r,lower}$ the upper and lower bounds, respectively, of the estimated confidence interval for a parameter estimate in the r-th replication. CIW is an indicator of the statistical efficacy of the imputation methods. There are no rules of thumb for interpreting this measure, but comparing the relative width of 95% Confidence Intervals after imputation with the ones obtained with their Gold Standard value is quite informative. In general, for the same confidence interval coverage, imputation methods that have smaller CIW are considered more efficient. Single imputation methods will tend to provide narrower CIs compared to Gold Standard estimates, while proper multiple imputations will tend to provide wider CIs, reflecting the greater uncertainty regarding a parameter value due to presence of missing values.

Euclidean Distance   Both bias and CIC are computed for individual parameter estimates. When many parameters are involved in the analysis model these measures become cumbersome to compare for every parameter. Hence, we have also used multivariate measures to assess the overall statistical validity of the analysis models of interest after imputation.

For a given m imputation approach, a multi-parameter measure of bias is computed as the Euclidean Distance between the t-dimensional vector of true model parameters values $\boldsymbol{\theta}$ and the vector of MCMC-average after-imputation estimates $\bar{\hat{\boldsymbol{\theta}}}^m$:

$$d_{bias}^m(\boldsymbol{\theta}, \bar{\hat{\boldsymbol{\theta}}}) = \sqrt{\sum_{i=1}^{t} (\theta_i - \hat{\theta}_i)^2} \qquad (10)$$

where $\hat{\theta}_i$ is an element of the t-dimensional $\bar{\hat{\boldsymbol{\theta}}}$, itself computed as $\frac{\sum_{r=1}^{R} \hat{\boldsymbol{\theta}}_r}{R}$, with $\hat{\boldsymbol{\theta}}_r$ the vector of the model parameters estimates in the r-th repetition.

A multi-parameter measure of Confidence Interval Coverage is computed as Euclidean Distance between a t-dimensional vector ($\boldsymbol{CIC}^n$) of nominal coverage values (.95) and the vector of MCMC actual coverages for all model parameters $\boldsymbol{CIC}^a$:

$$d_{CIC}^m(\boldsymbol{CIC}^n, \boldsymbol{CIC}^a) = \sqrt{\sum_{i=1}^{t} (CIC_i^n - CIC_i^a)^2} \qquad (11)$$

CIC_i^a computed as in equation 8 for each of the t parameters to be estimated.

3.4 Results

3.4.1 Experiment 1

Saturated Model Figure 1 reports the Percentage Relative Bias computed for each parameter estimate in the saturated model described above: item means, variances, and covariances for the six variables with missing values.

Focusing first on the item means (top row), all methods achieve a bias that is smaller than the 10% threshold for all item means, in all conditions. Looking at relative performances, we can see how IURR and MI-PCA always result in the smallest estimation bias. Bridge competes with these two methods in the low dimensional conditions (1 and 3), but it does lose ground in the higher dimensional conditions (2 and 4).

Moving to the item variances (central row) we notice that IURR, Blasso, and the tree based methods are giving the lowest item variance biases across all conditions, even in the most challenging one (condition 4). Somewhat surprisingly, directly using regularized regression within the imputation models (DURR) leads to a bias larger than 10% in the high dimensional condition with high proportion of missing values.

It is interesting to note that while, the single imputation method, missForest, leads to extreme negative bias (above 10%) for the item variances, MI-PCA overestimates the variance of the variables with missing values. As a single imputation method, missForsest tends to reduce the variance of the imputations while MI-PCA reflects a lot of uncertainty regarding the imputations. While not ideal, the latter behaviour is certainly less alarming than the former.

Finally, the third row in table 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: MI-PCA is the only method with all covariances PRB lower than 10% in the high dimensional conditions. Indirect Use of Regularized Regression (IURR) performs noticeably better than all other methods, but it seem to struggle with a high bias for the majority of the 15 covariances in condition 4.

Figure 2 reports the Confidence Interval Coverage computed for each parameter. The pattern of performances is quite similar to that described by bias with MI-PCA and IURR outperforming most other methods in the high dimensional conditions, with coverages that are consistently between the .93 and .97 Burton thresholds. This suggests that MI-PCA and IURR are the imputation methods doing the best job at preserving the integrity of hypothesis testing performed on the treated data.

However, there are a some peculiarities to note:

- MI-PCA is the only method that does not seriously undercover means and covariances in condition (4);

- compared to all other methods, MI-PCA tends toward over- rather than under-coverage, and it does so to a lesser degree;

- while the bias for item variances obtained with IURR was quite small in condition 4, the method seems to undercover the true values of this type of parameter, suggesting confidence intervals that are bit smaller than they should.
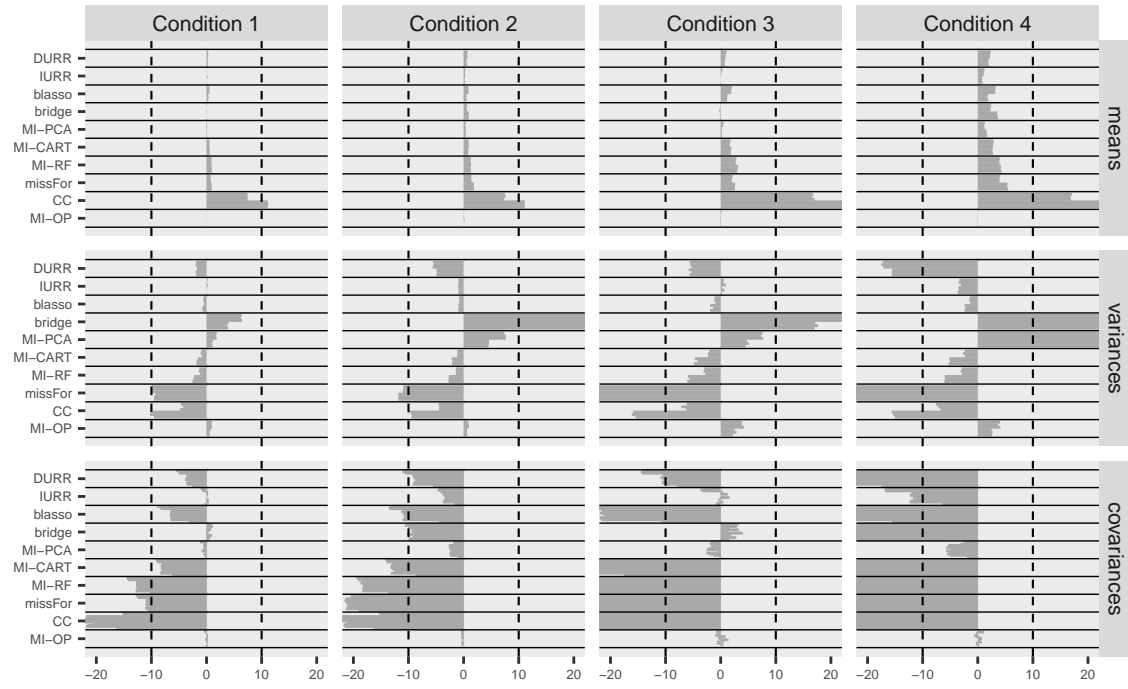
Figure 1: Percent Relative Bias (PRB) for item means, variances, and covariances broken down by method
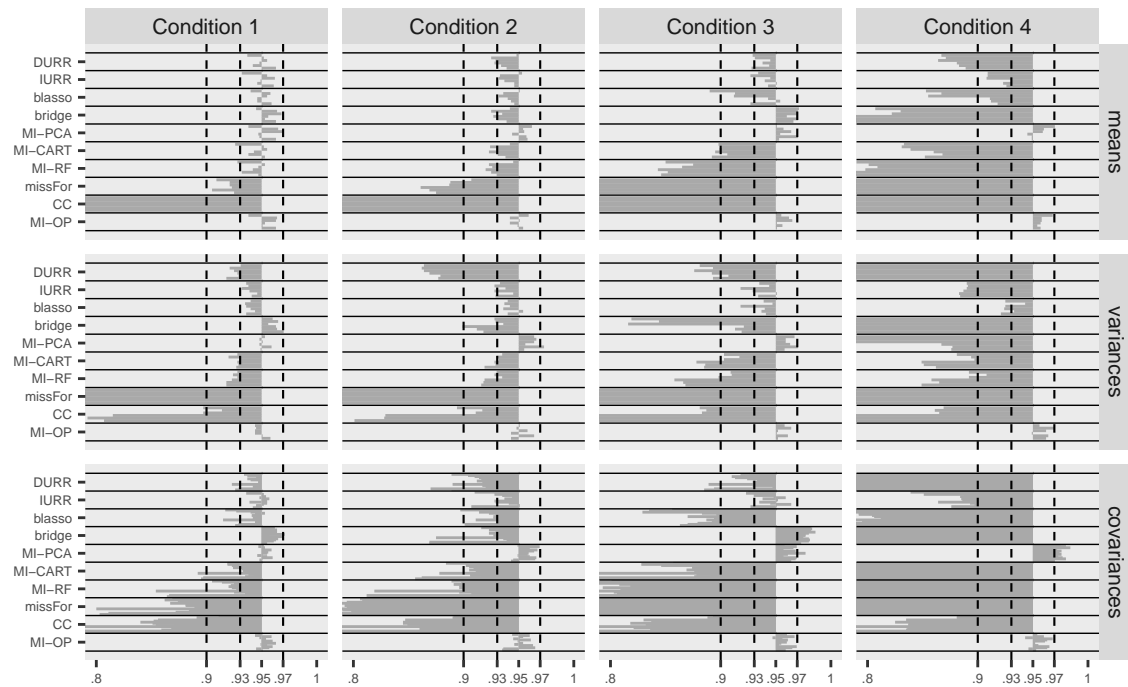

Figure 2: Confidence Interval Coverage (CIC) for item means, variances, and covariances broken down by method.

3.4.2 Experiment 2

Saturated Model Results form the first experiments held mostly constant in experiment 2, indicating that the presence and strength of a latent structure does not perturb the methods' relative performances.

Figure 3 reports the bias of the saturated model parameter estimates for the first four conditions of experiment 2 (high factor loadings). For both item variances and covariances PRB is reported, while for the item means we reported the SB. Items were generated around a mean of 0 making the computation of PRB for this parameter meaningless.

The least biased estimates for means and variances are obtained with IURR, in all conditions. Imputing missing values with the MI-PCA approach also grants low biases in all conditions. Bridge is also performing quite well with the exception of covariance estimates in condition 4.

In agreement with what was found in experiment 1, the MI-PCA approach is the one resulting in the lowest bias for all covariance estimates in all conditions. Surprisingly, the bias for the item variances that afflicted MI-PCA in the multivariate-normal set up disappears when the latent structure is strong (factor loadings larger than .9).

Figure 4 shows results for the confidence interval coverage in experiment 2. When factor loadings are high, we see that all multiple imputation methods lead to acceptable coverage for means and variances, in the conditions with low proportion of missing values, no matter the dimensionality of the data: in condition 1 and 3, for both item means and variances, confidence intervals coverage is approximately within .93 and .97 for all methods.

As the proportion of missing values increases we see a general deterioration in CIC performances, with IURR and MI-PCA still showing the most contained deviations from the target value. Again, MI-PCA tends to include the true parameter values more than it should (over-coverage), while most other methods show sings of under-coverage.

Given the large positive biases obtained by all methods for the covariances of the observed items, it comes to no surprise that most methods lead to under-coverage of these parameters in all conditions in experiment 2. MI-PCA is again the only exception providing acceptable coverage for all covariances.

The same patterns can be seen in the conditions with lower factor loadings (see figures 12 and 12 reported in appendix). However, in condition 8, the MI-PCA approach tends again to over-estimate the item variances (PRBs > 20%) and it also leads to extreme under-coverage of this parameters.
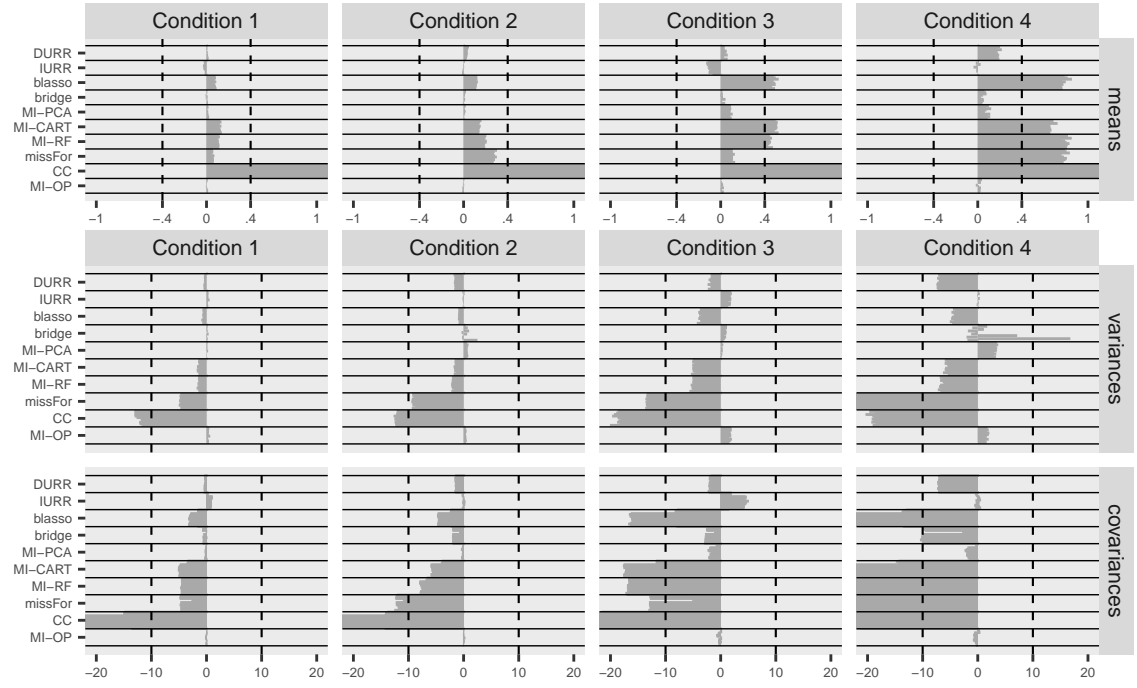
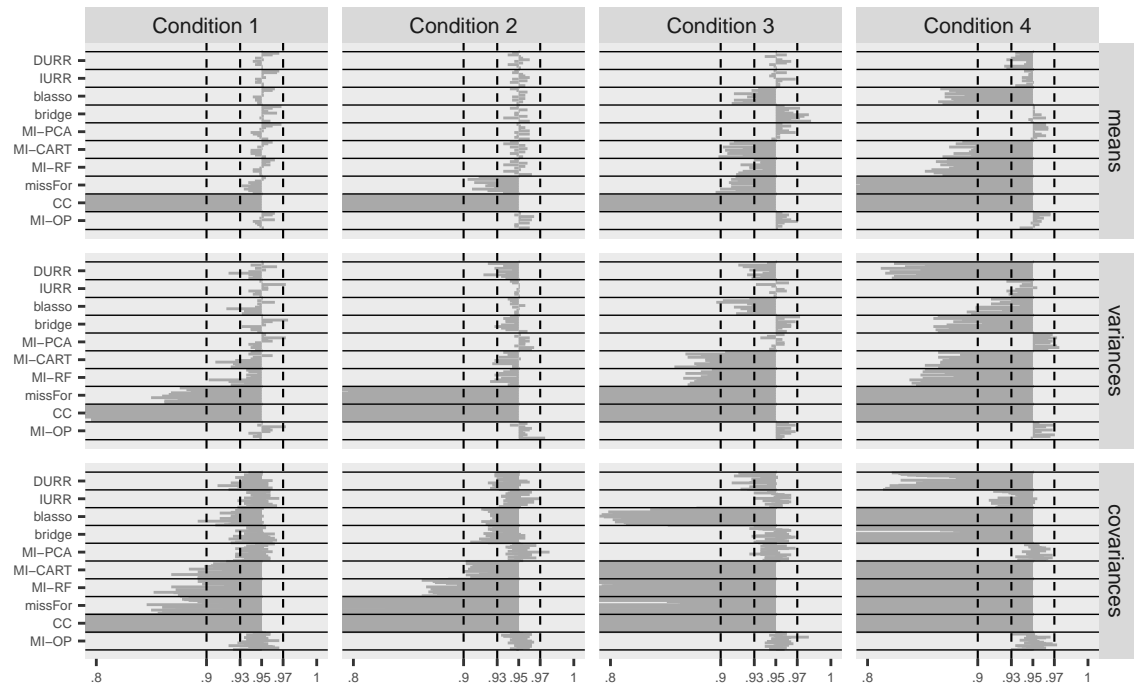Figure 3: Bias estimation for the means (SB), variances and covariances (PRB) for condition 1 to 4.

Figure 4: Confidence Interval Coverage (CIC) for the means, variances, and covariances for condition 1 to 4.

15

Confirmatory Factor Analysis Figures 5 shows the PRB values for all the factor loadings estimated by the Confirmatory Factor Analysis described above. Most MI-Methods are able to provide acceptably low biased estimates for these parameters in all conditions except the ones with both large proportion of missing values and high dimensional input data matrix (condition 4).

IURR and MI-PCA are again the two top performers giving virtually unbiased estimates of the factor loadings in all conditions. However, MI-PCA outperforms IURR when factor loadings are low, maintaining inconsequential biases even when data is high-dimensional and the proportion of missing values is high.
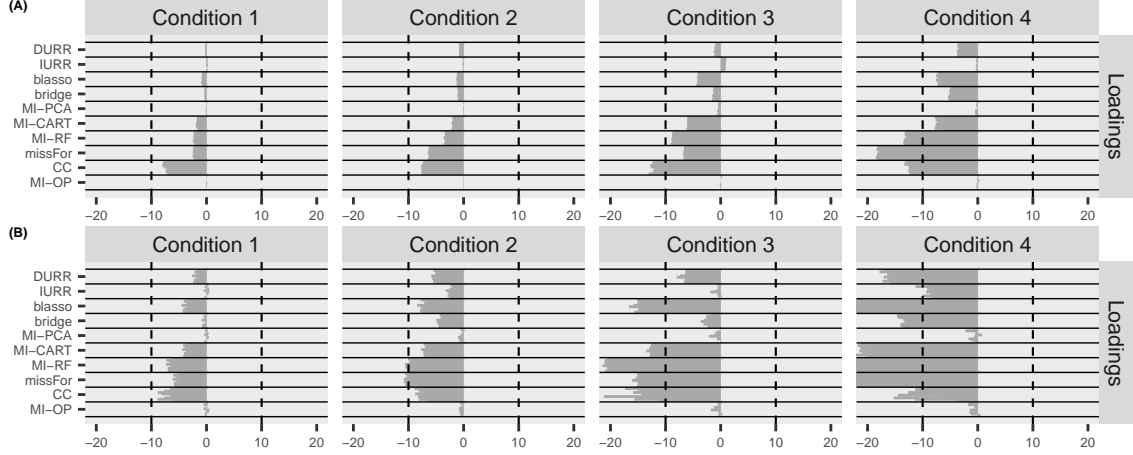
Figure 5: Percent Relative Bias (PRB) for the factor loadings conditions 1 to 4 (panel A) and conditions 5 to 8 (panel B).

4 Resampling Study

To test the ecological validity of findings in experiment 1 and 2 we have also designed a resampling study based on European Values Survey data. Variables in the EVS data are not generated artificially from continuous normal distributions but are discrete numerical and categorical items following a variety of distributions. By using data gathered with an actual survey, we can observe whether the relative performances of the imputation methods, displayed in the simulation studies, change when they are deployed for real data research.

4.1 Resampling Study Procedure

The resampling study follows a similar strategy to that used in the simulations. To assess the statistical validity of the different imputation methods we have repeated the following steps 500 times ($R = 500$):

1. Data generation - A bootstrap sample $\boldsymbol{X}^*$ is generated by sampling with replacement n observations from a pre-processed EVS data-matrix. Part of the pre-processing step is some form of imputation used to obtain a pseudo-fully observed input data matrix so that $\boldsymbol{X}^*$ is fully observed;

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

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

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

The average estimate, over the R repetitions, obtained with the Gold Standard approach are considered as "true" reference values of the parameters in the analysis models. The R estimates obtained with all other methods are used to obtain performance measures for each imputation method using the same criteria described for study 1 and 2 (see 3.3).

The code to run the simulation was written in the R statistical programming language (version 4.0.3). The resampling study was 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 the available cores (between 30). 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'

4.1.1 Data preparation

EVS is a standardized cross-sectional survey with a representative sample of more than 60,000 people, across more than 30 countries, interviewed via Web, post or face-to-face. For this study we have 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 european founding countries included in the data (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). The full cleaning process is more systematically described in the appendix.

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 considered as the population data.

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

4.1.2 Analysis models

To define plausible analysis models we have searched an EVS database, available on their website, looking for suitable analysis models to test the effectiveness of the different imputation algorithms. As a results, we have defined two linear regression models of the same form:

$$y_1 \sim \beta_{0,1} + \beta_{1,1}x_{1,1} + \boldsymbol{\beta}_{-1,1}\boldsymbol{X}_{-1,1} \tag{12a}$$

$$y_2 \sim \beta_{0,2} + \beta_{1,2}x_{1.2} + \boldsymbol{\beta}_{-1,2}\boldsymbol{X}_{-1,2} \tag{12b}$$

The first version of the linear model in 12a we used is inspired by Köneke (2014). The dependent variable y_1 is a 10-point EVS item measuring euthanasia acceptance ('Can this always be justified, never be justified, or something in between?'); the predictor of interest $x_{1,1}$ is 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 such as measures of trust, education, and socio-economic status.

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

The second version of the linear model in 12b is inspired by Immerzeel et al. (2015). The dependent variable y_2 is 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 $x_{1,2}$ is 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 included the usual socio-economic background information, the same measure of religiosity used in model 12a, and some measures of political interest.

A researcher might fit this model and look at $\beta_{1,2}$, the 'nativist attitude' regression coefficient, value and standard error to test a theory 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 are y_1 and y_2, the two dependent variables in models 12a and 12b, religiosity ($x_{1,1}$ in model 12a, and part of $X_{-1,2}$ in model 12b), 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 5 and 3 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).

Conditions   There were only two conditions for the resampling study: low and high dimensional 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 of the procedure outlined in section 4.1. We chose only two values for n, namely 1000 and 300, corresponding to the low and high dimensional condition.

4.2   Results

4.2.1   Bias

PRB   Figure 6 reports the PRB for the regression coefficients $\beta_{1,1}$ and $\beta_{1,2}$, in model 12a and 12b (henceforth model 1 and 2), respectively. Most of the MI methods result in negligible biases ($PRB < 10\%$) for both parameters in all conditions. The only two exceptions are bridge and MI-RF: the former is very competitive in condition 1, the low dimensional one, but leads to extreme bias in the high dimensional condition 2 for both parameters; the latter provides the highest PRB among the MI methods across the board, and it is consistently outperformed even by Complete Case analysis.

DURR and IURR are giving inconsequential biases for both parameters in all conditions, with PRBs that are often at least half in size as the ones obtain with the other methods, even outperforming MI-OP in the high dimensional condition for $\beta_{1,2}$
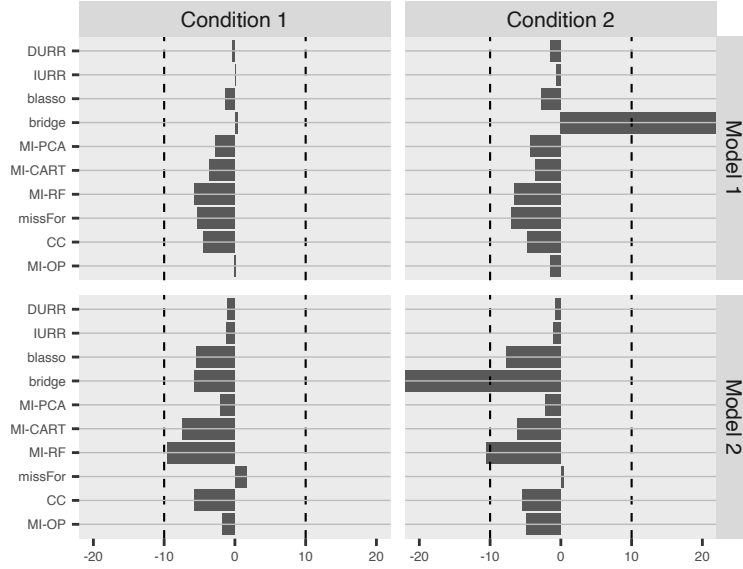
Figure 6: Bias for single parameter of interest in the two different models

Euclidean Distance Figure 7 reports the Euclidean Distance between the vector of estimated regression coefficients for model 1 and 2.

IURR and DURR yielded the vectors of parameters estimates that are closer to the vector of true values. Their advantage over other methods is stark for model 2 while it is less marked in model 1, where MI-CART and Blasso perform equally well. For model 1, Bridge performs better than DURR and IURR in the low dimensional condition, but shows the same extreme deterioration of performance in the high-dimensional condition as described for the single parameter of interest case.

MI-PCA seems to struggle with bias for model 1, ranking last among the multiple imputation models for $d_{bias}(\boldsymbol{\theta}, \bar{\bar{\boldsymbol{\theta}}})$. In model 2, it grants a $d_{bias}(\boldsymbol{\theta}, \bar{\bar{\boldsymbol{\theta}}})$ that is lower than that of tree-based methods, although not as low as DURR and IURR.

4.2.2   Confidence Interval Coverage

Confidence Interval Coverage Researchers interested in testing their theories based on models like the ones described will be interested in the confidence interval coverage of the estimates of interest. Figure 8 reports the CIC for $\beta_{1,1}$ and $\beta_{1,2}$, the focal regression coefficients in the two models.

For $\beta_{1,2}$ in model 2, both IURR and DURR remain fairly competitive with CIC close to nominal levels, but the advantage they showed in terms of bias is not carried over to this criterion. Both MI-PCA and Blasso provides CICs that are either equal or closer to nominal than the ones obtained with IURR and DURR in almost all conditions.

For $\beta_{1,1}$ in model 1, condition 2 shows interesting patterns:

- all MI methods show signs of undercoverage with CIC smaller than the threshold value .93;

- all MI methods perform *worse* than Complete Case analysis, which covers the true value more frequently than with imputation (more on this below);

- all MI methods perform *better* than the Gold Standard, which covers the true value in approximately only 90% of the replications performed.
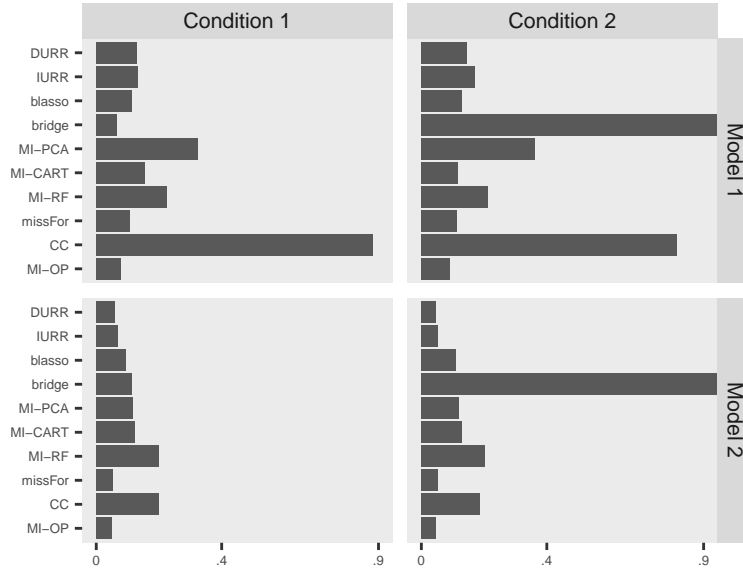
Figure 7: Euclidean distance between the vector of estimated regression coefficients and the vector of their true value

Euclidean Distance   When looking at the more general pattern described by d_{CIC}, reported in 9, IURR and DURR return to the top of the leader-board, providing the vectors of Confidence Interval Coverages closest to nominal levels. Furthermore, they are the only methods providing coverage equivalent to the Gold Standard method in all conditions for both models.

Confidence Interval Width   The seemingly competitive coverage showed by CC in figure 9 is achieved with much wider 95% Confidence Intervals than all other methods. Figure 10 shows the average Confidence Interval Width across model parameters. Apart from bridge in the high-dimensional conditions, CC is always showing the widest confidence intervals. This is of course due to the lower number of data observations used to obtained estimates, and it makes clear how the close to nominal coverage showed by CC in figure 9 is an artifice of the wider confidence intervals.

By a similar logic, we note that the less ideal behaviour of MI-PCA showed in figure 9, compared to IURR, is not due to wider or narrower confidence intervals. The difference is more plausibly due to the larger bias that MI-PCA tends to produce (see figure 7).

4.2.3   Imputation Time

Figure 11 shows the average imputation time across the different methods. IURR and DURR are 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.
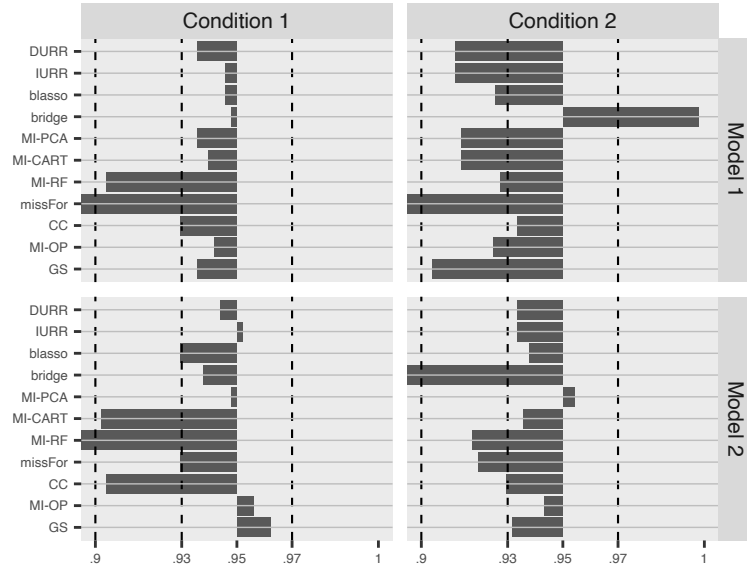
Figure 8: Confidence Interval Coverage for single parameter of interest in the two different models
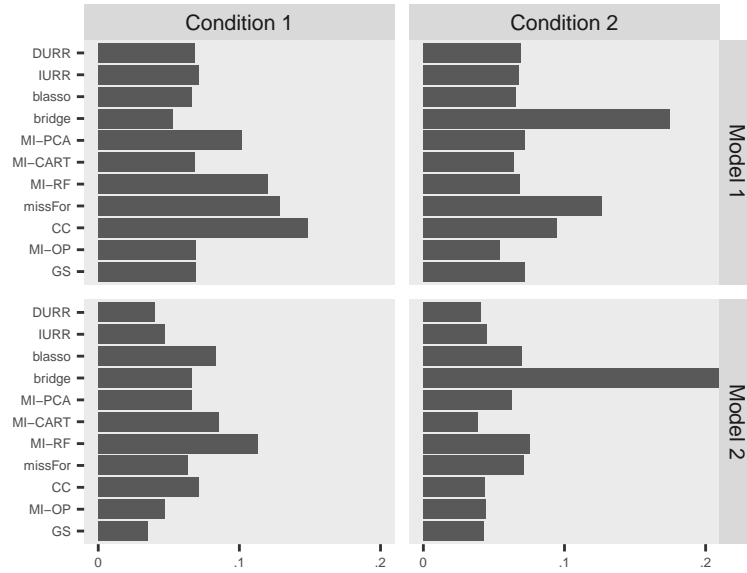

Figure 9: Euclidean distance between the vector of confidence coverages and the vector of nominal coverage
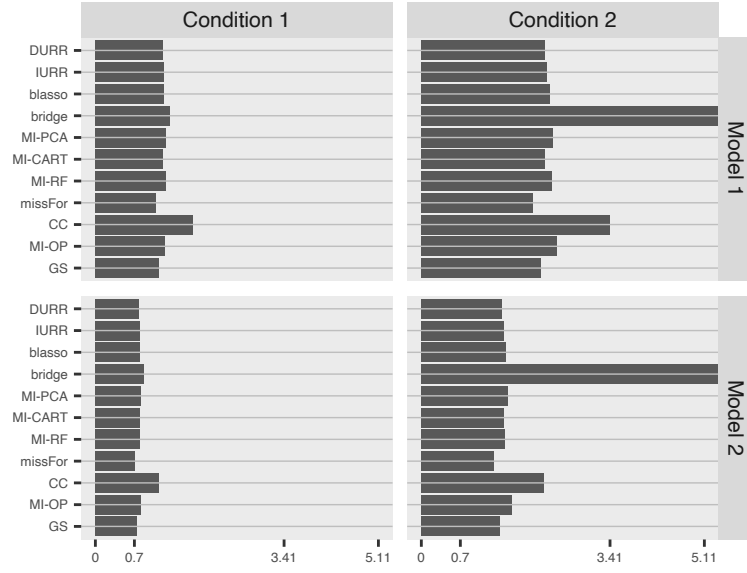
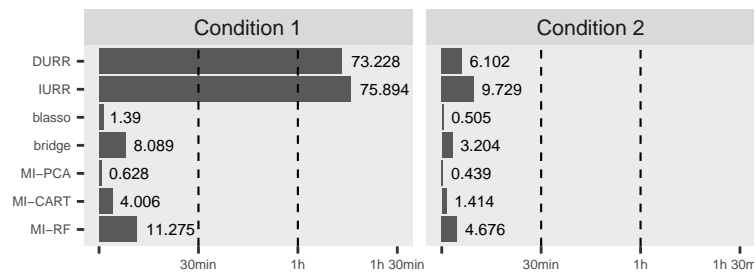Figure 10: Average Model Parameter Confidence Interval Width

Figure 11: Average imputation time for each method

5   Discussion

In what follows, the results obtained with the simulation and the resampling studies are discussed to provide an overall picture of how the methods compare to each other.

IURR and DURR effective but time consuming   Overall, DURR and especially IURR showed great performances in all experiments according to all performance measures. However, using them requires a lot more imputation time than other well performing methods. For example, while both IURR and MI-PCA have weaknesses, overall, they showed similar performances for most parameters and performance measures. However, MI-PCA obtained these results in a fraction of the time taken by IURR. Although these time measurement are specific to our set up, with six variables to be imputed, the relative differences are striking and bound to have a large impact in applied research.

MI-PCA struggle with item variances   In the simulation studies, MI-PCA tended to over-estimate (positive bias) the item variances while it was the only method with acceptable bias for the covariances. In other words, it seems that MI-PCA includes a lot of uncertainty regarding the imputed values, but it recovers correctly the relationships between variables.

   Furthermore, the bias for the item variances that afflicted MI-PCA in the multivariate-normal set up appears to be related to the strength of the latent structure: when the latent structure is absent (experiment 1) or weak (experiment 2, conditions 5 to 8, factor loadings between .5 and .6) item variances are biased, especially in the high-dimensional conditions; when the latent structure is prominent (experiment 2, conditions 1 to 4), the variances are estimated with negligible bias, even in the high-dimensional conditions.

MI-PCA IURR trade off   In the simulations, the two better performing methods, IURR and MI-PCA, exhibit different weaknesses and strengths in terms of bias. The two simulations studies, showed that, in the high dimensional condition, MI-PCA struggles with correctly recovering item variances but returns very lowly biased item covariances, while IURR has exactly the opposite behaviour.

   The resampling study showed that slightly more biased estimates of the parameters in model 1 and 2 are produced using MI-PCA rather than IURR, while coverage rates and confidence interval widths did not substantively differ. This difference did not show when looking at the focal regression coefficients, where MI-PCA performed as well as IURR, if not better.

   Overall, the simulation studies seemed to suggest that MI-PCA could recover more accurately the relationships between variables (less biased covariances) than IURR, while in the resampling study this conclusion was not supported to the same extent: when looking at bias and coverage rates for $\beta_{1,1}$ and $\beta_{1,2}$, MI-PCA performed equally well or better than DURR and IURR, while looking at the overall measures of bias and coverages IURR and DURR outperformed MI-PCA.

Bridge inadequacy for high-dimensional set ups   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 lead to extreme bias in the high dimensional conditions.

MI-CART and MI-RANF   [UNFINISHED] A few words on these methods. Probably need to pay more attention to them in the result section as well.

6 Conclusions

Recommendations / Take-home message [UNFINISHED] Give recommendations / take-home message in one or two paragraphs

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 data with any distribution: IURR and DURR have already been discussed for categorical data imputation, and MI-PCA can be performed with any standard imputation model for categorical data. Blasso has not been formally developed for multi-categorical imputation target variables yet, which limited the study to work with missing values on variables that are either continuous in nature or usually considered as such in practice. Furthermore, the interesting inclusion of interactions and squared terms in the imputation models was not explored as their inclusion as not been developed to the same extent across the different methods

[UNFINISHED] The resampling study compared results only for two analysis models and showed some variation in which methods were top performer.

References

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.

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.

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.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big data & society*, 4(2):2053951717745678.

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.

Jutte, D. P., Roos, L. L., and Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual review of public health*, 32:91–108.

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.

Kozyrskyj, A., HayGlass, K., Sandford, A., Pare, P., Chan-Yeung, M., and Becker, A. (2009). A novel study design to investigate the early-life origins of asthma in children (sage study). *Allergy*, 64(8):1185–1193.

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.

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

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.

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. 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.

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

7 Appendix

[UNFINISHED] - This appendix needs to include extra plots; EVS data cleaning description; algorithms?
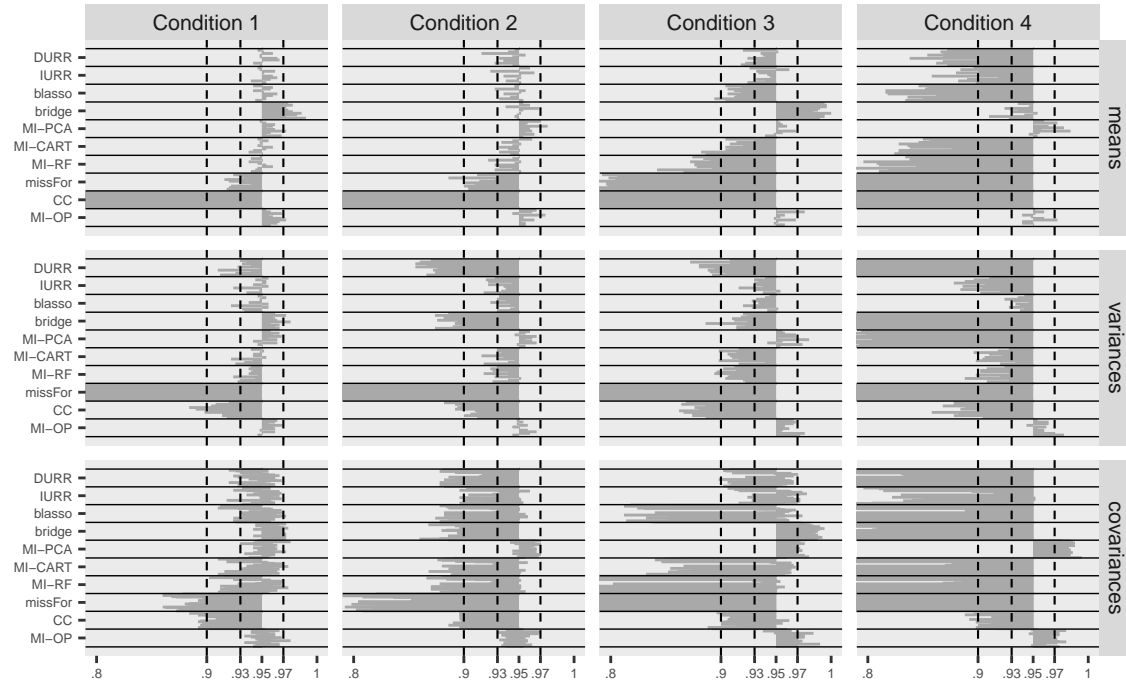

Figure 12: Bias estimation for the means (SB), variances and covariances (PRB) for condition 5 to 8.
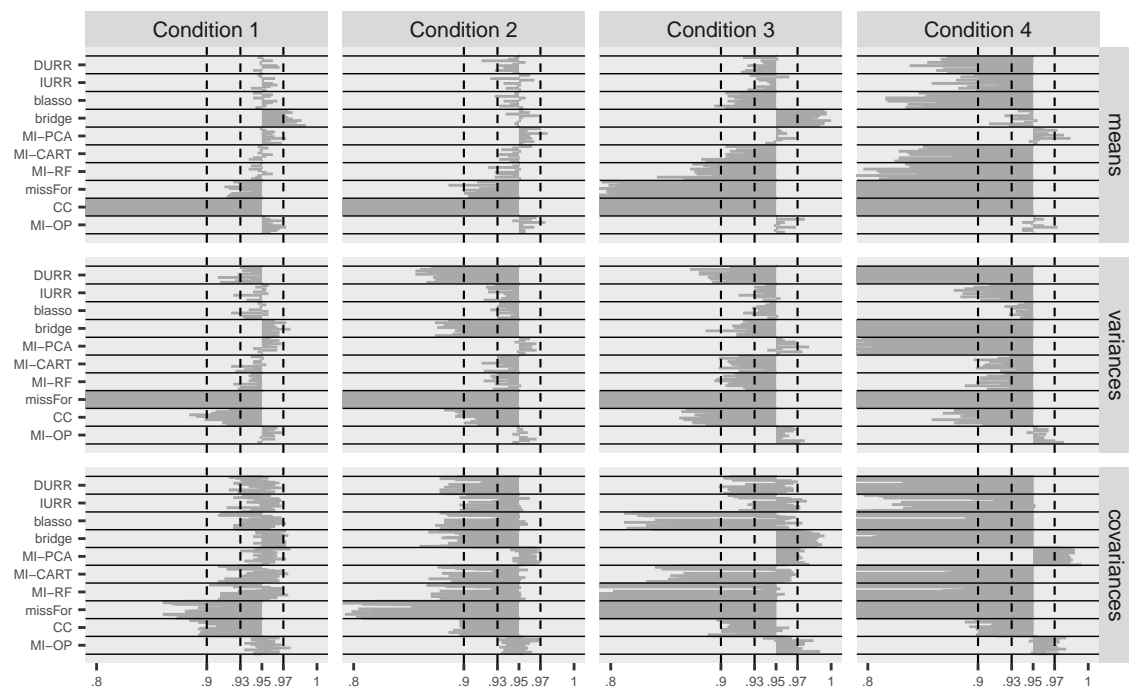
Figure 13: Confidence Interval Coverage (CIC) for the means, variances, and covariances for condition 5 to 8.