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

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

- **Introduction:** Frame problem; Discuss background literature; Focus/Reason to write paper; Content Summary.
- **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.
- **Simulation Studies**
  - Methods for Study 1 (MVN) + Study 2 (Latent Structure)
    - Data generation
    - Missing data imposition
    - Analysis models
    - Criteria
    - Procedure: Summary of crossed conditions, describe sequentially what happens during each replication
  - Results: distinguish by type of performance measure
- **Resampling Study (EVS)**
  - Methods
    - 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.
    - Missing data imposition
    - Analysis models
    - Criteria
    - Procedure: Summary of crossed conditions, describe sequentially what happens during each replication
  - Results: again divide by type.
- **Discussion:** Synthesize findings, make parallels and comparisons.
- **Conclusions:** Short take home message, limitations, future directions (hint at MY future work)
- Appendices - methods details - EVS quirks

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

Edoardo Costantini

November 25, 2020

1 Introduction

(Frame the problem) Today's social and behavioral scientists are blessed with a wealth of large, high-quality and publicly available social scientific datasets such as the Longitudinal Internet Studies for the Social Sciences (LISS) Panel and the European Values Study (EVS), with initiatives being undertaken to link and extend these datasets into a full system 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.

The tools researchers working with these data sets need to correct for the bias introduced by nonresponses require special attention. In general, when performing Multiple Imputation, data handlers tend to prefer including more, rather than less, predictors in the imputation models. This reduces the chances of uncongenial imputation and analysis models (Meng, 1994) and of leaving out important predictors of missingness. On top of this standard source of dimensionality, the large number of items recorded in surveys, 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 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 imputation algorithms, such as MICE (van Buuren, 2012).

High-dimensional data imputation settings represent both an obstacle and an opportunity in this sense: 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, there are other methods, that could potentially suit well the purpose, but have been tested only in low-dimensional settings (Burgette and Reiter, 2010; Doove et al., 2014; Howard et al., 2015).

(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 methods based on their ability to allow inferential statements that are as valid as if they were made on a dataset without missing data. The comparison is developed both through simulation studies and a real survey data application.

(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 row that are also part of the substantive model of 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 target variables 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 their posterior distribution, and imputations are drawn from the posterior predictive distribution.

After initialization of the missing values, at each m-th iteration, performs the following sampling steps 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 of singular matrices. When the sample covariance matrix is singular, it is not invertible, an operation that is key to the sampling of parameters in (1) (Schafer, 1997). By adding a biasing ridge penalty, singularity is circumvented and the sampling scheme described above is possible even on data affected by high collinearity and/or with a higher number of columns than rows ($p > n$).

2.1.2 MICE with Bayesian lasso (blasso)

A Bayesian hierarchical BLasso 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).

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 $\boldsymbol{Z}^{*(m)}_{j,obs}$ as the observed part of z^*_j and the corresponding values on the other variables in $\boldsymbol{Z}^*$, respectively. Suffix m is used to clarify that at each iteration $\boldsymbol{Z}^{*(m)}_{j,obs}$ 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 $\boldsymbol{z}_{j,obs}$ as outcome and $\boldsymbol{Z}^{*(m)}_{j,obs}$ as set of 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 the parameters' posterior distribution conditioned on the observed part of the data (1).

- Predict $\boldsymbol{z}_{j,mis}$, the missing values on target variable z_j, 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).

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 $\boldsymbol{z}_{j,obs}$ as dependent variable and $\boldsymbol{Z}^{(m)}_{j,obs}$ 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 $\boldsymbol{z}_{j,obs}$ 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.

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

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

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 $\boldsymbol{Z}$ that are not part of set T
- Create a new data matrix $\boldsymbol{Z}^{'}$ 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 $\boldsymbol{Z}^{'}$ 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* 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 $\boldsymbol{Z}_{j,obs}^{(m)}$ 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 $\boldsymbol{Z}_{j,mis}^{(m)}$ 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 imputation model predictors (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 case.

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.

3.0.1 Procedure

To assess the statistical validity of the different imputation methods we have repeated the following steps 500 times ($R = 500$) for each experiment

- 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 also depend on experimental factors described below.
- 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.
- Imputations - Each method described in section 2 to deal with missing values was used to impute NAs.
- Analysis - Different analysis models were fitted to the differently treated data. Parameters estimates 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 R estimates obtained with the Gold Standard approach are averaged and 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 (see below).

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'.

3.1 Methods: two simulation studies

3.1.1 Data generations

Experiment 1 The $\boldsymbol{X}_{n\times p}$ data matrix was generated by drawing from a multivariate normal model with a vector $\boldsymbol{\mu_0}$ of mean p 0s and a covariance matrix $\boldsymbol{\Sigma}_0$, with diagonal elements (variances) equal to 1. $\boldsymbol{\Sigma}_0$ was 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.

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 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:

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

where $\boldsymbol{x}_i$ is a vector of $5 \times l$ scores on each observed item for observation $i = 1, ..., n$, $\boldsymbol{\Lambda}$ 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.

Latent scores 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 load on only 1 factor (5 items for each latent variable). Both the items and latent factor variance are set to 1, $var(x_i) = 1$ and $\Psi_{ii} = 1$, so that the measurement error is defined as $var(\delta) = 1 - \lambda^2 1$. Factor loadings λ_{ij}, with $i = 1, ..., n$ and $j = 1, ..., l$, are hence defined as standardized values that range between 0 and 1.

If all values in $\mathbf{\Lambda}$ are 0s, there is no latent structure and items are simply drawn from multivariate distribution centered around the item means with covariance matrix $\mathbf{\Theta}$. If all values in $\mathbf{\Lambda}$ 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 Missing data imposition

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

$$p(x_t = MISSING|X) = \mathbf{\Phi}(\tilde{X}\boldsymbol{\theta}) \tag{5}$$

with x_t a variable target of missing data imposition, $\mathbf{\Phi}(.)$ being the logistic cumulative distribution function, $\tilde{\boldsymbol{X}}$ 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 3 in the block of lowly correlated variables (x_t with $t = 1, ..3, 6, ..., 8$). Item non-response was imposed following equation (4.1) with 4 variables included in $\tilde{X}$: two fully observed variables from both the high and lowly correlated group of variables (x_r with $r = 4, 5, 9, 10$).

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 (4.1).

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, which is a required to avoid imputing under Missing Not At Random.

3.1.3 Analysis model(s)

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

Experiment 2 The same saturated model is chosen to be fitted to the data with a latent structure to obtain means, variances, and covariances on the observed items. Furthermore, we an oracle Confirmatory Factor Analysis can be fitted to this data to see how the factor scores are recovered after imputation.

3.1.4 Criteria

To compare the performance of the different imputation methods, several outcome measures were considered.

Bias First, we used Percent Relative Bias (PRB) and Standardized Bias (SB) to quantify the bias introduced by the imputation procedure.

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

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

where θ represents the reference value (the "true" value) of the focal parameter, $\bar{\hat{\theta}}$ represents its estimate averaged over the MCMC replications, $SD_{\hat{\theta}}$ represents the empirical standard deviation of θ, and $I(.)$ is the indicator function that returns 1 if the argument is true and 0 otherwise.

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

$$CIR = \frac{\sum_{n=1}^{R} I(\hat{\theta} \in \widehat{CI}_r)}{R} \tag{8}$$

where R is the total number of MCMC repetitions, $\hat{CI}_r$ is the confidence interval for the focal estimates in a given repetition.

CIR below .9 are considered problematic for 95% confidence intervals (Van Buuren, 2018, p. 52). Confidence interval coverage higher than .95 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.

Euclidean Distance Both bias and CIC are computed for individual parameters. When many parameters are involved in the analysis model these measures become cumbersome to compare. Hence, we have also decided to use multivariate measures to assess the degree to which estimates obtained after employing MI miss the target.

The Euclidean Distance between ...

3.1.5 Conditions

Data generation happened based on different conditions defined by design factors specific to each experiment. The procedure outline above was run for each of the conditions in table 1.

condition	n	p	l	pm	λ range
Experiment 1					
1	200	50	0	.1	-
2	200	500	0	.1	-
3	200	50	0	.3	-
4	200	500	0	.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

Experiment 1 Two experimental factors were considered: p, the number of columns in the dataset, which are all fed to the imputation algorithms, and the target proportion of per variable missing cases.

Experiment 2 The dimensionality of the data was controlled based on the number of latent variables. 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.

In experiment 2, we have also varied two other factors: (1) the proportion of missing values as in experiment 1, and (2) the strength of the latent structure by drawing the values of the factor loadings used to generate data from a range of low (between .5 and .6) or high values (between .9 and .97).

3.2 Results

3.2.1 Experiment 1

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

Focusing first on the means, all methods achieve a bias that is smaller than the 10% threshold 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 (rows 1 and 3), but it does lose ground in the higher dimensional conditions (2 and 4)

Moving to the variances, we notice that the single imputation method, missForest, leads to extreme negative bias (above 10%) in all conditions. We also notice that IURR and Blasso are giving some of the lowest biases across all conditions, even in the most difficult one, with high dimensionality and proportion of missing values (row 4). Somewhat surprisingly, directly using regularized regression within the imputation models (DURR) leads to a bias larger than 10% for the high dimensional condition with high proportion of missing values.

It is interesting to note that MI-PCA over estimates the variance of the variables with missing values. This means that using such method we are incorporating more uncertainty regarding the values to be imputed than we should. While not ideal, this is certainly a more conservative mistake to make compared to imputing values with too much certainty.

Finally, the third column in table 1 shows values of the estimation bias obtained for covariances between variables with missing values. As covariances depend on two variables, recovering the correct estimates after imputation is inherently more difficult than with means and variances. These explains the generally worse performances reported in the figure. We can see that MI-PCA is the only method that allows to recover the covariances with a bias smaller than 10 % in all conditions. Indirect Use of Regularized Regression (IURR) also performs well, and noticeably better than all other methods, but it seem to struggle with covariances in condition 4.

The two most promising methods, in terms of bias, show different weaknesses and strengths: MI-PCA struggles with recovering correctly variances but returns very lowly biased covariances in the high dimensional condition; IURR shows exactly the opposite behaviour.

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 most conditions. However, there are a few exceptions and peculiarities to note:

- 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;

- compared to all other methods, MI-PCA tends toward over- rather than under-cover, and it does so to a much more contained degree.

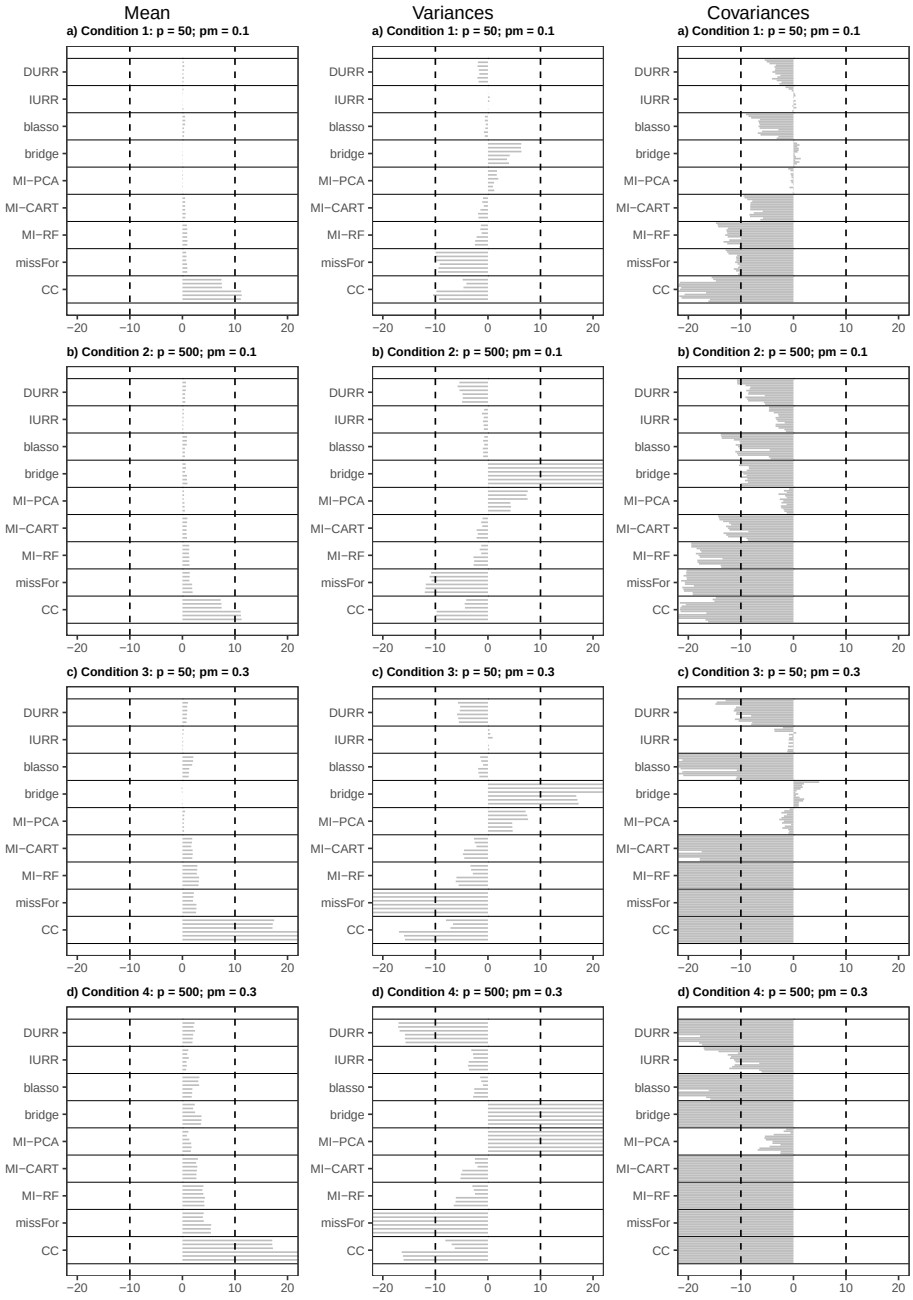

Figure 1: Percent Relative Bias (PRB) for the means, variances, and covariances broken down by method. Each row represents a different condition. Each column is dedicated to a different parameter type.

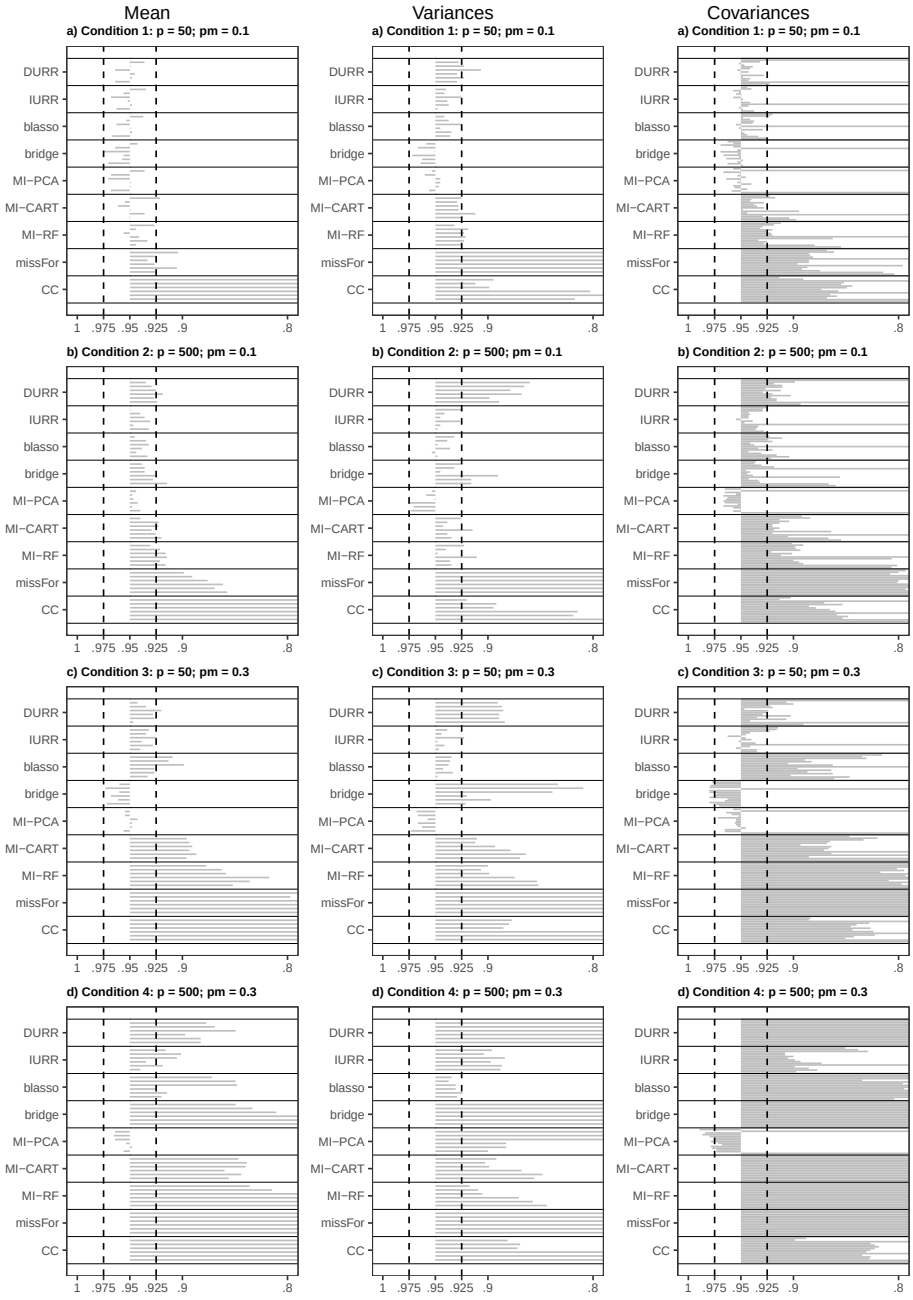

Figure 2: Confidence Interval Coverage (CIR) for the means, variances, and covariances broken down by method. Each row represents a different condition. Each column is dedicated to a different parameter type.

3.2.2 Experiment 2

Saturated Model Results form the first experiments held mostly constant in experiment 2. Figure 3 reports the bias of the Saturated Model parameters estimates for the first four conditions of experiment 2. For both variances and covariances, PRB is reported, while for the means we reported the SB. Items were generated around a mean of 0 making the computation of PRB for such parameter meaningless.

The least biased estimates for means and variances are obtained with IURR, in most 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 for condition 4

In agreement with what was found in experiment 1, the MI-PCA approach is the only one resulting in acceptable bias for all covariance estimates in all conditions.

Figure 4 shows results for the confidence interval coverage in experiment 2. When factor loadings are high, conditions 1 to 4, 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. For both means and variances confidence intervals coverage is within .925 and .975 for al methods. As the proportion of missing values increases we see a general deterioration in CIR performances, with IURR and MI-PCA still showing the most contained deviations from the target value. Furthermore, MI-PCA tends to include the true parameter values more than it should (over-coverage), while most other methods show sings of under-coverage.

The same pattern can be seen in the conditions with lower factor loadings (reported in appendix). However, the MI-PCA approach also leads to extreme under-coverage of variances (as do most other methods) in condition 8.

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 the true parameter value in all conditions in experiment 2.

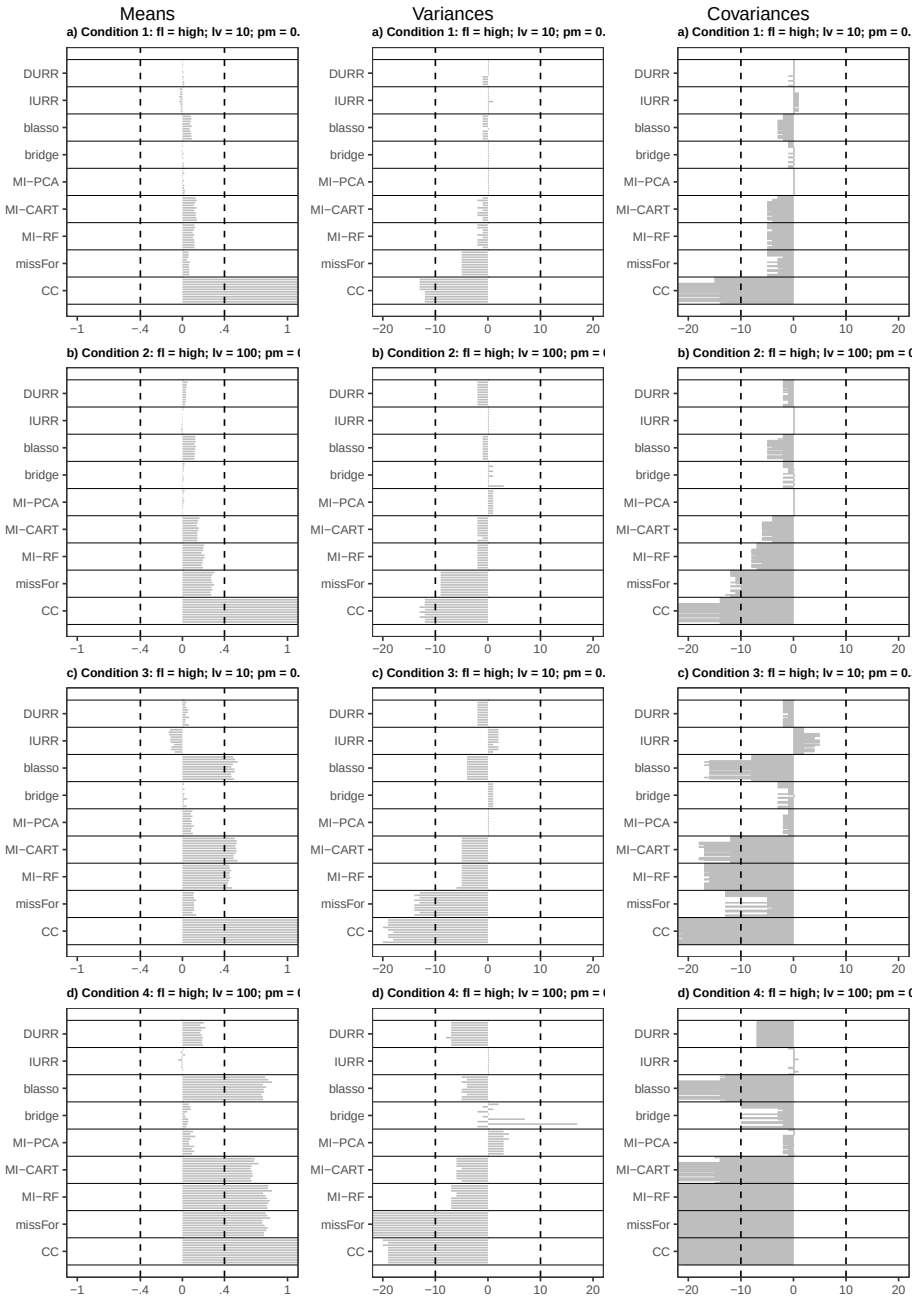

Figure 3: Bias estimation for the means (SB), variances and covariances (PRB) for condition 1 to 4. Each column is dedicated to a different parameter type.

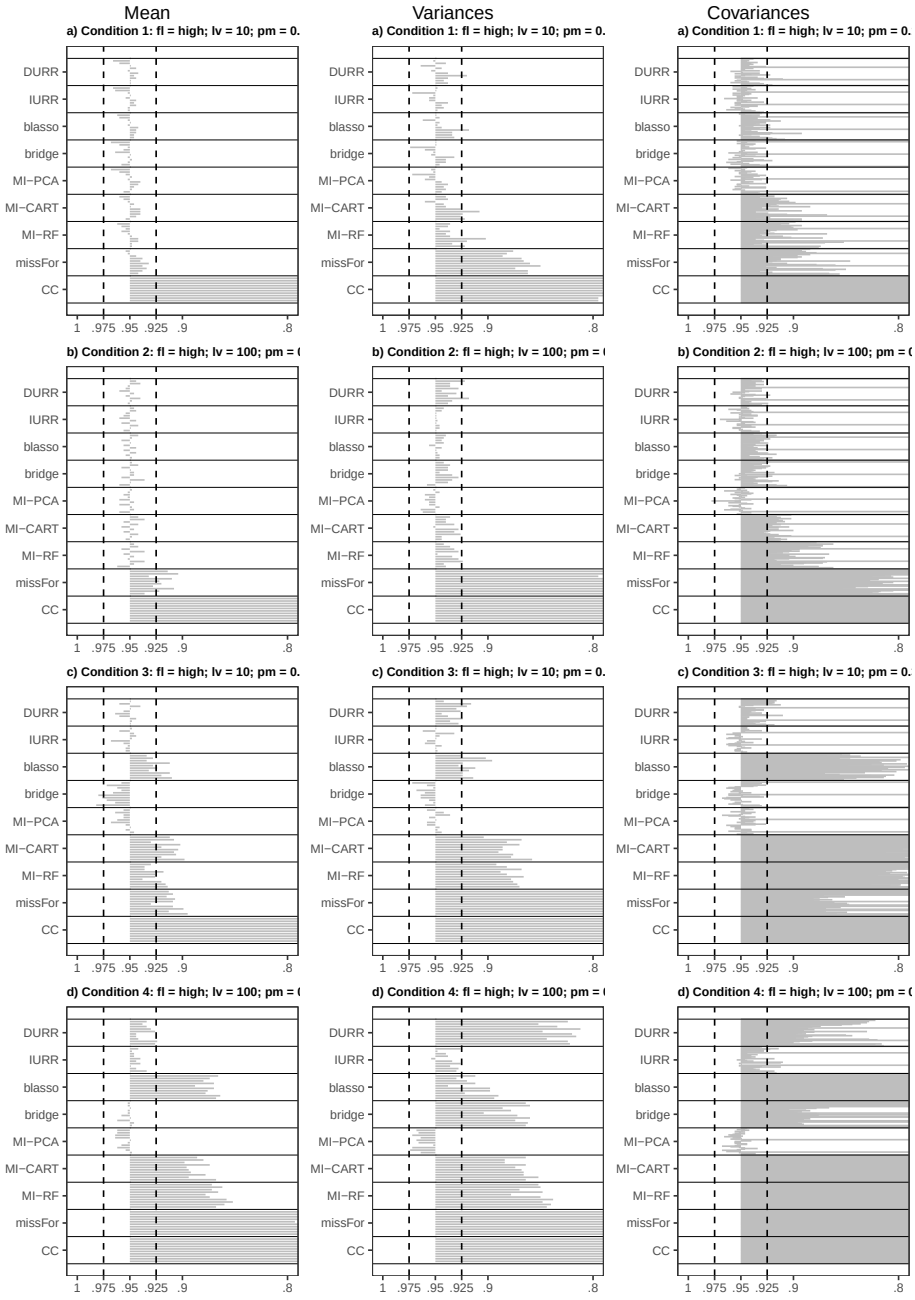

Figure 4: Confidence Interval Coverage (CIR) for the means, variances, and covariances for condition 1 to 4. Each column is dedicated to a different parameter type.

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

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.

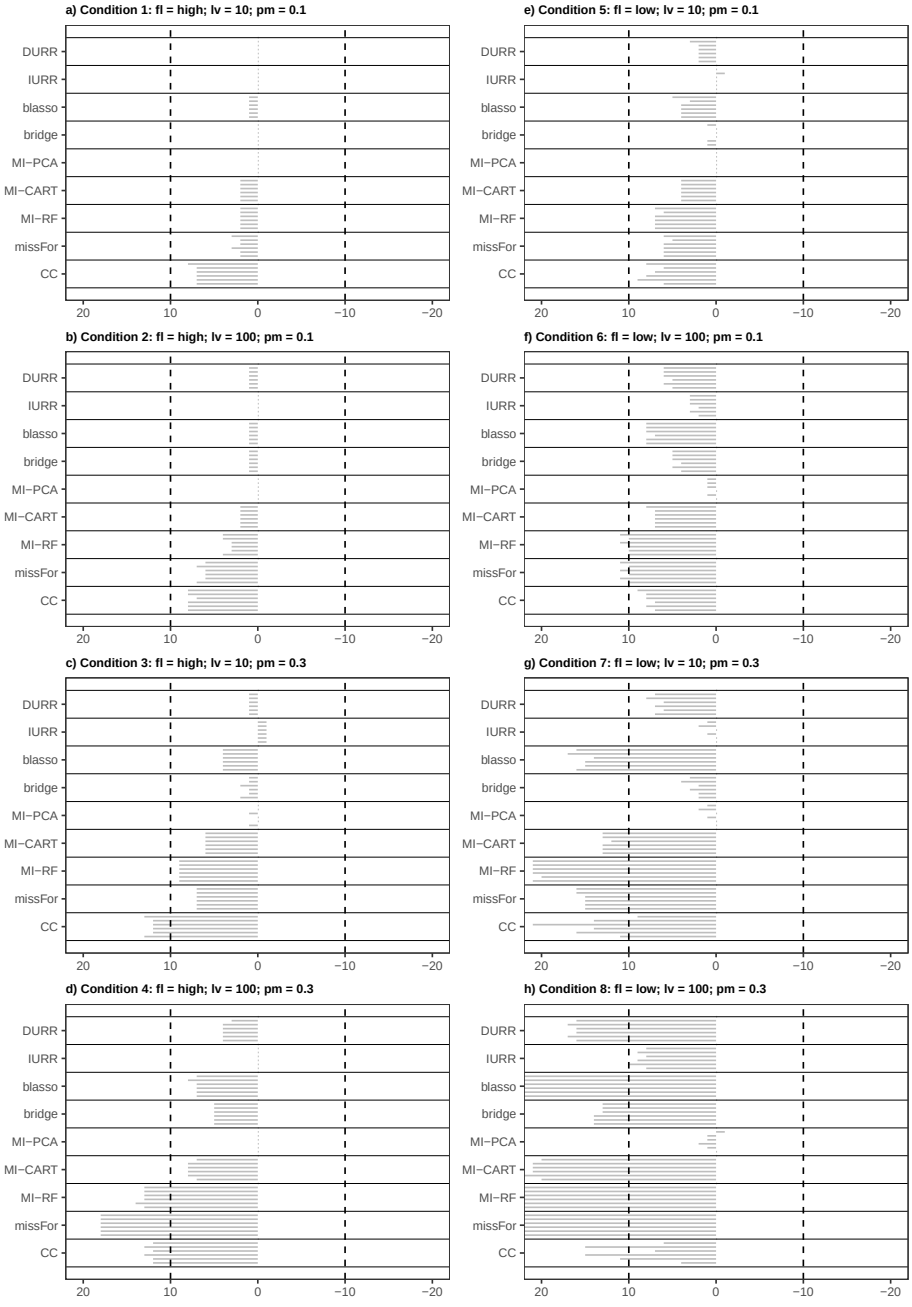

Figure 5: Percent Relative Bias (PRB) for the factor loadings.

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. By using data gathered for an actual survey, we can mostly observe whether the relative performances of the imputation methods change when they are deployed for real data research. Variables in the EVS data are not generated artificially from continuously normal distributions but are discrete numerical items that are treated as such by researchers.

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 1000 times ($R = 1000$):

- Data generation - A bootstrap sample $\boldsymbol{X}^*$ was generated by sampling with replacement n observations from a pre-processed EVS data-matrix. Part of the pre-processing step was some form of imputation used to obtain a pseudo-fully observed input data matrix.
- Missing data imposition - Missing values were imposed on a given number of target variables in $\boldsymbol{X}^*$, according to some response model.
- Imputations - Each method described in section 2 to deal with missing values was used to impute NAs.
- Analysis - Two analysis models were fitted to the differently treated data. Parameters estimates 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 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 3.1.4).

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 Methods

Data preparation EVS is a standardizes 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 linked to meta data (e.g. time of interview, mode of data collection). All missing values were filled in with a run of a single imputation predictive mean matching (PMM) which allowed us to obtain a pseudo fully-observed dataset. PMM was chosen for the task as it is an effective, flexible imputation method that maintains the distributional characteristics of the original data. The full cleaning process is more systematically described in the appendix.

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).

Analysis model(s) To define plausible analysis models we have searched the EVS database for articles using such data looking for suitable analysis models on which to test the effectiveness of the different imputation algorithm. We have defined two linear regression models of the form:

$$\boldsymbol{y} \sim \beta_0 + \boldsymbol{\beta_1 x}_1 + \boldsymbol{\beta}_{-1}\boldsymbol{X}_{-1} \tag{9}$$

In model 1, inspired by Köneke (2014), the dependent variable 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 a 4-point item measuring the self-reported importance of religion in one's life. A variety of covariates, such as measures of trust, education, and socio-economic status, were included in $\boldsymbol{X}_{-1}$ as control variables.

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

In model 2, inspired by Immerzeel et al. (2015), the dependent variable is an harmonized variable constructed by EVS to describe the respondents' tendency to vote on a 10-point left-to-right continuum; the predictor of interest is a composite mean scale measuring respondents attitudes toward immigrants and immigration ('nativist attitudes scale'). Respondents expressed how much they agreed on a scale from 1 to 10, with the following statements: '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 1, and some measures of political interest.

A research might fit this model to their data and look at the regression coefficient of the nativist attitudes scale to test some theory regarding its effect on voting for right wing parties.

Missing data imposition Missing data were imposed on 6 variables according to the same strategy as in 3.1.2. The target variables we identified were the two dependent variables in models 1 and 2, religiosity (predictor in both models) and the three items making up the nativist attitudes scale (predictor in the second model).

The response model form is the same as in and 3 variables were included in $\tilde{X}$: age, education, and an item measuring trust in new people. These are plausible variable that influence response tendencies in participants: older people usually have higher item non-response rates than younger; so do lower educated compared to higher educated people; people that have less trust in new people are assumed to withhold more information from the interviewer.

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 ($p = 250$), the dimensionality of the data is changed by defining different sizes for the sample taken from the pseudo-fully observed data. 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 of interest in model 1 and 2. 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 (a), the low dimensional one, but leads to extreme bias in the high dimensional condition (b); the latter provides, in all conditions, the highest PRB among the MI methods, it is consistently outperformed even by Complete Case analysis, and results in a 10% PRB for the regression coefficient of nativist attitudes in model 2.

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.

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. The advantage in using these methods over others is stark for model 2 while it is less marked in model 1. While in the low dimensional condition, IURR does not seem to provide a lower bias than the other methods, its relative performance improves as the dimensionality of the data increases.

MI-PCA seems to struggle with bias for model 1, ranking last among the multiple imputation models. Nevertheless, it does provide a stark advantage compared to mean imputation and Complete Case analysis.

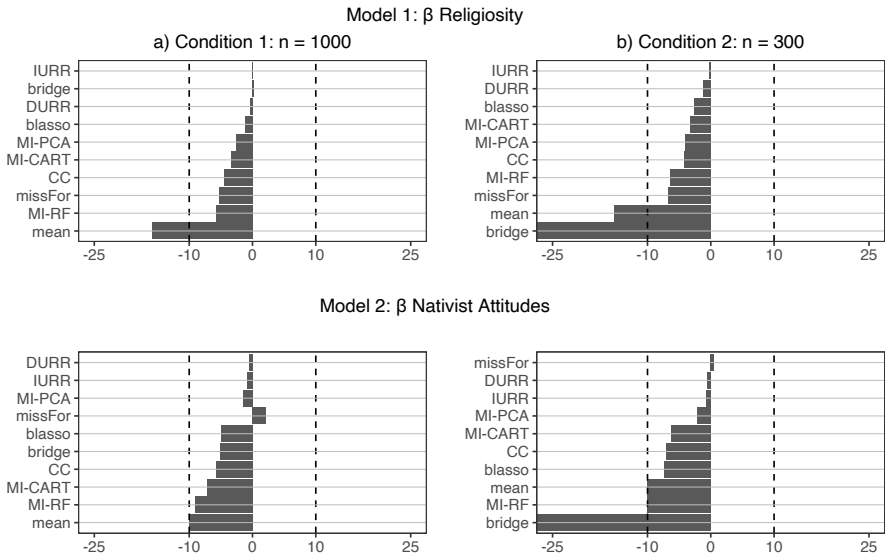

Figure 6: Bias for single parameter of interest in the two different models

Single imputation missForest is also able to provide a competitive vector of estimates, at least in model 2.

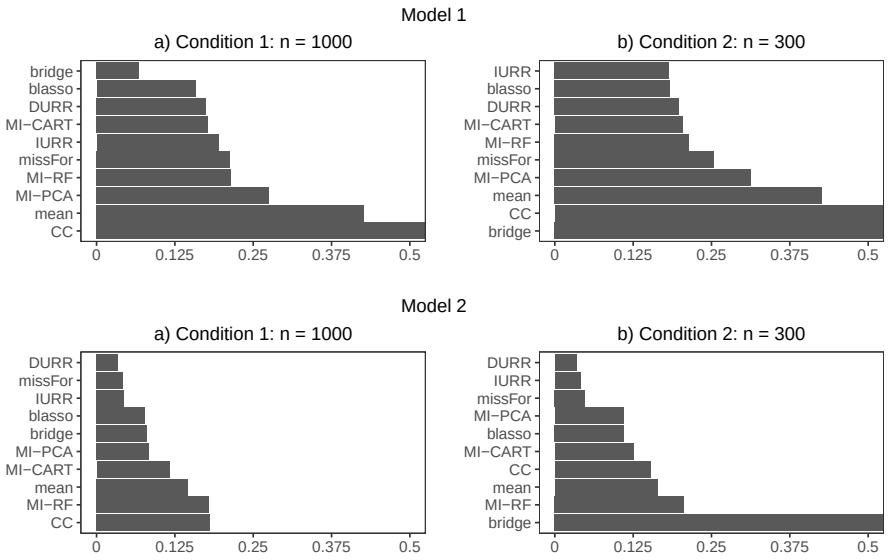

Figure 7: Euclidean distance between the vector of estimated regression coefficients and the vector of their true value

4.2.2 Confidence Interval Coverage

Confidence Interval Coverage Researchers interested in testing their theories with the inferential models described, will be interested in the confidence interval coverage of the estimates of interest. Figure 8 reports the CIR for the focal regression coefficient in the two models.

Both IURR and DURR remain fairly competitive in terms of CIR, but the advantage they showed in terms of bias is not carried over to this criterion. Both MI-PCA and Blasso outperform IURR and DURR in almost all conditions, granting confidence intervals that are noticeably closer to nominal coverage.

Euclidean Distance However, when looking at a more general pattern reported in 9, IURR and DURR return to the top of the leader-board, providing the closest vectors of Confidence Interval Coverages for model parameters to nominal levels.

Confidence Interval Width Plot?

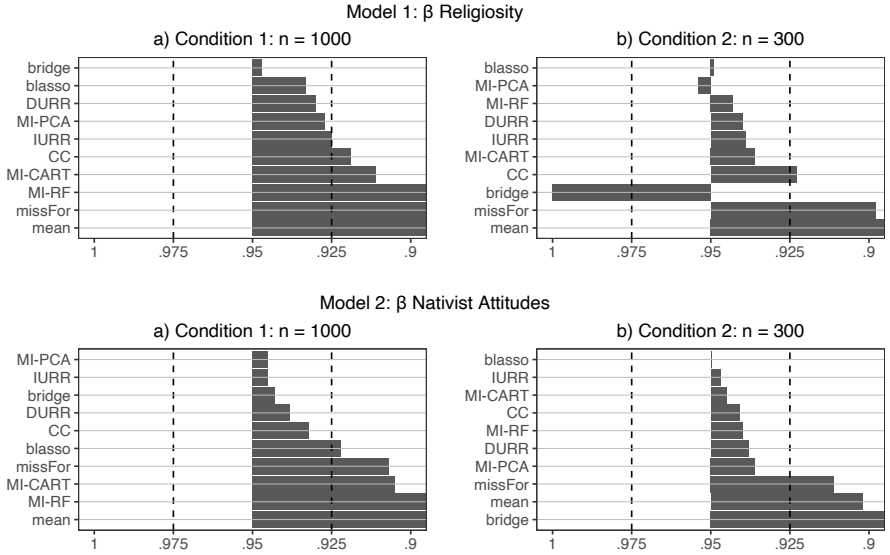

Figure 8: Confidence Interval Coverage for single parameter of interest in the two different models

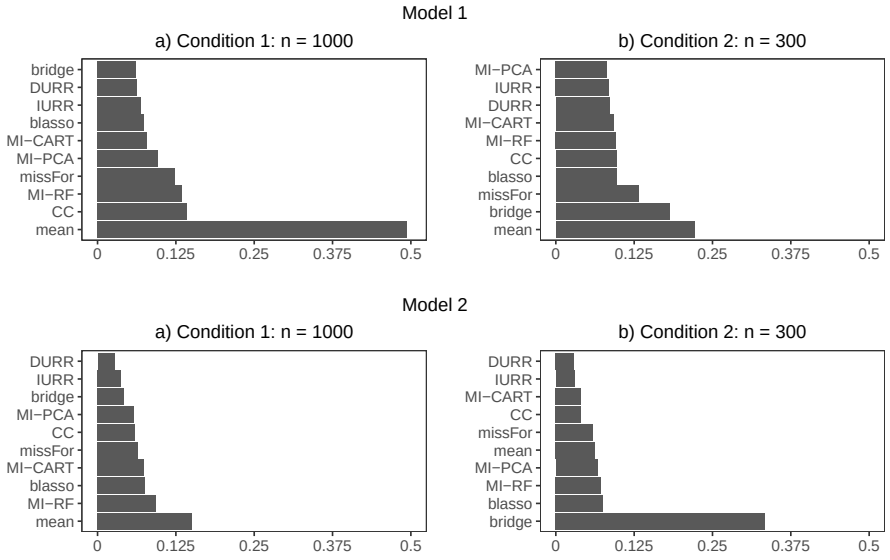

Figure 9: Euclidean distance between the vector of confidence coverages and the vector of nominal coverage

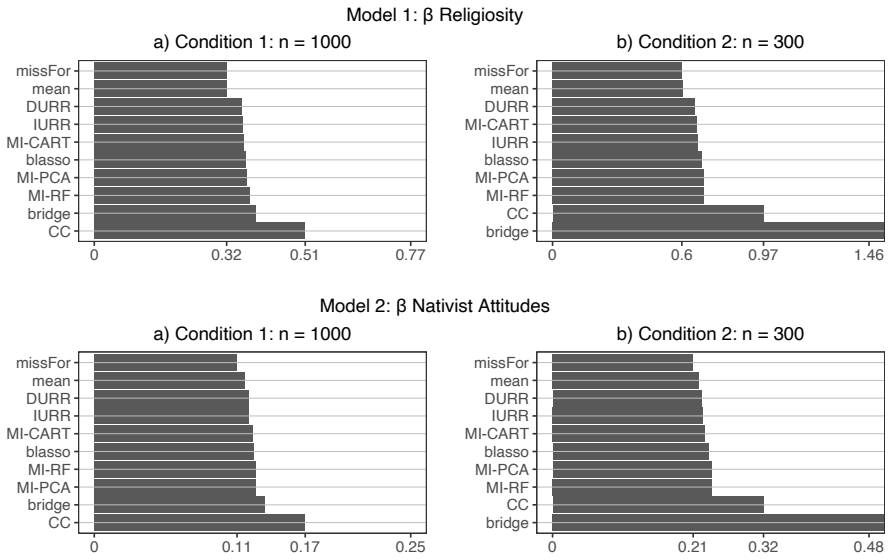

Figure 10: Confidence Interval Width for given parameter

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, the IURR and DURR are not as time-intensive, but still require more then ten times the time of MI-PCA and blasso imputation.

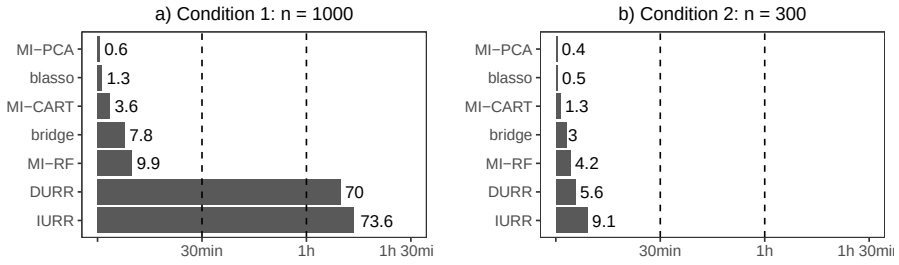

Figure 11: Average imputation time for each method

5 Discussion

Make parallels and comparisons between methods. Try to portray the general pattern that comes out of the combined results from the three studies.

6 Conclusions

Take-home message Give the take-home message in one or two paragraphs

Limitations and future directions Describe limitation with specific focus on what are your planned next steps in this line of research.

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.

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.

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.

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, volume 72. Chapman & Hall/CRC, Boca Raton, FL.

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.