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

Edoardo Costantini[a] , Kyle M. Lang[a], Tim Reeskens[b], and Klaas Sijtsma[a]

[a]Tilburg University, Department of Methodology and Statistics; [b]Tilburg University, Department of Sociology

ARTICLE HISTORY

Compiled May 11, 2021

CONTACT Edoardo Costantini. Email: e.costantini@tilburguniversity.edu

R Code for the project can be found at the main author's GitHub page: `https://github.com/EdoardoCostantini/imputeHD-comp`

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

ABSTRACT

Including a large number of predictors in the imputation model underlying a Multiple Imputation (MI) procedure is one of the most challenging tasks imputers face. A variety of high-dimensional MI techniques (HD-MI) can facilitate this task, but there has been limited research on their relative performance. In this study, we investigate a wide range of extant HD-MI techniques that can handle both large numbers of predictors in the imputation model and general missing data patterns. We assess the relative performance of seven HD-MI methods with two Monte Carlo simulation studies and a resampling study based on real survey data. The performance of the methods is defined by the degree to which they facilitate unbiased and confidence-valid estimates of the parameters of complete data analysis models. We find that using regularized regression to select the predictors used in the MI model, and using principal components analysis to reduce the dimensionality of auxiliary data produce the best results.

KEYWORDS

Keywords: Multiple Imputation; High-dimensional; Regularized Regression; Principal Components; CART; random forests

1. Introduction

Today's social, behavioral, and medical scientists have access to large multidimensional data sets that can be used to investigate the complex roles that social, psychological and biological factors play in shaping individual and societal outcomes. Large social scientific data sets—such as the World Values Survey and the European Values Study (EVS)—are easily accessible to researchers, but making use of the full potential of these data requires dealing with the crucial problem of multivariate missing data.

Rubin's Multiple Imputation (MI) approach (Rubin, 1987) was developed to address missing responses in surveys. An MI-based analysis is a three-step process that entails an imputation phase, an analysis phase, and a pooling phase. The fundamental idea of the imputation phase is to replace each missing data point with m plausible values sampled from the posterior predictive distribution of the missing data given

the observed data. This procedure generates m completed versions of the original data that are analyzed separately during the analysis phase, using standard complete data analysis models. Finally, in the pooling phase, the m estimates of any parameter of interest are pooled following Rubin's rules (Rubin, 1987) to create the MI parameter estimate.

Since Rubin's seminal work, two main strategies have become popular for multiple imputation of multivariate missing data: joint modelling (JM; Schafer, 1997, ch. 4) and full conditional specification (FCS), also known as Multiple Imputation by Chained Equations (MICE; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). JM relies on defining a joint distribution for the missing data, deriving conditional distributions for each missing data pattern, and obtaining samples from these distributions by means of a Markov Chain Monte Carlo algorithm. FCS defines conditional distributions for each incomplete variable and performs iterative imputations on a variable-by-variable basis. Compared to the JM approach, FCS can more easily accommodate different distributions of the incomplete variables. It is also easier to preserve unique features in the data, such as interactions between variables or skip patterns in questionnaires when using FCS.

Both the JM and FCS rely on the crucial *missing at random* (MAR) assumption. Meeting this assumption requires specifying imputation models that include all observed correlates of the missingness. Omitting a variable that relates to both the missingness and the incomplete variables induces *missing not at random* (MNAR) data. MI under MNAR can lead to substantial bias in the MI parameter estimates, and invalidates hypothesis testing involving the imputed variables. As a result, when choosing which auxiliary variables (i.e., covariates included purely to support the missing data analysis) to include in the imputation model, an inclusive strategy (i.e., including many auxiliary variables) is generally preferred to restrictive approach (i.e., including few or no auxiliary variables). An inclusive approach reduces the chances of omitting important correlates of missingness, thereby making the MAR assumption more plausible. Furthermore, the inclusive strategy has been shown to reduce estimation bias and increase efficiency (Collins, Schafer, & Kam, 2001) and to reduce the chances of specifying uncongenial imputation and analysis models (Meng, 1994).

Specifying the imputation models for a FCS MI procedure remains one of the most challenging steps in dealing with missing values for large multidimensional data sets. In practice, the inclusive strategy faces identification and computational limitations. One serious risk of an inclusive strategy is the occurrence of singular covariance matrices within the imputation algorithm. When the predictor matrix is high-dimensional (i.e., the number of recorded units, n, is not substantially larger than the number of recorded variables, p) or afflicted by high collinearity (i.e., some variables are linear combinations of other variables) the covariance matrix of the predictors will be singular. Singular matrices cannot be inverted, which is a fundamental part of estimating the imputation model in most parametric imputation procedures. As a result, the possible high dimensionality of the predictor matrix resulting from an inclusive strategy can prevent a straightforward application of MI or force researchers to make arbitrary choices regarding which variables to include in the imputation model.

1.1. Prior work on high-dimensional imputation

Recent developments in high-dimensional MI techniques represent interesting opportunities to embrace an inclusive strategy, without facing its downsides. High-dimensional Single Imputation methods have been developed in an effort to improve the accuracy of the individual imputations (e.g., D'Ambrosio, Aria, & Siciliano, 2012; Kim, Golub, & Park, 2005; Stekhoven & Bühlmann, 2011). However, the main task of social scientists is to make inference about a population based on a sample, and Single Imputation (SI) is inadequate for this purpose because it does not support statistically valid inference (Rubin, 1996). The relevant conceptualization of statistical validity was defined by Rubin (1996) as capturing two features of estimation. First, the point estimate of a parameter of interest must be unbiased, and second, the actual confidence interval (CI) coverage of the true parameter value must be equal or greater than nominal the coverage rate. SI strategies might meet the first requirement but cannot meet the second because they do not account for uncertainty regarding the imputation model parameters. MI, on the other hand, was designed to provide statistically valid inference and, therefore, is more suitable for social scientific research.

The combination of MI with high-dimensional prediction models has been directly

tackled by algorithms that combine FCS with shrinkage methods (Deng, Chang, Ido, & Long, 2016; Zhao & Long, 2016). Other researchers have proposed the use of dimensionality reduction to avoid the obstacles of an inclusive strategy. However, these solutions were either limited to the JM approach (Song & Belin, 2004), or tested exclusively in low-dimensional settings (Howard, Rhemtulla, & Little, 2015). Finally, tree-based FCS strategies also have the potential to overcome the limitations of inclusive strategies. The nonparametric nature of decision trees bypasses the identification issues most parametric methods face in high-dimensional contexts.

1.2. Scope of the current project

Including shrinkage methods, Principle Component Analysis (PCA), and decision trees within the FCS framework has the potential to simplify the decisions social scientists need to make when dealing with missing values. However, the lack of comparative research on the performance of these methods makes it difficult for social scientists working with large data sets to decide which imputation method to adopt. In this article, we provide a comparison of the state-of-the-art in high-dimensional imputation algorithms. We compared seven promising imputation methods in terms of their ability to support statistically valid analyses. These comparisons were based on three numerical experiments: two Monte Carlo simulation studies and a resampling study using real survey data.

In what follows, we first introduce the missing data treatments that we compared in our study. Then we present the methodology and results of the three numerical experiments. We discuss the implications of the results and provide recommendations for applied researchers. We conclude by describing the limitations of the study and recommending future research directions.

2. Imputation methods and Algorithms

We will use the following notation: scalars, vectors, and matrices are denoted by italic lowercase, bold lowercase, and bold uppercase letters, respectively. A scalar belonging to an interval is indicated by $s_1 \in [s_2, s_3]$, while a scalar taking the values in a set is

represented as $s_1 \in \{s_2, s_3\}$.

Consider an $n \times p$ data set, $\boldsymbol{Z}$, comprising variables $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, ..., $\boldsymbol{z}_p$. Assume that the first t variables of $\boldsymbol{Z}$ have missing values, and that these t variables are the targets of imputation. Denote the columns of $\boldsymbol{Z}$ containing $\boldsymbol{z}_1$ to $\boldsymbol{z}_t$ as the $n \times t$ matrix, $\boldsymbol{T}$. The remaining $(p-t)$ columns of $\boldsymbol{Z}$ contains variables that are not targets of imputation. These variables constitute a pool of possible *auxiliary* variables that could be used to improve the imputation procedure. Let $\boldsymbol{A}$ denote this set of auxiliary variables so that $\boldsymbol{Z} = (\boldsymbol{T}, \boldsymbol{A})$. For a given $\boldsymbol{z}_j$, with $j = (1, ..., p)$, denote its observed and missing components as $\boldsymbol{z}_{j,obs}$ and $\boldsymbol{z}_{j,mis}$, respectively. Let $\boldsymbol{Z}_{-j} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_{j-1}, \boldsymbol{z}_{j+1}, ..., \boldsymbol{z}_p)$ be the collection of $p-1$ variables in $\boldsymbol{Z}$ excluding $\boldsymbol{z}_j$. Denote by $\boldsymbol{Z}_{-j,obs}$ and $\boldsymbol{Z}_{-j,mis}$ the components of $\boldsymbol{Z}_{-j}$ corresponding to the data units in $\boldsymbol{z}_{j,obs}$ and $\boldsymbol{z}_{j,mis}$, respectively.

2.1. Multiple imputation by chained equations

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

More precisely, the MICE algorithm takes the form of a Gibbs sampler. At the mth iteration $(m = 1, ..., M)$, samples are drawn for the jth target variable $(j = 1, ..., t)$ from the following distributions:

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

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

where $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $\boldsymbol{z}_{j,mis}^{(m)}$ are draws from the parameter's full conditional posterior distribution (1) and the missing data posterior predictive distribution (2), respectively. After convergence, d sets of values are sampled from (2) and used as imputations. Any substantive model can then be fit to each of the d completed data sets, and the estimates can be pooled using Rubin's rules (Rubin, 1987).

Generally speaking, for each target of imputation, $\boldsymbol{z}_j$, the researcher needs to define a set of variables that will be included in $\boldsymbol{Z}_{-j}^{(m)}$. The high-dimensional imputation methods compared in this paper all follow the general MICE framework, but they differ in the elementary imputation methods they use to define equations (1) and (2). Each method has a different way of processing the auxiliary variables provided to the imputation algorithm, but all of them are designed to support an inclusive strategy while avoiding its usual obstacles.

2.1.1. MICE with a fixed ridge penalty

MICE with a fixed ridge penalty (bridge) uses the Bayesian normal linear model described by Van Buuren (2018, p. 68, algorithm 3.1) as its elementary imputation method (i.e., the model used to define equations (1) and (2)). In this approach, the sampling of each $\hat{\boldsymbol{\theta}}_j^{(m)}$ in equation (1) relies on inverting the cross-products matrix of $\boldsymbol{Z}_{j,obs}^{(m)}$. Adding a biasing ridge penalty, κ, to the diagonal of this cross-products matrix circumvents singularity, and allows sampling even when $\boldsymbol{Z}_{j,obs}^{(m)}$ is afflicted by collinearity or when n is not substantially larger than p.

The value of κ is usually chosen to be close to zero (e.g. $\kappa = 0.0001$), because values larger than 0.1 may introduce systematic bias (Van Buuren, 2018, p. 68). However, larger values may be necessary to invert the cross-products matrix in certain scenarios. In the present work, we choose the value of κ by means of cross-validation.

2.1.2. MICE with Bayesian lasso

Zhao and Long (2016) proposed the MICE with Bayesian lasso imputation algorithm (blasso), an MI procedure that uses the Bayesian lasso as its elementary imputation method. Bayesian lasso is a regular Bayesian multiple regression model with priors on the slope coefficients that allow interpreting the mode of the slopes' posterior distribution as lasso estimates (Hans, 2009; Park & Casella, 2008). Given data with a sample size n, consider a dependent variable $\boldsymbol{y}$, and a set of predictors $\boldsymbol{X}$. The Bayesian lasso specification that we used for the blasso imputation algorithm is that

specified by Hans (2010b):

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

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

$$\sigma^2 \sim \text{Inverse-Gamma}(a, b) \tag{5}$$

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

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

Equation (3) represents the density function of a multivariate normal random variable with mean $\boldsymbol{X}\boldsymbol{\beta}$ and covariance matrix $\sigma^2\boldsymbol{I}_n$ evaluated at $\boldsymbol{y}$. Equation (4) represents the expansion of the Park and Casella (2008) double exponential prior developed by Hans (2010b) to accommodate the uncertainty regarding model sparsity. Finally, equations (5) to (7) represent hyper priors for the residual variance, σ^2, the penalty parameter, τ, and the sparsity parameter, ρ. Our implementation of blasso imputation replaces equation (1) with the blasso model defined by equations (3) to (7) with $\boldsymbol{y} = \boldsymbol{z}_{j,obs}$ and $\boldsymbol{X} = \boldsymbol{Z}_{-j,obs}$.

The R code used to perform the blasso imputation was based on the R Package *blasso* (Hans, 2010a) and can be found on the main author's GitHub page. For a detailed description of the Bayesian lasso MI algorithm in a univariate missing data context see Zhao and Long (2016).

2.1.3. Direct use of regularized regression

As proposed by Zhao and Long (2016) and Deng et al. (2016), frequentist regularized regression can be directly used in a MICE algorithm. At iteration m, for a target variable $\boldsymbol{z}_j$, the Direct Use of Regularized Regression imputation algorithm (DURR) replaces equations (1) and (2) with the following two steps:

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

from equation (1).

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

2.1.4. Indirect use of regularized regression

While DURR simultaneously performs model regularization and parameter estimation in equation (1), the Indirect Use of Regularized Regression imputation algorithm (IURR) uses regularized regression exclusively for model trimming. The selected predictors are then used in a standard MI procedure (Deng et al., 2016; Zhao & Long, 2016). At iteration m, the IURR algorithm performs the following steps for each target variable, $\boldsymbol{z}_j$:

(1) Fit a linear regression model using a regularized method that does variable selection (e.g., lasso). Take $\boldsymbol{z}_{j,obs}$ as the dependent variable and $\boldsymbol{Z}_{-j,obs}^{(m)}$ as the predictors (unlike DURR, IURR uses the original data, not a bootstrap sample). The regression coefficients that are *not* shrunk to 0 define the active set of variables that will be used as predictors in the actual imputation model.

(2) Obtain the maximum likelihood estimates of the regression coefficients and the error variance from the linear regression of $\boldsymbol{z}_{j,obs}$ onto the active set of predictors defined in step 1. Then, sample new values of these parameters from a multivariate normal distribution parameterized by the MLEs[1]:

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

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

(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, $(\hat{\boldsymbol{\theta}}_j^{(m)}, \hat{\sigma}_j^{(m)})$.

2.1.5. MICE with PCA

By extracting principal components (PCs) from the auxiliary variables, it is possible to summarize the information contained in these auxiliaries with just a few components.

[1]The sampling notation is the same used by Deng et al. (2016).

These PCs can then be used as predictors in a standard, low-dimensional application of MICE. The MICE with PCA (MI-PCA) procedure can be summarized as follows:

(1) Extract the first PCs that cumulatively explain at most 50% of the variance in the auxiliary variables, $\boldsymbol{A}$, and collect them in a new matrix, $\boldsymbol{A}'$;
(2) Replace the auxiliary variables, $\boldsymbol{A}$, in $\boldsymbol{Z}$ with $\boldsymbol{A}'$ to obtain $\boldsymbol{Z}' = (\boldsymbol{T}, \boldsymbol{A}')$;
(3) Use the standard MICE algorithm with the Bayesian normal linear model (Van Buuren, 2018, p. 68, algorithm 3.1) as elementary imputation method to obtain multiply imputed datasets from $\boldsymbol{Z}'$.

If the auxiliary variables are incomplete, their missing values can be imputed with stochastic SI in a pre-processing step. Extracting PCs does not require the estimation of standard errors, so MI is unnecessary and SI suffices. The MI-PCA method was inspired by Howard et al. (2015) and the *PcAux* R package (Lang, Little, & PcAux Development Team, 2018) that implements and developed its ideas.

2.1.6. MICE with classification and regression trees

MICE with classification and regression trees (MI-CART; Burgette & Reiter, 2010) is a MICE algorithm that uses classification and regression trees (CART) as the elementary imputation method. Given an outcome variable $\boldsymbol{y}$ and a set of predictors $\boldsymbol{X}$, CART is a nonparametric recursive partitioning technique that models the relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$ by sequentially splitting observations into subsets of units with relatively homogeneous $\boldsymbol{y}$ values. At every splitting stage, the CART algorithm searches through all variables in $\boldsymbol{X}$ to find the best binary partitioning rule to predict $\boldsymbol{y}$. The resulting collection of binary splits can be visually represented by a decision tree structure where each terminal node (or *leaf*) represents the conditional distribution of $\boldsymbol{y}$ for units that satisfy the splitting rules.

For each $\boldsymbol{z}_j$, the mth iteration of MI-CART proceeds as follows:

(1) Train a CART model to predict $\boldsymbol{z}_{j,obs}$ from the corresponding $\boldsymbol{Z}^{(m)}_{-j,obs}$.
(2) Assign each element of $\boldsymbol{z}_{j,mis}$ to a terminal node by applying the splitting rules from the fitted CART model to $\boldsymbol{Z}_{-j,mis}$.
(3) Create imputations for each element of $\boldsymbol{z}_{j,mis}$ by sampling from the pool of $\boldsymbol{z}_{j,obs}$

in the terminal node containing $\boldsymbol{z}_{j,mis}$. This procedure corresponds to sampling from the missing data posterior predictive distribution in Equation (2).

This approach does not consider uncertainty in the imputation model parameters since the tree structure is not perturbed between iterations. Therefore, MI-CART cannot produce proper imputations in the sense of Rubin (1986). The implementation of MI-CART used in this paper corresponds to the one presented by Doove, Van Buuren, and Dusseldorp (2014, p. 95, algorithm 1) and the *impute.mice.cart()* function from the *mice* package.

2.1.7. MICE with random forests

MICE with random forests (MI-RF) is a MICE algorithm that uses random forests as the elementary imputation method. The random forest algorithm entails fitting many decision trees (e.g., CART models) to subsamples of the original data. These subsamples are derived by resampling rows with replacement and sampling subsets of columns without replacement. The random forest algorithm results in an ensemble of fitted decision trees that generate a sample of predictions for each outcome value.

For each $\boldsymbol{z}_j$, the mth iteration of MI-RF proceeds as follows:

(1) Generate k bootstrap samples from $\boldsymbol{Z}_{-j,obs}$.
(2) Use these bootstrap samples to fit k single trees predicting $\boldsymbol{z}_{j,obs}$ from a random subset of the variables in $\boldsymbol{Z}_{-j,obs}$.
(3) Generate a pool k terminal nodes for each element of $\boldsymbol{z}_{j,mis}$ by applying the splitting rules from each of the k fitted trees to the appropriate columns of $\boldsymbol{Z}_{-j,mis}$.
(4) Create imputations for each element of $\boldsymbol{z}_{j,mis}$ by sampling from the $\boldsymbol{z}_{j,obs}$ contained in the pool of terminal nodes defined above.

Bootstrapping and random input selection introduce uncertainty regarding the imputation model parameters (i.e., the tree sructure), as required by a proper MI procedure. For more details on the MI-RF algorithm, see Doove et al. (2014, algorithm A.1, p. 103). The programming of our implementation of the algorithm was heavily inspired by the *impute.mice.rf()* function in the R package *mice*.

2.1.8. MICE optimal model

When dealing with a large set of possible predictors for the imputation model, a common recommendation in the MI literature is to decide which predictors to include by following three criteria (Van Buuren, 2018, p. 168):

(1) include all variables that are part of the analysis models;
(2) include all variables that are related to the nonresponse;
(3) include all variables that are correlated with the targets of imputation.

In practice, researchers can never be sure that the second requirement is entirely met, as there is no way to know exactly which variables are responsible for missingness. However, with simulated data, we know which variables are involved in the missing data mechanisms. MICE optimal model (MI-OP) is an ideal specification of the MICE algorithm that uses this knowledge to include only the relevant predictors in the imputation models. The imputations were generated using the Bayesian normal linear model as the elementary imputation method.

2.2. Non-MI strategies

2.2.1. missForest

Most research on high-dimensional imputation has focused on applications for genomics data, where the goal is to prepare large data sets (e.g., DNA microarray data) for high-dimensional prediction algorithms, rather than inferential analysis. For this reason, a variety of SI methods based on machine learning algorithms have been proposed and compared (e.g., de Andrade Silva & Hruschka, 2009; Stekhoven & Bühlmann, 2011).

In this study, we consider the missForest (missFor) imputation method proposed by Stekhoven and Bühlmann (2011), which is a popular nonparametric imputation approach that can accommodate for a large number of predictors, can handle missing variables of the mixed data type, and has been robustly implemented in a popular R-package (Stekhoven, 2013). The approach consists of an iterative imputation that first trains a random forest on observed values, and then uses the trained random forest to

impute the missing values by averaging the predictions from its different trees. As a single imputation method we do not expect it to perform well for inferential tasks, at least compared to the MI methods discussed above.

2.2.2. Complete case analysis

By default, most data analysis software either fails in the presence of missing values or defaults to listwise deletion wherein only complete cases are used for the analysis (pandas development team, 2020; R Core Team, 2020). As the default behavior of most analysis tools, complete cases analysis (CC) remains a popular missing data treatment in the social sciences, despite its well-known flaws (Rubin, 1987, p. 8; Van Buuren, 2018, p. 9, Baraldi and Enders, 2010). Therefore, we include CC as a negative reference point.

2.2.3. Gold standard

Finally, we fit the analysis models directly to the fully observed data before imposing any missing data. In the following, we refer to the results obtained in this fashion as the gold standard (GS). These results represent the counterfactual analysis that would have been performed if there had been no missing data.

3. Experiment 1: simulated data from multivariate normal distribution

In the first simulation experiment, we focused on an ideal setting where data came from a known multivariate normal distribution and imputation was required to estimate the means, variances, and covariances of six items with missing values. We investigated the relative performance of the methods described above across a set of conditions defined by two experimental factors: the number of columns in the dataset, p, taking values 50 or 500; and the proportion of missing cases on each variable, pm, taking values 0.1 or 0.3. Table 1 summarizes the four resulting crossed conditions. Data with sample size $n = 200$ were independently generated $S = 1,000$ times for each condition. For each sth replicate, missing values were imposed on six target items, and then all missing data treatment methods described above were used to obtain estimates of the

item means, variances, and covariances.

[Table 1 about here.]

3.1. Simulation study procedure

3.1.1. Data generation

At every replication, a data matrix $\boldsymbol{Z}_{n\times p}$ was generated according to a multivariate normal model centered around a vector of fives and covariance matrix $\boldsymbol{\Sigma}_0$. The diagonal elements of $\boldsymbol{\Sigma}_0$ (variances) were equal to 1, and the off-diagonal elements were used to define three blocks of variables. The first five variables were highly correlated among themselves ($\rho = 0.6$); variables 6 to 10 were weakly correlated with variables in block 1 and among themselves ($\rho = 0.3$), and all the remaining $p - 10$ variables were uncorrelated.

3.1.2. Missing data imposition

Missing values were imposed on six items in $\boldsymbol{Z}$: three variables in the block of highly correlated variables ($\boldsymbol{z}_j$ with $j = 1, 2, 3$), and three in the block of lowly correlated variables ($\boldsymbol{z}_j$ with $j = 6, 7, 8$). Item nonresponse was imposed by sampling from a Bernoulli distribution with individual probabilities of nonresponse defined by

$$p_{miss} = p(z_{i,j} = miss|\tilde{\boldsymbol{Z}}) = \frac{exp(\gamma_0 + \tilde{\boldsymbol{z}}_i\boldsymbol{\gamma})}{1 + exp(\gamma_0 + \tilde{\boldsymbol{z}}_i\boldsymbol{\gamma})} \tag{9}$$

where $z_{i,j}$ is the ith subject's response on the jth target of missing data imposition, $\tilde{\boldsymbol{z}}_i$ is a vector of responses to the set of missing data predictors for the ith individual, γ_0 is an intercept parameter, and $\boldsymbol{\gamma}$ is a vector of slope parameters. $\tilde{\boldsymbol{Z}}$ was specified to include two fully observed variables from the strongly correlated set and two from the weakly correlated set ($\boldsymbol{z}_r$ with $r = 4, 5, 9, 10$). The probability of nonresponse to a variable never depended on the variable itself. Therefore, by using all columns of $\boldsymbol{Z}_{-j,obs}$ as predictors in the MI procedures, the elements of $\tilde{\boldsymbol{Z}}$ act as predictors in the imputation models, and the MAR assumption is satisfied. All slopes in $\boldsymbol{\gamma}$ were fixed to 1, while the value of γ_0 was chosen with an optimization algorithm that minimized

the difference between the actual and desired proportion of missing values.

3.1.3. Imputation

Missing values were treated with all methods described in Section 2. To evaluate convergence of the imputation models, we ran ten replications of the high-dim-high-pm condition and checked trace plots of the means of the imputed values. These checks suggested that the imputation algorithms converged within 50 iterations. The only exception was blasso, which required approximately 2,000 iterations for convergence.

The ridge penalty used in the bridge algorithm was fixed across iterations. We chose the value of the penalty term by means of a cross-validation procedure, wherein penalty values of $10^{-1}, 10^{-2}, ..., 10^{-8}$ were used to impute data with bridge. We then selected the value that resulted in the smallest average fraction of missing information (FMI; Rubin, 1987, eq. 3.1.10) across the analysis model parameters. Both IURR and DURR could have been implemented with a variety of penalties (e.g., lasso, Tibshirani, 1996; elastic net, Zou & Hastie, 2005; adaptive lasso, Zou, 2006). In this study, we used lasso as it is computationally efficient, and it performed well for imputation in Zhao and Long (2016) and Deng et al. (2016). A 10-fold cross-validation procedure was used at every iteration of DURR and IURR to choose the penalty parameter.

To maintain consistency with previous research, we specified the blasso hyperparameters in equations (5), (6), and (7) as in Zhao and Long (2016): $(a, b) = (0.1, 0.1)$, $(r, s) = (0.01, 0.01)$, and $(g, h) = (1, 1)$. In the MI-PCA algorithm, we extracted enough components to explain 50% of the total variance in the data. For the single random forest imputations, we used the *missForest* R package (Stekhoven, 2013) which implements Algorithm 1 from Stekhoven and Bühlmann (2011). When evaluating convergence, we found that the stopping criterion for the missFor algorithm was usually met within the first 10 iterations, and we fixed the maximum number of iterations to 20. Stekhoven and Bühlmann (2011) recommended growing 100 trees per forest; therefore, we used this value in our study.

3.1.4. Analysis

The substantive model of interest in Experiment 1 was a saturated model that estimated means, variances, and covariances of the six variables with missing values. Therefore, our analysis model estimated six means, six variances, and 15 covariances.

3.2. Comparison criteria

We compared the different missing data treatments in terms of bias and confidence interval coverage.

3.2.1. Bias

For a given parameter of interest θ (e.g., mean of item 1, variance of item 2), we used the percent relative bias (PRB) to quantify the estimation bias introduced by the imputation procedure:

$$PRB = \frac{\bar{\hat{\theta}} - \dot{\theta}}{\dot{\theta}} \times 100 \tag{10}$$

where $\dot{\theta}$ is the true value of the focal parameter defined as $\sum_{s=1}^{S} \hat{\theta}_s^{GS}/S$, with $\hat{\theta}_s^{GS}$ being the Gold Standard parameter estimate for the sth repetition. The averaged focal parameter estimate under a given missing data treatment was computed as $\bar{\hat{\theta}} = \sum_{s=1}^{S} \hat{\theta}_s/S$, with $\hat{\theta}_s$ being the estimate obtained from the treated incomplete data in the sth repetition. Following Muthén, Kaplan, and Hollis (1987), we considered $|\text{PRB}| > 10$ as indicative of problematic estimation bias.

3.2.2. Confidence intervals coverage

To assess the performance in hypothesis testing and interval estimation, we evaluated the confidence interval coverage (CIC) of the true parameter value:

$$CIC = \frac{\sum_{s=1}^{S} I(\dot{\theta} \in \widehat{CI}_s)}{S} \tag{11}$$

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

CICs below 0.9 are usually considered problematic for 95% CIs (Van Buuren, 2018, p. 52) as they imply inflated Type I error rates. High CICs (e.g., 0.99) indicate CIs that are too wide, implying inflated Type II error rates. Therefore, we considered CIs to show severe under-coverage (over-coverage) if CIC < 0.9 (CIC > 0.99). From a testing perspective, a CIC can be considered as significantly different from the nominal coverage rate if the magnitude of its difference from the nominal coverage probability, p_0, is more than two times the standard error of p_0, $SE(p_0) = \sqrt{p(1-p)/S}$ (Burton, Altman, Royston, & Holder, 2006). Therefore, we considered 95% CI coverages outside the interval $[0.94, 0.96]$ to be significantly different from the nominal coverage rate.

3.3. Results

We computed both PRB and CIC for each of the 27 parameters in the analysis model (six means, six variances, and 15 covariances). To summarize the results, we focus on the typical and extreme values of these measures. In Figures 1 and 2, we report the average, minimum, and maximum absolute PRB and CIC for each missing data treatment method and each parameter type. In the supplementary material, we include figures reporting the raw PRB and CIC for every parameter estimate.

3.3.1. Means

The largest |PRB| for the means was below 10 for all imputation methods. Only CC produced problematic degrees of bias. However, looking at the relative performances, IURR and MI-PCA resulted in smaller biases than all other methods (except MI-OP). In terms of CIC, only bridge and MI-PCA showed consistently strong performance. Neither method demonstrated any extreme under-/over-coverage (i.e., CIC $\notin$ [0.9, 0.99]), and MI-PCA showed nonsignificant deviations from nominal coverage for almost all estimates. DURR, IURR, and blasso demonstrated reasonable coverages when *pm* was low, but tended to under-cover to problematic degrees when *pm* was high. The tree-based methods and CC performed most poorly. These methods led to CICs sig-

nificantly different from nominal coverage rates in all conditions and demonstrated extreme under-coverage in many conditions.

3.3.2. Variances

IURR, blasso, and the tree-based MI methods resulted in low biases (i.e., $|\text{PRB}| < 10$) across all conditions. For blasso, these low biases were paired with low deviations from nominal coverage rates. IURR only demonstrated problematic CICs for the high-dim-high-pm condition where it produced extreme under-coverage. MI-CART and MI-RF only produced reasonable coverage rates when *pm* was low. MI-PCA showed acceptable biases and reasonable coverage rates in all but the high-dim-high-pm condition where it showed large biases and extreme under-coverage. DURR showed poor performance with regard to the item variances. Although DURR produced acceptable levels of bias in all but the high-dim-high-pm condition, it produced extreme CI under-coverage in all but the low-dim-low-pm condition. Bridge, missFor, and CC performed poorly in nearly all conditions. These methods tended to demonstrate substantial biases and extreme under-coverage. Although bridge produced reasonable CICs with low *pm*, the extreme biases it produced negate any benefits of good coverage properties.

3.3.3. Covariances

MI-PCA was the only method that showed consistently strong performance when estimating covariances. MI-PCA showed negligible bias and minimal deviations from nominal coverage in all conditions. The MI-PCA never produced extreme under-/over-coverage, and when the CIC was significantly different from the nominal rate, the CIs showed mild *over*-coverage (i.e., CICs greater than 0.96 but smaller than 0.99). After MI-PCA, IURR demonstrated the second strongest performance, with negligible bias and acceptable coverage in all but the high-dim-low-pm condition. In the high-dim-high-pm condition, IURR produced large biases and extreme under-coverage.

Bridge displayed low bias and acceptable coverage in the low dimensional conditions (columns 1 and 3 of Figures 1 and 2) but borderline-acceptable to unacceptable biases and deviations from nominal coverage in the high dimensional conditions (columns 2 and 4 of Figures 1 and 2). DURR, blasso, and the tree-based MI methods

tended to produce unacceptable biases in all but the low-dim-low-pm condition, with accompanying under-coverage of the true covariance values. MissFor and CC showed extreme bias and under-coverage in all conditions.

[Figure 1 about here.]

[Figure 2 about here.]

4. Experiment 2: simulated data with latent structure

In the second simulation experiment, we generated data from a confirmatory factor analysis (CFA) model. The data social scientists analyze often comprise items measuring different latent constructs, a characteristic that is likely to impact imputation performance. In Experiment 2, we varied three design factors: dimensionality of the data, size of the factor loadings, and proportion of missing data. We controlled the dimensionality of the data by the number of latent variables, $l \in \{10, 100\}$. We generated five items to measure each latent variable, resulting in either 50 or 500 total items. We held the sample size constant at $n = 200$, so conditions with $l = 10$ resulted in a low-dimensional data, while conditions with $l = 100$ resulted in high-dimensional data. We defined the factor loadings as a two-level random factor. We sampled the high factor loadings from Unif(0.9, 0.97), and we sampled the low factor loadings from Unif(0.5, 0.6). We defined the proportion of missing values as either $pm = 0.1$ or $pm = 0.3$. Table 2 summarizes the eight resulting conditions. We conducted $S = 1000$ replications of each condition. For each replicate, missing values were imposed, and then each of the missing data treatments used in Experiment 1 was used to obtain estimates of the parameters of a the analysis models described below.

[Table 2 about here.]

4.1. Simulation study procedure

4.1.1. Data generation

For each replication, we created an $n \times p$ data matrix, $\boldsymbol{Z}$, based on a CFA model. Each of l latent variables was measured by five items, for a total of $p = 5 * l$ columns in $\boldsymbol{Z}$. Values of the items for the ith observation were generated with the following measurement model:

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i \tag{12}$$

where $\boldsymbol{z}_i$ is a vector of $5 * l$ item scores for observation $i = 1, ..., n$; $\boldsymbol{\Lambda}$ is the $(5 * l) \times l$ matrix of factor loadings; $\boldsymbol{\xi}_i$ is a vector of l latent variable scores for observation i, and $\boldsymbol{\delta}_i$ is a vector of $5 * l$ measurement errors sampled from a multivariate normal distribution with means of zero and a diagonal covariance matrix, $\boldsymbol{\Theta}$. We rescaled all items to have means of five. For notation and model specification, the interested reader may refer to Bollen (1989).

We sampled the latent scores in $\boldsymbol{\xi}_i$ from a multivariate normal distribution with means of zero and covariance matrix $\boldsymbol{\Psi}$. We defined the off-diagonal elements of $\boldsymbol{\Psi}$ such that the first four latent variables were highly correlated ($\rho = 0.6$), the second block of four latent variables were weakly correlated ($\rho = 0.3$), and the remaining $l - 8$ latent variables were uncorrelated.

$\boldsymbol{\Lambda}$ defined a simple latent structure where each item loaded onto only one factor. Both the item and latent factor variances were set to 1, so the measurement error variance was defined as $var(\delta) = 1 - \lambda^2$. This specification allowed the factor loadings to be interpreted as standardized values between 0 and 1. For each repetition, we sampled the exact values of the factor loadings from a uniform distribution with bounds defined by the condition.

4.1.2. Missing data imposition

Item nonresponse was imposed on 10 items in $\boldsymbol{Z}$ using Equation (9) to define the probability of nonresponse. The indicators of the first two highly correlated latent

variables ($l = \{1, 2\}$) were candidates for missing data imposition. The latent scores for the other two highly correlated latent variables ($l = \{3, 4\}$) acted as the missing data predictors comprising the columns of $\tilde{\boldsymbol{Z}}$.

4.1.3. Imputation

The missing values were treated with all the methods previously described. The imputation methods were parameterized as in Experiment 1. The same convergence checking procedure used in Experiment 1 suggested that 50 iterations were sufficient for convergence of all MI methods except blasso, which required approximately 2,000 iterations for convergence.

4.1.4. Analysis

The substantive model of interest in Experiment 2 was a saturated model that estimated means, variances, and covariances of the raw items with missing values. Furthermore, we estimated the true CFA model for the same items to evaluate the estimation of the factor loadings. Results for the CFA model are reported in the supplementary material.

4.2. Results

To assess the performance of the methods, we used the same comparison criteria described for Experiment 1. Figures 3 and 4 report the maximum, average, and minimum |PRB| and CIC obtained with each missing data treatment method for each parameter type (mean, variance, and covariance) in the conditions with high factor loadings. Figures 5 and 6 report the same results for the conditions with low factor loadings. We included figures reporting the raw PRBs and CICs for every parameter in the supplementary materials.

4.2.1. Means

All imputation methods resulted in unbiased estimates of the item means in all conditions. In particular, IURR, MI-PCA, and MI-OP exhibited the smallest bias, and

only CC produced any meaningful bias of the means. MI-PCA was the only method that produced reasonable coverages in all conditions. When *pm* was low, all methods expect missFor and CC produced reasonable coverages, regardless of the factor loading size. When factor loadings were large, DURR and IURR resulted in minimal deviations from nominal coverage, while blasso, bridge, MI-CART, MI-RF, and missFor led to problematic under-coverage when the *pm* was high. In the conditions with lower factor loadings, all methods performed worse, but relative performances remain unchanged. IURR, MI-PCA, and MI-OP again exhibited the smallest biases and deviations from nominal coverage rates. However, MI-PCA and MI-OP were the only methods that produced reasonable coverage rates in the high-dim-high-pm-low-λ condition. CC produced extreme levels of under-coverage across all conditions.

4.2.2. Variances

4.2.2.1. High factor loadings. All MI methods, except bridge, resulted in acceptable estimation bias for the item variances in the conditions with large factor loadings. The bridge estimates were unbiased in the low-dimensional conditions but biased in the high-dimensional conditions. IURR, MI-PCA, and MI-OP produced the least biased estimates, while missFor and CC tended to produce large biases. In the low-pm conditions, all methods expect for missFor and CC produced reasonable coverages. In the high-pm conditions, only IURR, MI-PCA, and MI-OP maintained reasonable coverages. Bridge produced good coverage in the low-dim-high-pm condition, wherein blasso and DURR produced borderline-acceptable coverages. All three methods, however, produced problematic under-coverage in the high-dim-high-pm condition. The tree-based MI methods produced extreme under-coverage whenever *pm* was high. CC and missFor again showed extreme under-coverage in all conditions.

4.2.2.2. Low factor loadings. IURR, blasso, MI-CART, MI-RF, CC, and MI-OP produced unbiased variance estimates in all low-λ conditions. IURR, blasso, and the tree-based MI methods showed the lowest biases. Bridge produced unacceptable biases in the low-dim-high-pm condition, while DURR and MI-PCA produced large biases in the high-dim-high-pm condition. missFor tended to produce problematic biases in

all conditions. Only blasso and MI-OP produced reasonable coverages in all low-λ conditions. IURR and MI-PCA produced reasonable coverage in all but the high-dim-high-pm condition where they both led to substantial under-coverage. The tree-based MI methods produced good coverage with low *pm* and borderline-acceptable coverages in the high-pm conditions. DURR produced extreme under-coverage in all but the low-dim-low-pm condition, while missFor and CC tended to produce problematic under-coverage in all conditions.

4.2.3. Covariances

As in Experiment 1, MI-PCA resulted in the lowest biases and the smallest deviations from nominal coverage. Other than MI-OP, MI-PCA was the only method that produced unbiased covariance estimates and reasonable coverages across all conditions. IURR produced unbiased estimates and acceptable to borderline-acceptable CICs in all but the high-dim-high-pm-low-λ condition. For all the conditions with high factor loadings, DURR showed acceptable biases but under-covered in the high-dim-high-pm-high-λ condition. However, DURR produced large covariance biases and under-coverage in all the low-λ conditions expect for the low-dim-low-pm-low-λ condition. Except in the low-pm-high-λ conditions, all other methods tended to produce large biases and problematic degrees of under-coverage.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

5. Experiment 3: EVS resampling study

In the third experiment, we performed a resampling study based on the EVS data to assess how well the results obtained from Experiments 1 and 2 carried over to real data applications. EVS is a large-scale, cross-national survey on human values administered in around 50 countries across Europe. It covers a wide range of human values regarding

family, work, environment, perceptions of life, politics, society, religion, morality, and national identity. EVS is a high-quality survey widely used for comparative studies between European countries. Furthermore, it is accessible free of charge and it represents the type of data social scientist regularly analyze. Variables in the EVS data are discrete numerical and categorical items following a variety of distributions.

In Experiment 3, we treated the original EVS data as the population from which we drew the samples used in the resampling study. We investigated the performance of the methods by resampling $S = 1000$ datasets of n units from this population. For each replicate, we imposed missing values, treated them with the same methods used in experiments 1 and 2, and pooled the analysis model parameter estimates. This procedure was repeated for a low-dimensional and a high-dimensional condition. As the number of predictors in the data was fixed at $p = 243$, we controlled the dimensionality of the data by defining different sizes for the samples taken from the EVS population data ($n \in \{1000, 300\}$).

5.1. Resampling study procedure

5.1.1. Data preparation and sampling

We used the third prerelease of the 2017 wave of EVS data (EVS, 2020a) to define a population dataset with no missing values. The original dataset contained 55,000 observations from 34 countries. We selected only the four founding countries of the European Union included in the dataset (France, Germany, Italy, and the Netherlands) and excluded all columns that contained duplicated information (e.g., recoded versions of other variables), or metadata (e.g. time of interview, mode of data collection).

We used the R package *mice* to run a single imputation with predictive mean matching (PMM) to fill the originally missing values. We employed the variable selection procedure described by Van Buuren, Boshuizen, and Knook (1999, pp. 687–688) and implemented in the *quickpred* function to select the predictors in the imputation models. We implemented the variable selection by setting the minimum correlation threshold in *quickpred* to 0.3. The number of MI iterations was set to 200. We used SI, and not MI, because this imputation procedure was used to obtain a set of pseudo-fully

observed data to act as the population in our resampling study and not for statistical inference. At the end of the data cleaning process, we obtained a (pseudo) fully observed dataset of 8045 observations across four countries with $p = 243$ variables. For every replicate in the resampling study, we generated a bootstrap sample by sampling n observations with replacement from this dataset.

5.1.2. Analysis models

To define plausible analysis models, we searched for models that have been used in published articles testing social scientific theories on the EVS data. The search was performed by screening the repository of publications using EVS data available on the EVS website (EVS, 2020b).

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

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\beta C} \tag{13}$$

where a dependent variable $\boldsymbol{y}$ is regressed onto a variable of interest $\boldsymbol{x}$ and a set of control variables $\boldsymbol{C}$. In this scenario, β_1 is a focal parameter that a researcher wants to use to test some hypothesis. We used the bias and CIC of the regression coefficients from these models as the outcome measures by which we evaluated the relative performance of the missing data methods.

Model 1 was inspired by Köneke (2014): the dependent variable was a 10-point EVS item measuring euthanasia acceptance ('Can [euthanasia] always be justified, never be justified, or something in between?'); the predictor of interest was a 4-point item measuring the self-reported importance of religion in one's life; the set of covariates included trust in the health care system, trust in the state, trust in the press, country, sex, age, education, and religious denomination. This model could be used to test a hypothesis regarding the effect of religiosity on the acceptance of end-of-life treatments.

Model 2 was inspired by Immerzeel, Coffé, and Van der Lippe (2015). The dependent variable was a harmonized variable constructed by EVS to describe the re-

spondents' tendencies to vote for left- or right-wing parties, expressed on a 10-point left-to-right continuum. The predictor of interest was a scale measuring respondents' attitudes toward immigrants and immigration ('nativist attitudes scale'). The scale was obtained by taking the average of respondents' agreement, on a scale from 1 to 10, with three statements: 'immigrants take jobs away from natives', 'immigrants increase crime problems', and 'immigrants are a strain on welfare system'. The control variables used were: attitudes toward law and order, attitudes toward authoritarianism, interest in politics, level of political activity, country, sex, age, education, employment status, socio-economic status, importance of religion in life, religious denomination, and the size of the town where interview was conducted. A researcher might fit this model to test a hypothesis regarding the effect of xenophobia on voting tendencies.

5.1.3. Missing data imposition

We imposed missing data on six variables using the same strategy described for experiments 1 and 2. The targets of missing data imposition were the two dependent variables in models 1 and 2 (i.e., euthanasia acceptance, and left-to-right voting tendency); religiosity (the focal predictor in model 1 and a control variable in model 2); and the three items making up the "nativist attitudes" scale (the focal predictor in model 2).

The response model was the same as in Equation (9), and three variables were included in $\tilde{\boldsymbol{Z}}$: age, education, and an item measuring trust in new people. We chose these predictors because older people tend to have higher item nonresponse rates than younger people, and lower educated people tend to have higher item non-response rates than higher educated people (De Leeuw, Hox, & Huisman, 2003; Guadagnoli & Cleary, 1992). We also assumed that people with less trust in strangers would have a higher nonresponse tendency as they are likely to withhold more information from the interviewer (a stranger).

5.1.4. Imputation

We treated the missing values with the same methods used in Experiments 1 and 2. The imputation methods were parameterized as in Experiments 1 and 2, and convergence

checks were performed in the same way. These convergence checks suggested that the imputation models had converged after 60 iterations.

5.2. Results

When estimating linear regression models, all partial regression coefficients can be influenced by missing values on a subset of the variables included in the model. Therefore, it is important to observe the estimation bias and CIC rates for all model parameters. Figure 7 reports the absolute PRBs for the intercept and all partial slopes from Model 2, under the different imputation methods, for both the low- and high-dimensional conditions. Figure 8 reports CIC results in the same way. Results for Model 1 are reported in the supplementary materials.

As seen in Figure 7, even MI-OP did not provide entirely unbiased parameter estimates. Around half of the estimates obtained with MI-OP had large biases ($|\text{PRB}| > 10\%$). The largest MI-OP biases were considerable: around 40 in the low-dimensional condition and 20 in the high-dimensional condition. In both the high- and low-dimensional conditions, DURR, IURR, blasso, MI-CART, and missFor showed only slightly larger |PRB|s than MI-OP. MI-PCA and MI-RF showed similar trends but produced somewhat larger |PRB|s. Bridge demonstrated the same results described in the simulation studies. It was competitive in low-dimensional scenarios, but it was inadequate with high-dimensional data (all but one |PRB| was larger than 50). For all other methods, |PRB|s were smaller in the high-dimensional condition.

DURR, IURR, and MI-CART maintained similar coverage patterns to MI-OP, with only a few significant deviations from nominal coverage rates. MI-PCA, blasso, and MI-RF significantly over-covered more than half of the parameters but did produce any extreme over-/under-coverage (except for a single parameter estimate by MI-RF). All MI methods led to CIC closer to the nominal rate in the high-dimensional condition. As expected, imputation by missFor led to significant under-coverage of most regression coefficients. Despite showing poor performance in terms of bias, CC manifested good coverage rates. However, this was a result of the smaller sample size used for estimating the analysis model, rather than a positive feature of the method. The smaller samples produced wider intervals which covered the true values even when

the point-estimates were biased. Notably, very few of the CICs fell into the range of extreme over-/under-coverage. Only the high-dimensional estimates from bridge consistently exhibited extreme under-/over-coverage.

[Figure 7 about here.]

[Figure 8 about here.]

5.2.1. Imputation time

Table 3 reports the average imputation time for the different methods. IURR and DURR were the most time-consuming methods, with imputation times above one hour in the low-dimensional condition. MI-PCA and blasso had imputation times of a minute or less. In the high-dimensional condition, IURR and DURR were not as time-intensive due to the smaller sample size but still took more than ten times longer than MI-PCA and blasso.

[Table 3 about here.]

6. Discussion

The inclusion of high-dimensional prediction techniques within the MICE framework has the potential to simplify the use of MI for social scientists. We studied the bias and CI coverage of parameter estimates after imputation by seven high-dimensional MI methods. Although extensive simulation studies had already been carried out by the researchers proposing these methods, no comparison study had been developed to assess their relative performance. Our research fills this gap and provides initial insights into applying such methods in social scientific research. In this section, we discuss the overall performance of the methods and we give recommendations for social scientists facing high-dimensional imputation problems.

6.1. Methods that do not work well

We found that bridge is an inadequate solution to high-dimensional data imputation problems. In both the simulation and the resampling study, the use of a fixed

ridge penalty within the imputation algorithm manifested the same undesirable performance. The method worked well when the imputation task remained low-dimensional. However, bridge led to extreme bias and unacceptable CI coverage in nearly all the high-dimensional conditions.

We have also confirmed that missFor—the only single imputation method we evaluated—leads to low estimation bias but results in severe CI under-coverage. Under-coverage coupled with unbiased estimates indicates that too little uncertainty is incorporated into the imputation procedure, which is to be expected from a single imputation approach. As a result, missFor should be avoided by social scientists who wish to draw inferential conclusions from their data analysis.

6.2. Methods that work well

IURR and MI-PCA were the two strongest performers. IURR produced the smallest estimation bias for item means, variances and regression coefficients. The method also produced small deviations from nominal coverage rates for these parameters. Furthermore, the large covariance estimation bias introduced by IURR in the high-dim-high-pm conditions only slightly exceeded the $|\text{PRB}| = 10$ threshold and the CIC was around 0.9. Comparatively, most of the other MI methods resulted in covariance |PRB|s larger than 20, and CICs well below 0.9.

IURR does not require the imputer to make choices regarding which variables are relevant for the imputation procedure. The only additional decision required of the imputer is selecting the number of folds to use when cross-validating the penalty parameter. As a result, an IURR imputation run is easy to specify, which makes IURR an appealing method for imputation of large social scientific datasets. However, IURR is relatively computationally intensive. If the number of variables with missing values is large, IURR might result in prohibitive imputation time. In such a scenario, a researcher might prefer to address imputation with the MI-PCA method.

MI-PCA showed low bias and good coverage for both item means and covariances in experiments 1 and 2. Although it exhibited a large bias of the item variances, the relationships between variables with missing values were always correctly estimated. It was the only method resulting in low bias and close-to-nominal CI coverage of the

true covariance values, even in the high-dimensional conditions. Finally, when the CICs obtained with MI-PCA deviated significantly from nominal rates, they over-covered. In most situations, this tendency is less worrisome than under-coverage as it leads to conservative, rather than liberal, inferential conclusions. Our results suggest that MI-PCA is a promising approach for data analysts interested in testing theories on large social scientific datasets with missing values. We are currently conducting research to explore—and extend—the capabilities of the MI-PCA approach more fully.

6.3. Methods with mixed results

DURR produced low bias and good CI coverage for item means, variances, and regression coefficients. However, compared to IURR, it suffered from greater performance deterioration when applied to high-dimensional data. As a result, our results suggest that DURR should not be preferred to IURR.

There was little difference in performance between the use of CART and random forests as elementary imputation methods within the MICE algorithm. In line with what Doove et al. (2014) found, when a difference was noticeable, the simpler CART generally outperformed the more complex random forests. Both MI-CART and MI-RF produced large covariance bias in experiments 1 and 2. Although bias for means, variances, and regression coefficients was acceptable, it was usually larger than that obtained by other MI methods. Furthermore, in terms of CI coverage, both methods showed large under-coverage of most parameters in the high-dim-high-pm conditions. Although the nonparametric nature of these approaches elegantly avoids over-parameterization of imputation models, these methods were still outperformed by IURR and MI-PCA.

Blasso resulted in low biases for item means and variances, even in the high-dimensional conditions. While the covariance bias was large in experiments 1 and 2, blasso performed well in the resampling study, where the overall bias levels were similar to those of MI-OP. In terms of CI coverage, blasso showed poor performance resulting in either CI under-coverage or CI over-coverage in almost all high-dimensional conditions.

The mixed performance of blasso is also accompanied by a few obstacles to its

application for social scientific research. Using Hans (2010b)'s Bayesian Lasso requires the specification of six hyper-parameters, which introduces more researcher degrees of freedom and demands a strong grasp of Bayesian statistics. Furthermore, the method has not currently been developed for multi-categorical data imputation, a common task in the social sciences. As a result, we do not recommend blasso for imputation of large social scientific datasets.

7. Limitations and future directions

The present work was aimed at comparing current implementations of existing imputation methods. As a result, the scope of the simulation and resampling studies was limited by the current development state of the different methods. For example, DURR, IURR, and MI-PCA allow imputation of any type of data: DURR and IURR have been developed for categorical data imputation (Deng et al., 2016), and MI-PCA can be performed with any standard imputation model for categorical data. However, blasso has not been formally developed for imputing multi-categorical variables yet. This limitation of blasso forced us to work with missing values on variables that are either continuous, or usually considered as such in practice (e.g., Likert-type scales). To maintain a fair comparison with blasso, all methods were implemented with the assumption that the imputed variables are continuous and normally distributed. However, IURR, DURR and MI-PCA could have performed differently in the resampling study if we had used their ordinal data implementations.

Another limitation of this study is the assumption of a linear missing data mechanism. In real social scientific data, the response mechanism might be nonlinear, a condition that could require including transformations of the raw variables (e.g., interactions, polynomial terms) in the imputation models. Non-linear response models were not part of the scope of this project. However, all of the high-dimensional imputation methods considered have the potential to account for more complex response mechanisms.

Finally, these results only apply to the specific implementations of the algorithms we used. Many of the methods discussed could have been implemented differently.

Zhao and Long (2016) proposed versions of IURR and DURR using the elastic net penalty (Zou & Hastie, 2005) and the adaptive lasso (Zou, 2006) instead of the lasso penalty. Although no substantial performance differences between penalty specifications emerged from the work of Zhao and Long (2016) or Deng et al. (2016), we must acknowledge that we did not investigate the impact of different types of regularization in the present study.

MI-PCA requires making a decision on the number of components to extract from the auxiliary variables. In this study, we decided to retain the first components that explained 50% of the total variance in the auxiliary variables. However, this decision was arbitrary. We plan on assessing its effect on the imputation accuracy as part of a project to expand and improve the use of principal components within the MICE framework.

As for blasso, we have not investigated the sensitivity of the results to different hyper-parameters choices. Furthermore, alternative implementations of Bayesian Lasso could be used within a MICE framework. In particular, the well-known Bayesian Lasso proposed by Park and Casella (2008) is a viable option.

We could have also implemented the random forests differently. We decided to use the Doove et al. (2014) version which is supported in the popular R package *mice*. However, Shah, Bartlett, Carpenter, Nicholas, and Hemingway (2014) independently developed another implementation of random forests within the MICE algorithm, which was available in the now archived R package CALIBERrfimpute (Shah, 2018). We are not aware of any evidence or theoretical reason to expect differences between the two implementations, but we did not verify this empirically.

Disclosure statement

No authors reported any financial or other conflicts of interest in relation to the work described.

Funding

This work was supported by the Tilburg School of Social and Behavioural Sciences. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

Data Availability

The data that support the findings of this study are openly available in GESIS Data Archive at https://doi.org/10.4232/1.13511, reference number ZA7500.

References

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

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Burgette, L. F., & 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., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in medicine*, *25*(24), 4279–4292.

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

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

de Andrade Silva, J., & Hruschka, E. R. (2009). Eacimpute: an evolutionary algorithm for clustering-based imputation. In *2009 ninth international conference on intelligent systems design and applications* (pp. 1400–1406).

De Leeuw, E. D., Hox, J. J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*, 153–176.

Deng, Y., Chang, C., Ido, M. S., & 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., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104.

EVS. (2020a). *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.

EVS. (2020b). *Evs bibliography.* (`https://europeanvaluesstudy.eu/education-dissemination-publications/evs-publications/publications/` [Accessed: 2020-09-30])

Guadagnoli, E., & Cleary, P. D. (1992). Age-related item nonresponse in surveys of recently discharged patients. *Journal of Gerontology*, *47*(3), P206–P212.

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

Hans, C. (2010a). blasso: Mcmc for bayesian lasso regression model [Computer software manual]. Retrieved from `http://www.stat.osu.edu/~hans/` (R package version 0.3)

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

Howard, W. J., Rhemtulla, M., & 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., & 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., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, *21*(2), 187–198.

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

Lang, K. M., Little, T. D., & PcAux Development Team. (2018). Pcaux: Automatically extract auxiliary features for simple, principled missing data analysis [Computer software manual]. Retrieved from `https://github.com/PcAux-Package/PcAux` (R package version 0.0.0.9013)

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

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*(3), 431–462.

pandas development team, T. (2020, February). *pandas-dev/pandas: Pandas.* Zenodo. Re-

trieved from https://doi.org/10.5281/zenodo.3509134

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

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94.

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

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* (Vol. 72). Boca Raton, FL: Chapman & Hall/CRC.

Shah, A. D. (2018). Caliberrfimpute: Imputation in mice using random forest [Computer software manual]. (R package version 1.0-1)

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & 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., & Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, *23*(18), 2827–2843.

Stekhoven, D. J. (2013). missforest: Nonparametric missing value imputation using random forest [Computer software manual]. (R package version 1.4)

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

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

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine*, *18*(6), 681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivarite imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data.

Statistical Methods in Medical Research, *25*(5), 2021–2035.

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

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

Table 1. Summary of conditions for Experiment 1.

condition	label	n	p	pm
1	low-dim-low-pm	200	50	0.1
2	high-dim-low-pm	200	500	0.1
3	low-dim-high-pm	200	50	0.3
4	high-dim-high-pm	200	500	0.3

Table 2. Summary of conditions for Experiment 2.

condition	label	n	p	l	pm	λ range
1	low-dim-low-pm-high-λ	200	50	10	0.1	[0.9, 0.97]
2	high-dim-low-pm-high-λ	200	500	100	0.1	[0.9, 0.97]
3	low-dim-high-pm-high-λ	200	50	10	0.3	[0.9, 0.97]
4	high-dim-high-pm-high-λ	200	500	100	0.3	[0.9, 0.97]
5	low-dim-low-pm-low-λ	200	50	10	0.1	[0.5, 0.6]
6	high-dim-low-pm-low-λ	200	500	100	0.1	[0.5, 0.6]
7	low-dim-high-pm-low-λ	200	50	10	0.3	[0.5, 0.6]
8	high-dim-high-pm-low-λ	200	500	100	0.3	[0.5, 0.6]

Table 3. Average imputation time in minutes for the MI methods compared in Experiment 3.

condition	DURR	IURR	blasso	bridge	MI-PCA	MI-CART	MI-RF	MI-OP
low-dim (n = 1000)	73.20	75.90	1.40	8.10	0.60	4.00	11.30	2.20
high-dim (n = 300)	6.10	9.70	0.50	3.20	0.40	1.40	4.70	1.90

[a] In both conditions, the number of data columns was 243; n refers to the sample size.

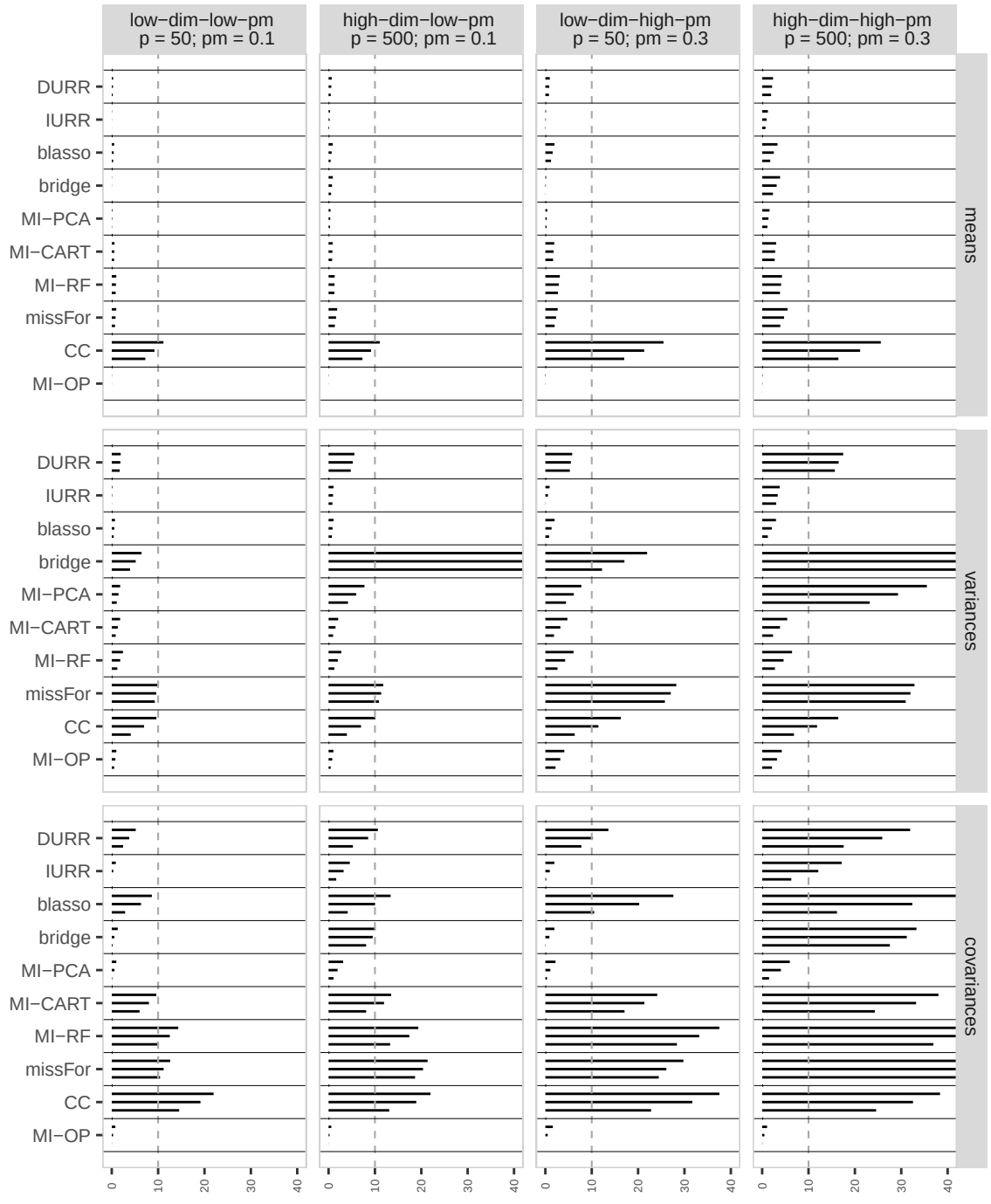

Figure 1. Maximum, average, and minimum absolute percent relative bias (|PRB|) for item means, variances, and covariances in Experiment 1.

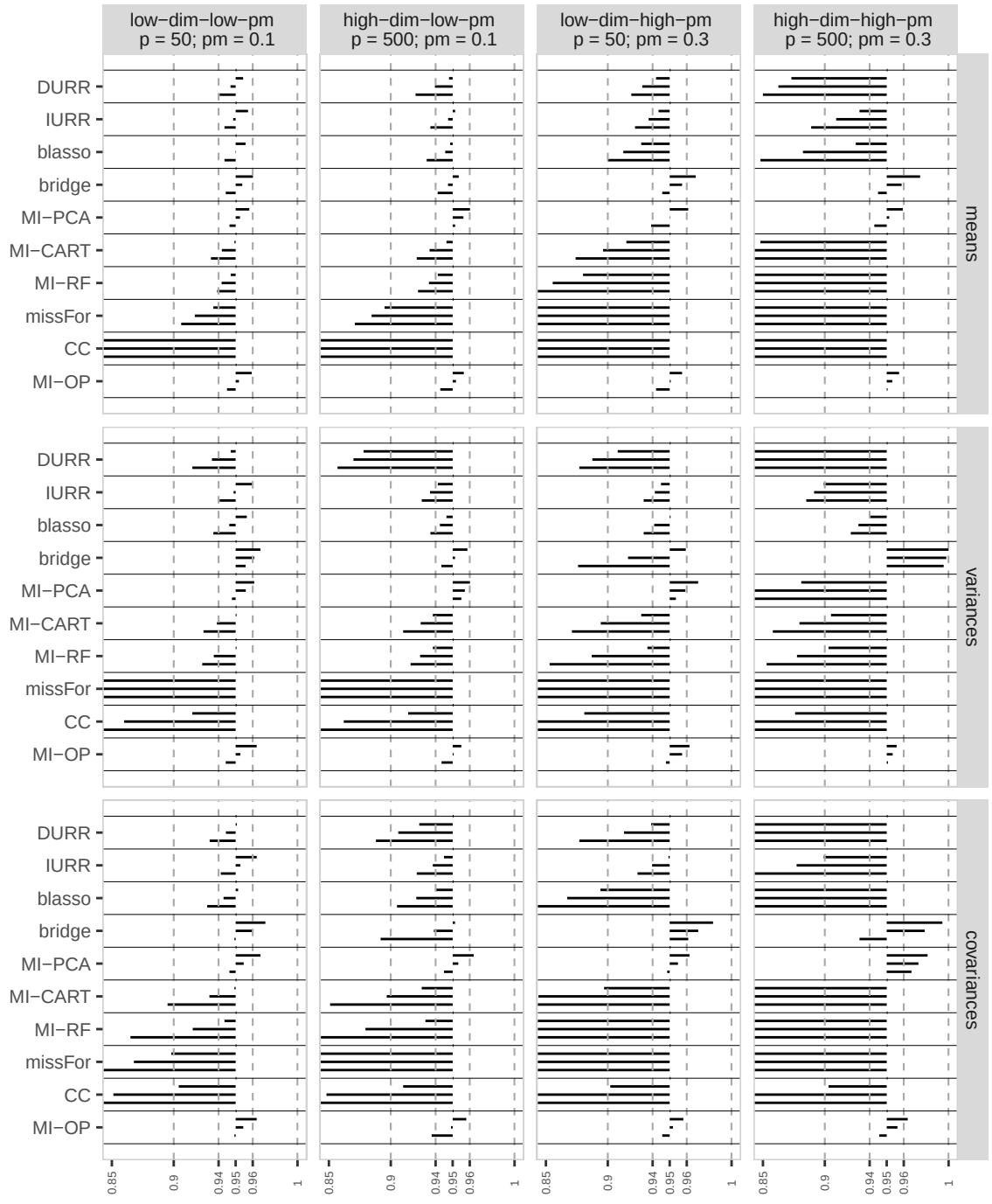

Figure 2. Maximum, average, and minimum CIC for item means, variances, and covariances in Experiment 1.

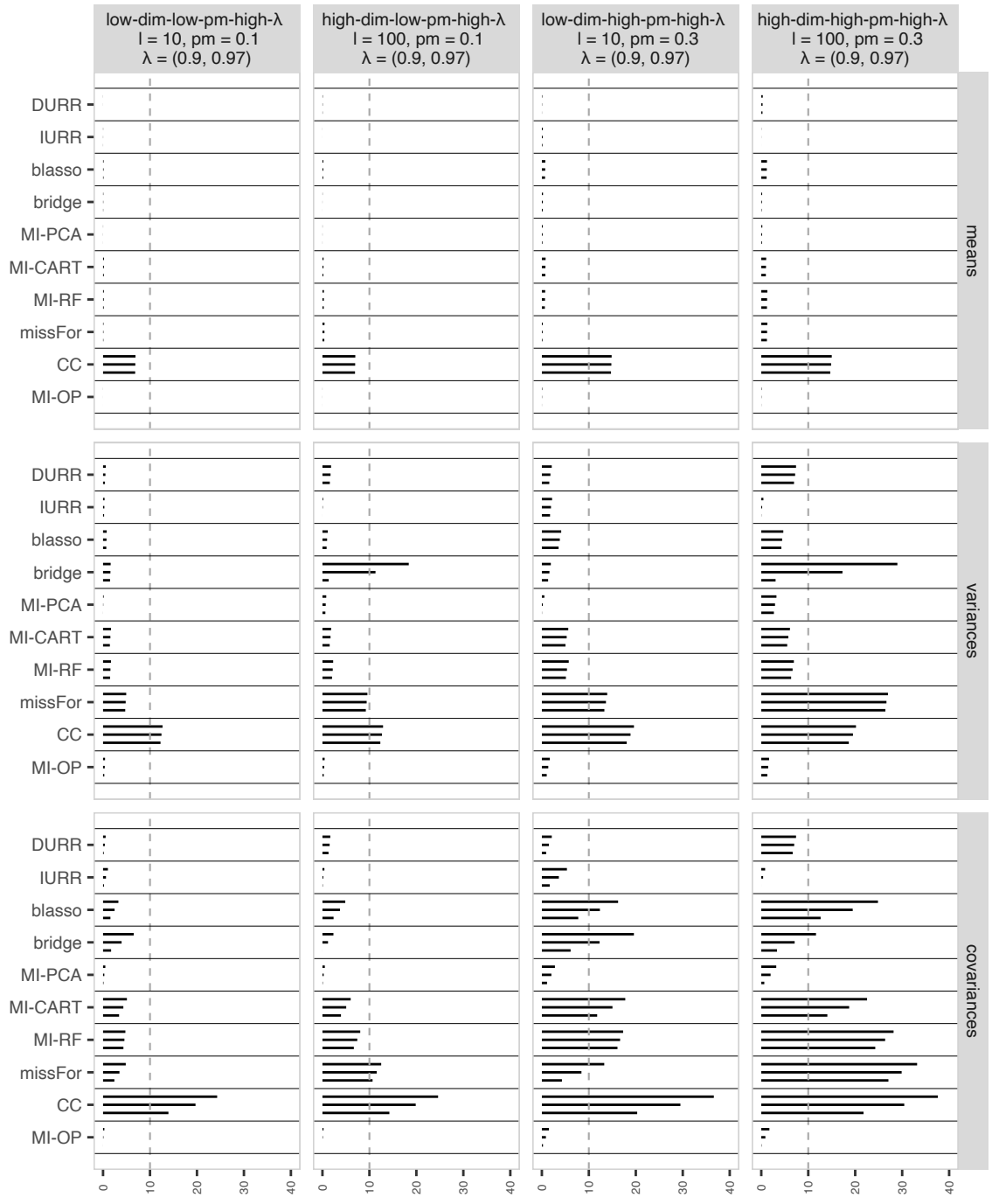

Figure 3. Maximum, average, and minimum absolute Percent Relative Bias (|PRB|) for item means, variances, and covariances in the conditions of Experiment 2 with high factor loadings.

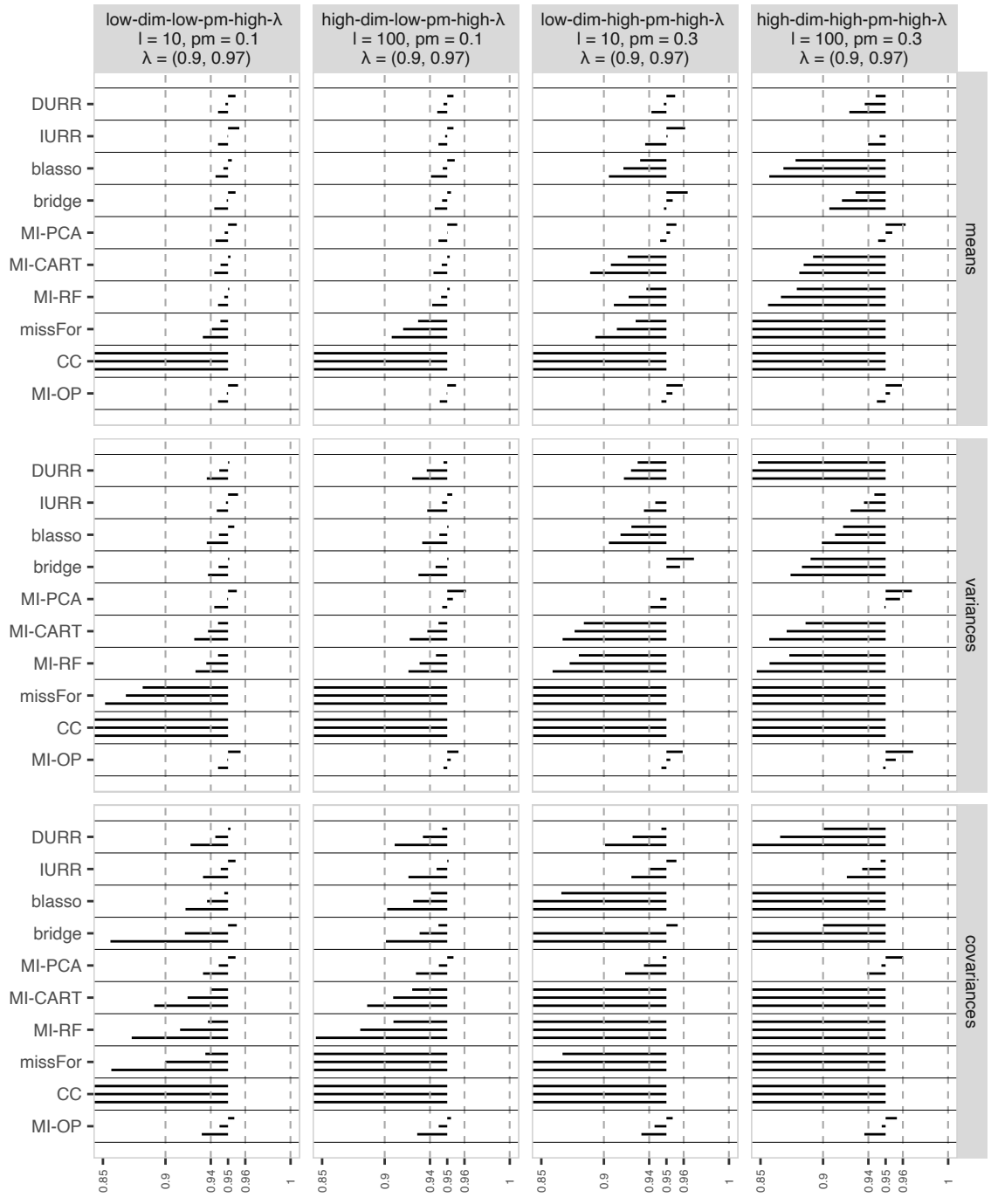

Figure 4. Maximum, average, and minimum CIC for item means, variances, and covariances in Experiment 2 in the conditions of Experiment 2 with high factor loadings.

Figure 5. Maximum, average, and minimum absolute Percent Relative Bias (|PRB|) for item means, variances, and covariances in the conditions of Experiment 2 with low factor loadings.

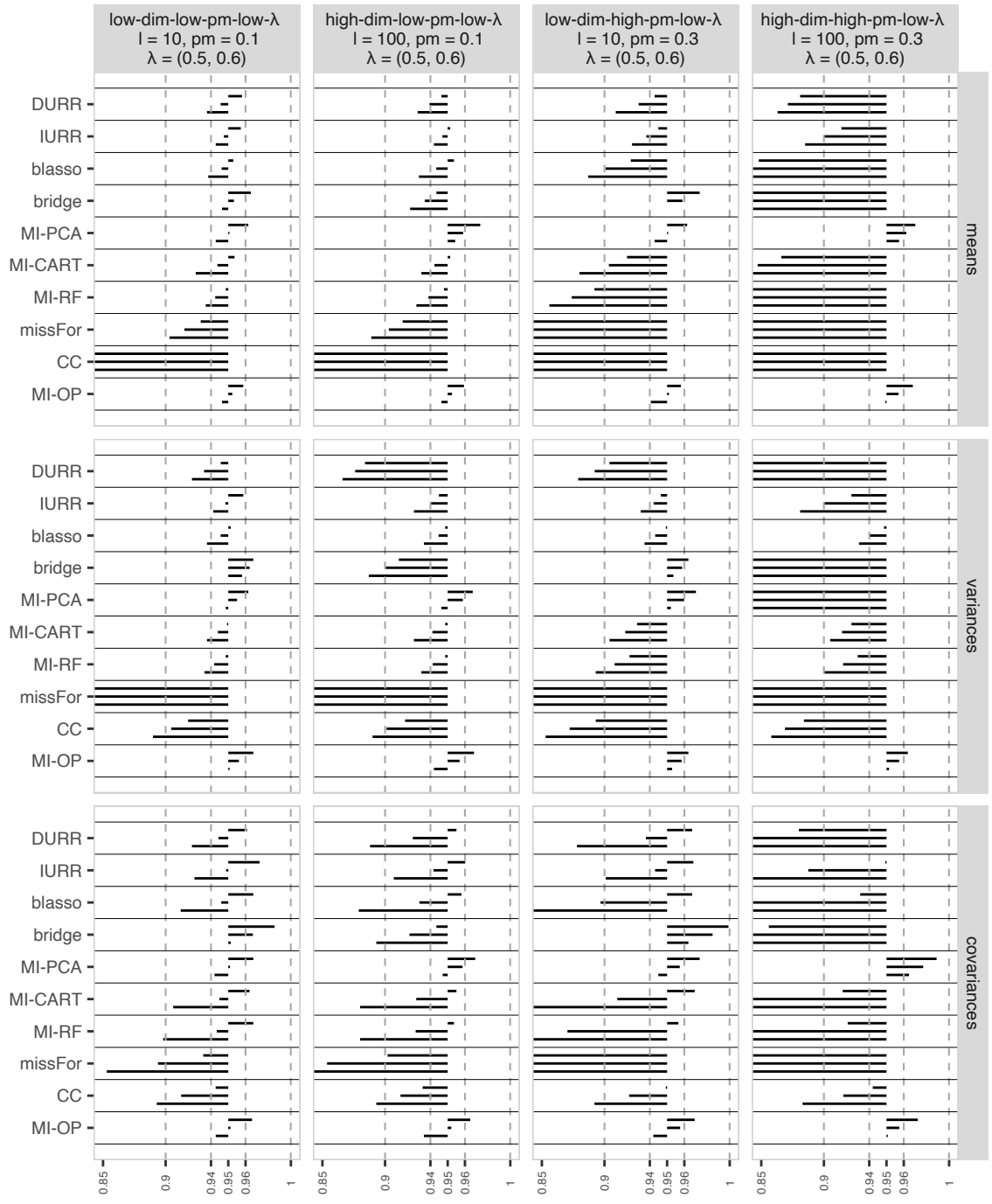

Figure 6. Maximum, average, and minimum CIC for item means, variances, and covariances in the conditions of Experiment 2 with low factor loadings.

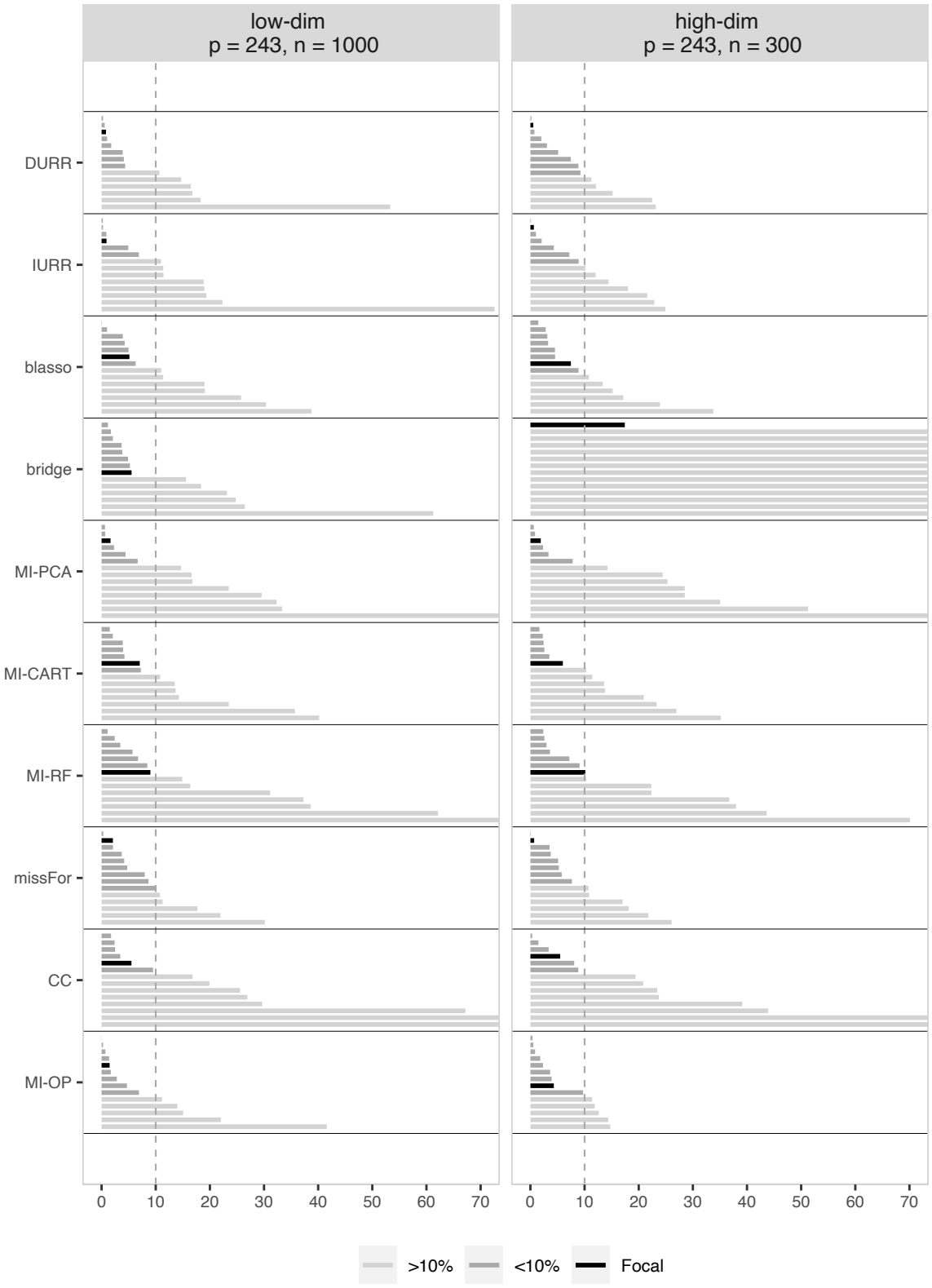

Figure 7. |PRB| for all the model parameters in model 2. For each method, the PRBs are reported by increasing absolute value. The |PRB| for the focal parameter estimate is reported in black.

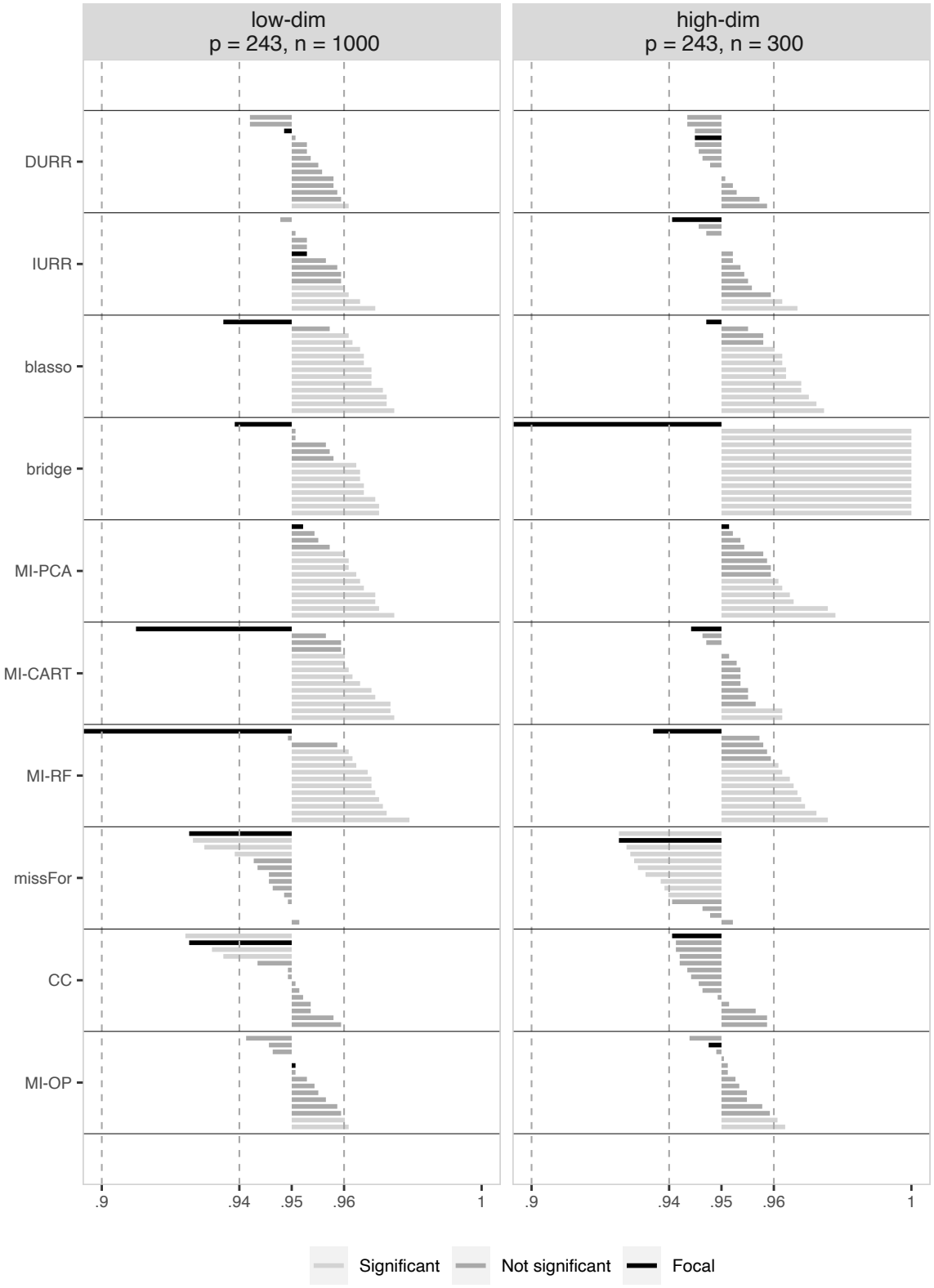

Figure 8. CIC for all model parameters in model 2. For each method, the CICs are reported by increasing value. The CIC of the focal parameter is reported in black.