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

Edoardo Costantini

Kyle M. Lang

Tim Reeskens

Klaas Sijtsma

Tilburg University

Abstract

Your abstract here.

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

Introduction

Today's social, behavioral and medical scientists have access to large multidimensional datasets that can be used to investigate the complex relationships between social, psychological and biological factors in shaping individual and societal outcomes. Large social science datasets, such as the World Values Survey, or the European Values Study (EVS), are easily available to researchers and initiatives have been undertaken to link and extend these datasets into a system of linked open data. Making use of the full potential of these data sets requires dealing with the crucial problem of multivariate missing data.

Rubin's Multiple Imputation (MI) approach (Rubin, 1987) was developed to specifically address the issue of missing responses in surveys. MI is a three-step process that entails an imputation, analysis, and pooling phase. The fundamental idea of the imputation phase is to replace each missing data point with m plausible values sampled from their posterior predictive distributions given the observed data. This procedure leads to the definition of m complete versions of the original data that can be analyzed separately using standard complete-data analysis models (analysis phase). Finally, the m estimates of any parameter of interest can be pooled following Rubin's rules (Rubin, 1987) (pooling phase).

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 Equation (MICE). (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). The first one relies on defining a multivariate distribution for the missing data, deriving conditional distributions for each missing data pattern, and obtaining samples by means of a Markov Chain Monte Carlo algorithm. The second method defines conditional densities for each incomplete variable and performs iterative imputations on a variable-by-variable basis.

When applied to large multidimensional datasets, there are at least two reasons to

prefer the FCS approach over the JM. First of all, the complexity of the FCS method increases with the number of variables with missing values, while the complexity of the JM approach increases with the number of missing data patterns that manifest in a dataset. Given a number of p variables, the FCS approach needs to specify at most p imputation models, while the JM approach requires an imputation model for each missing data pattern. Furthermore, the FCS approach allows a great level of flexibility to accommodate the peculiarities of surveys and other large multidimensional datasets. FCS can easily accommodate for the different distributions of variables and it can preserve unique features in the data, such as skip patterns or variables interactions.

Both the JM and the FCS approaches rely on the crucial *missing at random* (MAR) assumption. Meeting this assumption requires specifying imputation models for the MI procedure that include all observed variables that are correlates of missingness. Omitting from the imputation models an observed predictor related to both the missingness and the imputed variables creates a *missing not at random* (MNAR) problem. MI under MNAR leads to substantial bias in parameter estimation in the analysis and pooling phases, and invalidates hypothesis testing involving the imputed variables.

As a result, when it comes to defining the set of auxiliary variables for the imputation models within an MI procedure, an inclusive strategy (i.e. including numerous 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, making the MAR assumption more plausible. Furthermore, the inclusive strategy has been shown to reduce estimation bias and increase efficiency (Collins, Schafer, & Kam, 2001), as well as reducing 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 matrices within the imputation algorithm. When data 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. one or more of the variables is equal to a linear combination of the others) the data covariance matrix is singular. Singular matrices are not invertible, an operation that is fundamental in the estimation of the imputation models in any parametric imputation procedure. As a result, the possible high dimensionality of the observed data matrix, resulting from an inclusive strategy, can prevent a straightforward application of MI algorithms, such as MICE (van Buuren, 2012), or force researchers to make arbitrary choices regarding which variables to use.

Recent developments in high-dimensional data multiple imputation techniques represent interesting opportunities to embrace an inclusive strategy, without facing its downsides. Some statisticians and machine learning experts have focused on high-dimensional single imputation methods in an effort to improve the accuracy of 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 of observed data points, and single imputation is simply inadequate for this purpose: it does not provide statistically valid inference (Rubin, 1996). The concept of statistical validity as defined by (Rubin, 1996) is meant to capture two features of estimation. First, the point estimate of a parameter of interest must be unbiased, and second, the actual confidence interval coverage (CIC) must be equal or greater than nominal coverage. Single imputation strategies might meet the first requirement, but cannot meet the second as they do not take into account the uncertainty regarding the imputed values. Multiple Imputation, 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 combining full conditional specification of imputation models with shrinkage methods (Deng, Chang, Ido, & Long, 2016; Zhao & Long, 2016), but their application has been studied only for biomedical sciences. Other researchers have proposed FCS strategies using dimensionality reduction to avoid the obstacles of an

inclusive strategy in high-dimensional data imputation. However, these solutions were either limited to the Joint Modeling approach (Song & Belin, 2004), or tested exclusively on particularly low-dimensional settings (Howard, Rhemtulla, & Little, 2015). Finally, tree-based FCS strategies have the potential to overcome the limitations of inclusive strategies. The non-parametric nature of decision trees bypasses the identification issues most parametric methods face in high-dimensional contexts. However, these methods have been proposed to deal with other issues, such as imputation in the presence of interaction effects (Doove, Van Buuren, & Dusseldorp, 2014), or have been tested exclusively on biomedical datasets (A. D. Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014).

Scope. The inclusion of shrinkage methods, principal component analysis and non-parametric decision trees within the FCS framework have the potential of simplifying the decisions social scientists need to make when dealing with missing values. The lack of comparative research on these methods performances makes it difficult for social scientiest working with large multidimensional data sets to decide which imputation method to adopt. With this article, we provide a comparison of these state-of-the-art high-dimensional imputation algorithms. We compared the imputation methods based on the statistical validity of the complete-data analyses they allow to perform. The comparison was based on two simulation studies and a resampling study using real survey data.

Outline. This paper is organized as follows. Section 2 presents the general MICE framework, how the high-dimensional MI methods fit within it, and some single data missing data strategies considered for reference. Section 3 presents the two simulation studies, their design and results. Section 4 presents the resampling study performed on real survey data. Section 5 discusses the implication of the results of the simulation and resampling studies. Finally, section 6 provides concluding remarks, a description of the limitations of the study, and future research directions we want to take.

Imputation methods and Algorithms

Consider a dataset $\boldsymbol{Z}$ of dimensionality $n \times p$, with n observations (rows) and p variables (columns). Assume that the first t ($t \leq p$) variables of $\boldsymbol{Z}$ have missing values. These t variables are part of some substantive model of scientific interest (e.g. a linear regression model), and are target of imputation. The subset of $\boldsymbol{Z}$ containing variables z_1 to z_t is referred to as the $n \times t$ matrix $\boldsymbol{T}$. The remaining $n \times (p-t)$ subset of $\boldsymbol{Z}$ contains variables that are not target 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 $\boldsymbol{Z} = (T, A)$. For a given z_j variable, with $j = (1, ..., p)$, denote its observed and missing components by $z_{j,obs}$ and $z_{j,mis}$, respectively. Let $\boldsymbol{Z}_{-j} = (z_1, ..., z_{j-1}, z_{j+1}, ..., z_p)$ be the collection of $p-1$ variables in $\boldsymbol{Z}$ excluding z_j. Denote $\boldsymbol{Z}_{-j,obs}$ and $\boldsymbol{Z}_{-j,mis}$ the components of $\boldsymbol{Z}_{-j}$ corresponding to the data units in $z_{j,obs}$ and $z_{j,mis}$, respectively.

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 $P(z_1|\boldsymbol{Z}_{-1}, \boldsymbol{\theta}_1)$ ... $P(z_t|\boldsymbol{Z}_{-t}, \boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_1$ ... $\boldsymbol{\theta}_t$ are imputation model parameters specific to the conditional densities of each variable with missing values.

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

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

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

where $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ are draws from the parameters full conditional posterior distribution (1) and the missing data posterior predictive distribution (2), respectively. After convergence, D different sets of values sampled from the predictive distribution are

kept as imputations and D differently imputed data sets are obtained. Any substantive model can then be fit to each dataset, and estimates can be pooled appropriately using Rubin's rules (Rubin, 1987).

Generally speaking, for each variable z_j target of imputation, a researcher needs to define a set of observed variables that will be included in $\boldsymbol{Z}_{-j}^{(m)}$. The high-dimensional imputation methods compared in this paper and described below follow the general MICE framework, but they differ in the elementary imputation methods they use to define equation (1) and (2). Each of them has a different way of processing the large number of auxiliary variables provided to the imputation algorithm to allow a maximal inclusive strategy while avoiding its usual obstacles.

MICE with fixed ridge penalty (bridge). This approach uses as elementary imputation method the Bayesian imputation under the normal linear model procedure as presented by van Buuren (2012) (p. 68, algorithm 3.1).

In this approach, the sampling of each $\hat{\boldsymbol{\theta}}_j^{(m)}$ in (1) relies on the inversion of the cross-product of the observed data matrix $\boldsymbol{Z}_{j,obs}^{(m)}$. By adding a biasing ridge penalty κ, singularity of the cross-product matrix is circumvented and the sampling scheme is possible even if $\boldsymbol{Z}_{j,obs}^{(m)}$ is afflicted by high collinearity and n is not substantially larger than p.

The value of κ is usually chosen close to zero (e.g. $\kappa = 0.0001$), as values larger than 0.1 may introduce systematic bias (Van Buuren, 2018, p. 68). However, larger values may be necessary to invert the observed data matrix cross-product in certain scenarios. In the present work, the value of κ was decided by means of a cross-validation procedure described below.

MICE with Bayesian lasso (blasso). A high-dimensional Bayesian lasso imputation algorithm was proposed by Zhao and Long (2016), but it was tested only in a univariate missing data context. The method relies on the Bayesian lasso model, a regular Bayesian multiple regression with prior specifications that allow to interpret the mode of the posterior distribution of the regression coefficients as lasso estimates (Hans, 2009; Park & Casella, 2008). Given data with sample size n, consider the dependent

variable y and a set of predictors X. The Bayesian Lasso linear regression specification we used within the blasso imputation algorithm is that specified by Hans (2010b):

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

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

$$\exp\left(\frac{-\tau\left\|\beta\right\|_1}{\sigma}\right) \tag{5}$$

$$\sigma^2 \sim \text{Inverse-Gamma}(a, b) \tag{6}$$

$$\tau \sim \text{Gamma}(r, s) \tag{7}$$

$$\rho \sim \text{Beta}(g, h) \tag{8}$$

The expression in equation (3) represents the density function of a multivariate normal random variable with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, evaluated at y. The prior expressed in equation (5) is the expansion on the Park and Casella (2008) double exponential prior developed by Hans (2010b) to accommodate for uncertainty regarding the value of the regression coefficients and the model sparsity. Finally, equations (6) to (8) represent hyper priors for the residual variance σ^2, the penalty parameter τ, and the sparsity parameter ρ. The blasso imputation algorithm used here is a standard MI MCMC sampler that replaces equation (1) with the full conditional posterior distributions derived by Hans (2010b), based of the prior specifications in equations (6) to (8), and uses posterior parameters draws to sample plausible values from the predictive distributions of the missing data for equation (2).

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

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

MICE framework the following two steps:

- 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 considered as a sample from equation (1).

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

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

- Fit a multiple linear regression model using a regularized regression method with $\boldsymbol{z}_{j,obs}$ as dependent variable and $\boldsymbol{Z}_{-j,obs}^{(m)}$ as predictors (compared to DURR, the original data are used, not a bootstrap sample). In this model, the regression coefficients that are *not* shrunk to 0 identify the active set of variables that will be used as predictors in the actual imputation model.

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

$$(\hat{\boldsymbol{\theta}}_j^{(m)}, \hat{\sigma}_j^{(m)}) \sim N(\hat{\boldsymbol{\theta}}_{MLE}^{(m)}, \hat{\boldsymbol{\Sigma}}_{MLE}^{(m)}) \tag{9}$$

  so that equation (9) corresponds to equation (1) in the general MICE framework.

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

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

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

Note that if missing values are present in the set of auxiliary variables, one can fill them in with a stochastic single imputation (SI) algorithm of choice. MI is preferred to SI because it accounts for the uncertainty regarding the missing values when producing standard errors. As the extraction of PCs does not require the estimation of standard errors, SI suffices. This method is inspired by Howard et al. (2015) and the *PcAux* R-package (Lang, Little, & PcAux Development Team, 2018) that implements and developed its ideas.

MICE with regression trees (MI-CART and MI-RANF). The MI-CART imputation method (Burgette & Reiter, 2010) is a MICE algorithm that uses classification and regression trees (CART) to define the conditional distributions used in the MI Gibbs sampler. Given an outcome variable y and a set of predictors X, CART is a nonparametric recursive partitioning technique that models the relationship between y and X by sequentially splitting observations in subsets of units with relatively homogeneous y values. At every splitting stage, a CART algorithm searches through all predictor variables in X to find the best binary partitioning rule to predict y/minimize a homogeneity criterion. The collection of binary splits can be visually represented by a

decision tree structure where each terminal node (or *leaf*) represents the conditional distribution of y for units that satisfy the splitting rules.

In MI-CART, at the m-th iteration for a target variable z_j, a CART model to predict $z_{j,obs}$ based on $\boldsymbol{Z}^{(m)}_{-j,obs}$ is trained. For each missing value, a draw form the predictive distribution is taken by sampling from the elements of the terminal node where the unit of interest belongs to based on $\boldsymbol{Z}^{(m)}_{-j,mis}$. The implementation of MI-CART used in this paper corresponds to the one presented by Doove et al. (2014) (p. 95, algorithm 1) and the *impute.mice.cart()* R function from the *mice* package.

In MI-RANF, at the m-th iteration for a target variable z_j, k bootstrap samples are drawn from the complete dataset and k single trees are fitted to each using a small group of input variables to find the best split at each node. All trees are used to compose the pool of candidates from which imputations are drawn. Bootstrapping and random input selection introduce the model and imputation uncertainty in the imputation procedure, as required by a proper MI procedure. For greater details on the algorithms, the reader may consult Doove et al. (2014) (algorithm A.1, p. 103). The programming of the algorithm was heavily inspired by the *impute.mice.rf()* function in the R package *mice*.

MICE optimal model (MI-OP). When dealing with a large set of possible predictors for the imputation models, a common recommendation in the MI literature is to decide which predictors to include by following three criteria (van Buuren, 2012, p. 168):

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

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, if we knew which predictors were essential for the imputation models, we could use this information to specify optimal imputation models. With simulated data, we have perfect knowledge over which variables are involved in the missing data mechanisms.

MI-OP is an ideal specification of the MICE algorithm that uses as elementary imputation strategy a low dimensional univariate Bayesian imputation under the normal linear model and uses this knowledge to include only the relevant predictors in the imputation models.

Single data strategies

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

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

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

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

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

Experiment 1: Simulated Data from Multivariate Normal Distribution

In the first simulation experiment, we focused on an ideal setting where data come from a known multivariate normal distribution and imputation is required to estimate the mean, variance and covariances of six items with missing values. We investigated the relative performance of the methods described in Section 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 *per* variable missing cases pm, taking values 0.1 or 0.3. Table 1 summarizes the four crossed conditions. Data with sample size $n = 200$ were independently generated 1,000 times for each set of conditions. For each replicate, missing values were imposed and then all the missing data treatment methods described above were used to obtain estimates for the parameters of a substantive analysis model of interest.

Simulation Study Procedure

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

Missing Data Imposition. Missing values were imposed on six items in $\boldsymbol{Z}_{n\times p}$: three variables in the block of highly correlated variables (z_j with $j = 1, 2, 3$), and three in the block of lowly correlated variables (z_j with $j = 6, 7, 8$). Item non-response was

imposed by sampling from a Bernoulli distribution with individual probabilities defined by

$$p_{miss} = p(z_{i,t} = miss|\tilde{Z}) = \frac{exp(\gamma_0 + \tilde{Z}_i\boldsymbol{\gamma})}{1 + exp(\gamma_0 + \tilde{Z}_i\boldsymbol{\gamma})} \tag{10}$$

where $z_{i,j}$ is the i-th subject's response on the j-th variable target of missing data imposition, $\tilde{Z}_i$ is a vector of responses for the i-th individual to the set of predictors involved in the missing data mechanism, γ_0 is the intercept parameter, and $\boldsymbol{\gamma}$ is the vector of slope parameters. $\tilde{Z}$ was specified to include two fully observed variables from the highly correlated set, and two from the lowly correlated set (z_r with $r = 4, 5, 9, 10$). The choice of predictors in $\tilde{Z}$ is important to allow imputations under MAR: the probability of observing a response for a target variable did not depend on the variable itself, to avoid imputation under Missing Not At Random; and, as all features in the data are included in the MI procedures, the predictors in $\tilde{Z}$ are always allowed to be part of the imputation models. 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 a target proportion of missing values and its actual value.

Imputation. Missing values were treated with all the methods described in Section 2. For both experiments, convergence of the imputation models was assessed in a preprocessing step. Before running the actual simulation studies, 10 datasets were generated according to each experimental set up. Missing values in each dataset were imputed by running 5 parallel imputation chains for each Multiple Imputation method. Convergence was checked by plotting the mean of the imputed values for each variable in each stream, against the iteration number. In each parallel run, all the MI algorithms run for 250 iterations. In the simulation experiments, the imputation algorithms were considered to converge after 50 iterations, after which 10 imputed data sets were store and used for the subsequent standard complete-data analysis and pooling. The only exception was blasso, which required approximately 2000 iterations for convergence.

The ridge penalty used in the bridge algorithm is fixed across iterations and its value needs to be decided beforehand by the imputer. The value used in the simulation was determined by means of cross-validation in a pre-processing phase. The grid of possible values for the ridge penalty was $10^{-1}, 10^{-2}, ..., 10^{-8}$. For each of 100 data

repetitions, bridge imputation was performed with each of the different penalty parameters and used to obtain 10 differently imputed datasets. For each data replication, the Fraction of Missing Information (FMI) (Savalei & Rhemtulla, 2012) associated with each parameter in the analysis models of interest (see next section for details) was computed and then averaged across repetitions. The mean of these average parameter FMIs was used as a composite measure of FMI associated with each ridge penalty value. Finally, the penalty value with the smallest composite FMI was selected.

Both IURR and DURR could have been specified with a variety of penalty parameters. For example, one could use any of the following: ridge penalty (Hoerl & Kennard, 1970), lasso penalty (Tibshirani, 1996), elastic net penalty (Zou & Hastie, 2005), adaptive lasso (Zou, 2006). In this study we specified the regularization as a lasso penalty 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.

For blasso, in order to maintain consistency with previous research, the hyperparameters in equations (6), (7), and (8) were specified as in Zhao and Long (2016): $(a, b) = (0.1, 0.1)$, $(r, s) = (0.01, 0.01)$, and $(g, h) = (1, 1)$. In the MI-PCA algorithm, enough components were extracted to explain 50% of the total variance in the data. To impute data with the single imputation random forest approach we used the function *missForest* in the homonymous R package. This function implements algorithm 1 proposed by Stekhoven and Bühlmann (2011). The stopping criterion for the missForest algorithm was usually met within the first 10 iterations, but to make a conservative choice we fixed the maximum number of iterations to 20. Stekhoven and Bühlmann (2011) showed that increasing the number of trees grown in each forest has stagnating effects on the imputation error while linearly increasing the computation time. In their paper, the authors recommend growing 100 trees per forest, which offers a good compromise between imputation precision and computation time. Therefore, we used this value in our study.

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

values. This resulted in estimating six means, six variances, and 15 covariances.

Comparison Criteria

We compared methods in terms of bias an confidence interval coverage.

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 procedures:

$$PRB = \frac{\bar{\hat{\theta}} - \dot{\theta}}{\dot{\theta}} \times 100 \tag{11}$$

where $\dot{\theta}$ is the *true* value of the focal parameter computed as $\sum_{r=1}^{R} \hat{\theta}_r^{GS}/R$, with $\hat{\theta}_r^{GS}$ being the Gold Standard parameter estimate for the r-th repetition. The averaged focal parameter estimate under a given imputation method is computed as $\bar{\hat{\theta}} = \sum_{r=1}^{R} \hat{\theta}_r/R$, with $\hat{\theta}_r$ being the estimate obtained after using a given imputation approach in the r-th repetition. Following Muthén, Kaplan, and Hollis (1987), |PRB| $> 10\%$ was considered indicative of problematic estimation bias.

Confidence Intervals Coverage. To assess the correctness of hypothesis testing, the Confidence Interval Coverage (CIC) of the reference value was defined as

$$CIC = \frac{\sum_{r=1}^{R} I(\dot{\theta} \in \widehat{CI}_r)}{R} \tag{12}$$

where $\widehat{CI}_r$ is the confidence interval of the parameter estimate $\hat{\theta}_r$ 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% confidence intervals (Van Buuren, 2018, p. 52) as they imply inflated Type I error rates. A high coverage (e.g., 0.99) may indicate confidence intervals that are too wide, implying that the imputation method leads to conservative inferential conclusions. Therefore, Confidence Intervals were considered to show severe under-coverage (over-coverage) if they are below 0.9 or above 0.99.

Following Burton, Altman, Royston, and Holder (2006), a CIC can be considered as significantly different from the nominal coverage rate if it falls outside two Standard Errors of the nominal coverage probability ($SE_{(p)}$) from the nominal coverage rate. The

standard error of nominal coverage probability is defined as $SE(p) = \sqrt{p(1-p)/R}$, with p indicating the chosen nominal coverage probability. Therefore, for $R = 1000$, 95% CI coverages ($p = .95$) outside the range (0.94, .96) were considered as significantly different from nominal coverage.

Results

Both PRB and CIC were computed for the 27 parameters in the analysis model (6 means, 6 item variances, and 15 covariances). In this discussion we focus on the typical and extreme values of the measures to summarize the information. In Figures 1 and 2, we report the average, minimum, and maximum PRB and CIC achieved by the missing data treatment methods for each parameter parameter type. In the supplementary material, we included figures reporting the PRB and CIC for every parameter estimate.

Means. Focusing first on the item means (top rows), the largest PRB is within 10 percentage points from 0 for all imputation methods. However, looking at relative performances, in all conditions, IURR and MI-PCA resulted in the closet trend to MI-OP, the optimal but unachievable MI approach. Furthermore, the tree-based MI methods, missForest, and CC lead to CICs significantly different from nominal coverage rates, resulting in extreme under-coverage of the true values in all conditions. In the conditions with high proportion of missing values (columns 3 and 4), all methods showed some signs of either under-coverage or over-coverage of the true values, with all CICs outside of the interval (.94, .96). The only exceptions were MI-OP, and MI-PCA which showed non-significant deviations from nominal coverage for almost all estimates, with both the lowest and highest CIC falling within (.94, .96) in all conditions.

Variances. Moving to the item variances (central rows), IURR, blasso, and the MI tree-based methods resulted in the lowest biases across all conditions, even in the high-dim-high-pm condition. These low biases were mostly paired with low deviations from nominal coverage, except for the high-dim-high-pm condition where IURR and the tree-based methods resulted in significant under-coverage of the true item variances (highest $CIC < .94$). Apart from MI-OP, Blasso was the method with best coverage in

this final condition.

DURR showed poor performance with regard to the item variances: in all conditions but the first, it led to large (negative) bias accompanied by significant CI under-coverage. Bridge was the only MI method showing larger bias than DURR in all the high-dimensional conditions (2 and 4), with even the minimum $|PRB|$ exceeding the 20% threshold. MI-PCA also showed poor performance with a noticeable positive bias in all conditions that became extreme in the high-dim-high-pm condition (column 4), where all PRBs exceeded 20%. This poor performance was reflected in extreme confidence interval under-coverage of the true item variances in the final experimental condition. Single data imputation with missForest and complete case analysis led to substantial negative bias and CI under-coverage for all item variances, even in condition 1.

Covariances. Finally, the third row in Figure 1 shows the minimum, maximum, and average covariance estimation bias achieve by each methods. IURR performed noticeably better than most other methods, with negligible negative bias and acceptable coverage for all covariances in conditions 1, 2 and 3, but it struggled with a large negative bias and extreme under-coverage for most of the 15 covariances in the high-dim-high-pm condition (average $|PRB| > 10\%$). MI-PCA showed negligible negative bias for all the covariance estimates (with the maximum $|PRB| < 10\%$), and performed as well as MI-OP in all but the high-dim-high-pm condition. Furthermore, MI-PCA showed virtually no deviation from nominal coverage, with a CIC pattern similar to that of MI-OP, in all but the last condition, where it manifested only mild *over*-coverage of the items covariances.

All other methods, including DURR, showed absolute PRBs larger than the 10% threshold in all but the first condition, with persistently significant CI under-coverage of the true values. Bridge displayed acceptably low biases and coverage in the low dimensional conditions (columns 1 and 3), but extremely large biases and low CI coverage in all the high dimensional conditions (columns 2 and 4). MissForest and CC showed extreme bias and under-coverage for all the covariances (minimum $|PRB| > 10\%$), even in condition 1.

Experiment 2: Simulated Data with Latent Structure

In the second simulation experiment, we focused on data generated from a Factor Analysis model. The data social scientists analyse is often a collection of items measuring different latent constructs, a characteristic that is likely to impact imputation performances. For Experiment 2, we considered three experimental factors. First, the dimensionality of the data was controlled by the number of latent variables l (10, 100). 5 items were generated as measurements of each latent variable, resulting in either 50 and 500 total items. Second, factor loadings λ_{ij} were defined as a 2-level random experimental factor (high, low). High factor loadings were drawn from a uniform distribution between (0.9, 0.97), while low factor loadings were drawn from a uniform distribution between (0.5, 0.6). Third, the proportion of missing values was defined as a fixed experimental factor with two levels: 0.1 or 0.3. Data with sample size $n = 200$ were independently generated 1,000 times for each set of conditions. Table 2 summarizes the eight resulting conditions. On each replicate, missing values were imposed and then each missing data treatment described in Section was used to obtain estimates for the parameters of a substantive analysis model of interest. With a sample size fixed at 200, conditions with $l = 10$ resulted in a low-dimensional settings, while conditions with $l = 100$ resulted in a high-dimensional settings.

Simulation Study Procedure

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

$$\boldsymbol{z}_i = \boldsymbol{\Lambda}\boldsymbol{\xi}_{i.} + \boldsymbol{\delta}_{i.} \tag{13}$$

where $\boldsymbol{z}_i$ is a vector of $5 * l$ observed items scores, for observations $i = 1, ..., n$; $\boldsymbol{\Lambda}$ is the $(5 * l) \times l$ matrix of factor loadings; $\boldsymbol{\xi}_{i.}$ is a vector of l latent scores for observation i; and $\boldsymbol{\delta}_{i.}$ is a vector of $5 * l$ uncorrelated measurement errors sampled from a multivariate

normal distribution centered around a mean vector of 0s and with a diagonal covariance matrix $\mathbf{\Theta}$. All items are centered around a mean of 5. For notation and model specification the interested reader may refer to Bollen (1989).

The latent scores in $\boldsymbol{\xi}_{i\cdot}$ are sampled from a multivariate normal distribution centered around 0, and with a covariance matrix $\mathbf{\Psi}$, 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 weakly correlated ($\rho = .3$), while the remaining $l - 8$ latent variables are uncorrelated.

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

Item non-response was imposed on 10 items in $\boldsymbol{Z}_{n\times p}$. The items targeted by the missing data measured two highly correlated latent variables ($l = 1, 2$). The same strategy described in Subsection was used. The predictors included in $\tilde{Z}$ (see Equation 10) were the latent scores for the other two highly correlated latent variables ($l = 3, 4$).

Missing values were treated according to all the methods described in Section . The imputation methods were parametrized as in Experiments 1. 50 iterations were sufficient for convergence for all MI methods except blasso, which required approximately 2000 iterations for convergence.

The substantive model of interest in Experiment 1 was a saturated model that estimates means, variances, and covariances of the items with missing values. Furthermore, the true Factor Analysis model for the same items was estimated to see how

the factor loadings were recovered after imputation.

Results

The same comparison criteria defined in Subsection were used to assess the performances of the methods. Figures 3 and 4 report the average, minimum, and maximum PRB and CIC obtained with each missing data treatment method for each parameter type (means, variances, and covariances) in the conditions with high factor loadings. Figures 5 and 6 report the same results for the conditions with low factor loadings. In the supplementary material you may find the figures reporting the PRBs and CICs for every parameter.

Means. All the imputation methods provided unbiased estimates of the item means with PRBs that were almost 0 for all items. For high proportion of missing cases (column 3 and 4) there was a slight increase in PRB values for all methods except IURR, bridge, and MI-PCA. However, only Complete Case analysis led to unacceptable bias of the means. DURR, IURR and MI-PCA resulted in little to no deviations from nominal coverage in all conditions, while blasso, MI-CART, MI-RF, bridge, and missForest led to significant under-coverage of the true means when the proportion of missing cases was high (columns 3 and 4).

Variances. All MI methods, except Bridge, resulted in acceptable estimation bias for the item variances in all conditions, but the least biased estimates were obtained by MI-OP, IURR and MI-PCA. CIC decreased as the proportion of missing cases increased (from condition 1 and 2 to conditions 3 and 4). For high pm, only IURR and MI-PCA maintained CICs mostly within the range .94-.96, while blasso and the MI tree-based methods led to mild to extreme under-coverage (all CICs $< 90\%$). Single data approaches, missForest and CC, showed again extreme (negative) bias and CI under-coverage in almost all conditions.

The large positive bias (and low CIC) for the item variances that afflicted MI-PCA in the multivariate-normal set up (Figures 1 and 2) is not present in figures 3 and 4. However, that pattern reappeared when the factor loadings were small, as can be seen in

Figure 5 and 6.

Covariances. For all the conditions with high factor loadings in experiment 2, IURR and DURR showed acceptable covariance biases ($|\text{PRB}| < 10\%$). However, they led to large negative bias in all the conditions with low factor loadings (see Figures 5 and 6). The other methods followed the same pattern as in experiment 1: the MI-PCA approach resulted in the lowest bias and deviation from nominal coverage for the covariances of the observed items; all other methods led to large negative biases and mild-to-extreme under-coverage for all the covariances, in all conditions.

Factor Loadings. Figure 7 shows the average, minimum, and maximum PRB values for all the factor loadings estimated by the Confirmatory Factor Analysis described above. Most MI-Methods provided acceptably low bias for these estimates in all conditions except the one with both large proportion of missing values and high dimensional input data matrix (condition 4 and 8). MI-OP, IURR, and MI-PCA outperformed all other methods and produced virtually unbiased estimates of the factor loadings in all conditions. In particular, MI-PCA outperformed IURR when factor loadings were low (panel b), maintaining inconsequential biases even when data were high-dimensional and the proportion of missing values was high.

Experiment 3: EVS Resampling Study

In the third experiment, we performed a resampling study based on the EVS data to assess the how the result obtained for Experiments 1 and 2 carry over to real data applications. EVS is a large-scale, cross-national survey on human values administered in almost 50 countries across Europe. It covers a wide range of human values regarding family, work, environment, perceptions of life, politics and society, religion and morality, national identity. It is a high-quality survey widely used for comparative studies between European countries. Furthermore, it is accessible free of charge and it represents the type of data social scientist regularly analyse. Variables in the EVS data are discrete numerical and categorical items following a variety of distributions. By using data gathered for an actual survey, we could study whether the relative performances of the

imputation methods, displayed in the simulation studies, changed when deployed for real data research.

Resampling Study Procedure

Data preparation and generation. For this study, we used the third pre-release 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 of the data that were either duplicated information (recoded versions of other variables), or meta data (e.g. time of interview, mode of data collection). All originally missing values were filled in with a run of a single imputation predictive mean matching (PMM) algorithm to obtain a pseudo fully-observed dataset. This imputation step was completed by using the *mice* R package imputation procedure. PMM was chosen for the task as it is a flexible imputation method that maintains the distributional characteristics of the original data. Predictors for the imputation models were selected based on the variable selection procedure described in Van Buuren, Boshuizen, and Knook (1999, pp. 687–688) by using the *quickpred()* R function and setting the minimum correlation threshold to 0.3. The number of iterations was set to 200. This imputation procedure is used to obtain a pseudo-population dataset, and therefore it does not require multiple imputation itself.

At the end of this data cleaning process, we obtained a fully-observed dataset ($\boldsymbol{Z}$) of 8,045 observations (n), across 4 countries, and 243 variables (p). For every repetition of the first step in the resampling procedure, a bootstrap sample $\boldsymbol{Z}^*$ was generated by sampling with replacement n observations from the EVS population data $\boldsymbol{Z}$. We considered two conditions in the resampling study: low and high dimensional data. As the number of predictors in the data was fixed ($p = 243$), the dimensionality of the data was changed by defining different sizes for the sample taken from the pseudo-fully observed data in step 1. We chose two values for n, namely $1,000$ and 300, corresponding to a low and high dimensional condition.

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:

$$y = \beta_0 + \beta_1 x + \boldsymbol{\beta C} \tag{14}$$

where a dependent variable y is regressed on a variable of interest 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.

The first version of linear model (14), Model 1, was inspired by Köneke (2014): $y^{(1)}$, its dependent variable, was a 10-point EVS item measuring euthanasia acceptance ('Can [euthanasia] always be justified, never be justified, or something in between?'); the predictor of interest $x^{(1)}$ was a 4-point item measuring the self-reported importance of religion in one's life; the matrix of covariates $\boldsymbol{C}^{(1)}$ included trust in the health care system, trust in the state, trust in the press, country, sex, age, education, and religious denomination. This model represents a plausible analysis a researcher would perform to test a hypothesis regarding the effect of religiosity on the acceptance of end-of-life treatments.

Model 2, the second version of the linear model in equation 14, was inspired by Immerzeel, Coffé, and Van der Lippe (2015). The dependent variable $y^{(2)}$ was an harmonized variable constructed by EVS to describe the respondents' tendency to vote left or right-wing parties, expressed on a 10-point left-to-right continuum. The predictor of interest $x^{(2)}$ 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 town where interview was conducted. A researcher might fit this model and look at the value and standard error of $\beta_1^{(2)}$, the 'nativist attitude' regression coefficient, to test an hypothesis regarding the effect of xenophobia on voting tendencies.

Missing data imposition. Missing data were imposed on 6 variables according to the same strategy described in Subsection . The variables target of missing value imposition were the euthanasia acceptance item $y^{(1)}$, and the left-to-right voting tendency $y^{(2)}$, the two dependent variables in models 1 and 2; religiosity ($x^{(1)}$, focal and control variable in models 1 and 2 respectively), and the three items making up the "nativist attitudes" scale (focal predictor $x^{(2)}$ in the second model).

The response model form was the same as in Equation (10) and three variables were included in $\tilde{Z}$: age, education, and an item measuring trust in new people. Older people tend to have higher item non-response 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 assumed that people with less trust in strangers have a higher item non-response tendency as they are likely to withhold more information from the interviewer (a stranger).

Imputation. Missing values were treated according to all the methods described in Section 2. The imputation methods were parametrized as in Experiments 1 and 2. Here, we describe the key details in the algorithms specifications.

Convergence. As for the simulation studies, convergence of the imputations was assessed in a pre-processing step. Before running the actual resampling study, 10 datasets were sampled from the EVS populaton data and missing values were imposed. Missing values were then imputed by running 5 parallel imputation chains for each Multiple Imputation method. Convergence was checked by observing mean imputations trace-plots. This procedure was performed only for condition 2, the high-dimensional one, under the assumption that convergence would be faster in the low-dimensional set up. After the convergence check, we decided to run the algorithms for the regular experiment

run for 60 iterations before saving the multiply imputed datasets, although most methods achieved convergence well before that number of iterations.

Results

Single Parameter of interest. Figures 8 and 9 report the PRB and the CIC for parameters $\beta_1^{(1)}$ and $\beta_1^{(2)}$, the focal regression coefficients in the two models. Most of the MI methods resulted in negligible biases ($|PRB| < 10\%$) for both parameters in all conditions. The only two exceptions were bridge and MI-RF. The former was very competitive in the low-dimensional condition, but led to extreme bias and over-coverage in the high-dimensional condition for β_1 in Model 1; the latter provided the largest PRBs and worst CIs (under) coverage for the focal regression coefficients among the other MI methods, and it was consistently outperformed even by Complete Case analysis. missForest also displayed contained focal parameter biases. DURR and IURR gave inconsequential biases for both parameters in all conditions, with PRBs that were often at least half in size as the ones obtained with the other methods, outperforming even MI-OP in the high dimensional condition for β_1 in Model 2. For β_1 in Model 2, both IURR and DURR remained fairly competitive in terms of coverage with CICs close to nominal levels, but the advantage they showed in terms of bias was not carried over to this criterion. Both MI-PCA and Blasso provided CICs either equal or closer to nominal than the ones obtained with IURR and DURR in almost all conditions.

As for $\beta^{(1)}$ in Model 1, all MI methods led to under-coverage of the true parameter values with CIC smaller than the threshold value 0.94. Although coverage of the true values was not particularly good for any of the methods selected, the Gold Standard confidence intervals were also under-covering the true value of β_1 in Model 1. This was likely due to the right-skewed nature of the distribution of the dependent variable (euthanasia acceptance). Most MI imputation methods achieved coverages similar to that of the Gold Standard method, and more importantly their relative difference, compared to the GS coverages, was in line with what can be seen for β_1 in Model 2.

Overall Model Parameter Assessment. The estimated linear models were multiple regressions, and all regression coefficient estimates were influenced by the imputations. Therefore, it is important to observe the overall effects of the different methods on model estimation. Figure 10 reports the absolute values of the PRBs for every Model 2 parameter estimate, ordered by size, under each of the different imputation methods. (see figures **??** and **??** in the Appendix for Model 1 results). MI-OP showed that even having perfect information regarding the missing data mechanism and data structure, results in some bias for certain estimates. Although the bias for the intercept and the focal parameter were negligible, around half of the estimates obtained after using this imputation method showed large biases ($|PRB| > 10\%$), and the largest bias was considerable (around 40%, in the low dimensional condition, and 20%, in the high-dimensional one). In both the high- and low-dimensional conditions, MI using DURR, IURR, Blasso, and MI-CART, and SI using missForest showed fairly similar overall patterns to MI-OP, with only slightly larger PRBs. MI-PCA and MI-RF showed similar trends but they presented overall larger PRBs for those estimates exceeding the 10% threshold. However, none of these methods seemed to suffer from the increase in dimensionality. Bridge demonstrated the same results described in the simulation studies. It was a competitive method in low dimensional scenarios, but it was inadequate to deal with high-dimensional data imputation (all but one PRB are larger than 100%).

Figure 11 reports the CIC for each parameter estimate in model 2. When using MI-OP, CICs showed a deviation from nominal coverage for only two parameters, with a slight tendency toward over-coverage. While DURR, IURR, MI-CART and MI-PCA maintained a similar coverage pattern to MI-OP, blasso, MI-RANF, and missforest were either over- or under-covering many more parameters. The results showed by missForest were as expected. A single imputation approach, it underestimates uncertainty regarding values of the empty data cells, and tends to produce narrower confidence intervals. Despite showing poor performances in terms of bias, Complete Case analysis manifested good coverage. However, this was a result of the smaller sample size used for 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.

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 the hour in our low-dimensional conditions. MI-PCA and Blasso imputation 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 required more than ten times the time of MI-PCA and blasso imputation.

Discussion

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

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

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

The performances showed by DURR and IURR come at a large computational cost:

in our resampling study set up, they took significantly more time to perform on average than all other methods.

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

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

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

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

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

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

MI-PCA showed poor performances in terms of estimation bias of the item variances. In the high-*pm*-high-dimensionality condition of experiment 1, MI-PCA led to item variance PRBs larger than 20%. The bias for the item variances that afflicted MI-PCA in the multivariate-normal set up appeared to be related to the strength of the latent structure: when the latent structure was absent (experiment 1) or weak (experiment 2, conditions 5 to 8, factor loadings between .5 and .6) item variances were biased, especially in the high-dimensional conditions; when the latent structure was prominent (experiment 2, conditions 1 to 4), the variances were estimated with negligible bias, even in the high-dimensional conditions. One possible explanation for this is related to the CFA factor loadings specification. When data was generated with a CFA model with factor loadings close to 1 (experiment 2, conditions 1 to 4), variables measuring the same latent constructs became highly correlated. In the resampling study, there were 10 variables with missing values, measuring 2 latent constructs. These items were directly included in the imputation models, and not part of the set of auxiliary variables from which the Principal Components were extracted. So while the PC extraction picked up the correlates of missingness in the auxiliary set, the high correlation between the variables directly included in the imputation models improved the accuracy of the imputation and reduced the item variance estimation bias.

MI-PCA performed acceptably in the resampling study, with sufficiently low biases and close to nominal CI coverage for the focal parameters and the overall model parameter assessment. However, the great recovery of bivariate relationships manifested in experiment 1 and 2 (low covariance bias) did not directly translate in particularly low biases for regression coefficients in the resampling study. PCA is a tool to find a low-dimensional representation of a data set summarizing in a few components as much unique variation on each dimension of the data as possible. As such, PC extraction is

likely negatively affected by the nature of survey data where items are usually discrete, and their possible values are only a few integer values in relatively small ranges. This key aspect might be the root of the different performances obtained by MI-PCA in the simulation and the resampling study.

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

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

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

The use of Random Forests within a MICE algorithm could have been implemented differently. We decided to use Doove et al. (2014)'s versions as they are the ones implemented in the popular *mice* R package, while other versions are not currently supported by active R packages. For example, A. D. Shah et al. (2014) independently developed another integration of Random Forests within the MICE algorithm, which was available in the now archived R package *CALIBERrfimpute* (A. Shah, 2018). We are not aware of any evidence or theoretical reason motivating substantial differences in how the two methods would perform, but we did not verify this empirically.

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

the imputation procedure, which is to be expected from a single imputation approach.

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

Conclusions

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

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

Finally, the use of PCA to reduce the dimensionality of the data, as a pre-processing step followed by regular low-dimensional MICE imputation strategies, proved to be a fast and effective approach. It was especially effective in preserving relationships between variables with missing values: in simulation study 1 and 2, it

returned negligible covariance biases when all other methods failed to. However, its application on real survey data and the need for improvements to imputation accuracy suggest further research is needed.

The inclusion of State-of-the-art modern regression techniques within the MICE framework has been developed in recent years allowing imputers to include all available predictors in imputation models. These methods had not been compared among themselves and not all of them had been tested with general multivariate missing data patterns. Our research fills this gap and provides initial insights into applying such methods in social research. [MAKE THIS PARAGRAPH BETTER, ON THE LINE OF ZHAO LONG CONCLUSION]

Limitations and future directions. As this work aimed at comparing current implementations of different methods, some limitations to the scope of the simulation and resampling studies were imposed by the current state of development of the different methods. For example, both IURR/DURR and MI-PCA allow imputation of any type of data: IURR and DURR have been developed for categorical data imputation (Deng et al., 2016), and MI-PCA can be performed with any standard imputation model for categorical data. However, blasso has not been formally developed for multi-categorical imputation target variables yet, which limited the study to working with missing values on variables that are either continuous in nature or usually considered as such in practice. IURR, DURR and MI-PCA could have performed better had they been used in their ordered categorical data implementations, but to maintain a fair comparison ground with blasso, they were implemented with the assumption that the imputed variables are continuous and normally distributed.

Furthermore, in real survey data, the missing data mechanism might be non-linear, which would require including interactions between auxiliary variables and polynomial terms in the imputation models, and would lead to increase dimensionality even more. This factor was not taken into consideration as the inclusion of interactions and squared terms in the imputation models has as not been developed to the same extent across the different methods. However, all of the high-dimensional imputation methods considered

have great potential to allow the specification of much more complex response mechanisms than traditional methods ones.

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

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

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. doi: 10.1093/aje/kwq260

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. doi: 10.1037//1082-989X.6.4.330

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. doi: 10.1007/s00357-012-9108-1

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. doi: 10.1038/srep21689

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.

doi: 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.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

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. doi: 10.1080/00273171.2014.999267

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. doi: 10.1093/bioinformatics/bth499

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. Retrieved from `https://doi.org/10.5281/zenodo.3509134` doi: 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. (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.

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

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

Shah, A. (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. doi: 10.1002/sim.1867

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. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press. doi: 10.1201/b11826

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. doi: 10.1177/0962280213511027

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.

condition	n	p	pm
1	200	50	.1
2	200	500	.1
3	200	50	.3
4	200	500	.3

Table 1

Summary of conditions for experiment 1.

condition	n	p	l	pm	λ range
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 2

Summary of conditions for experiment 2.

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

Table 3

Average imputation time in minutes.

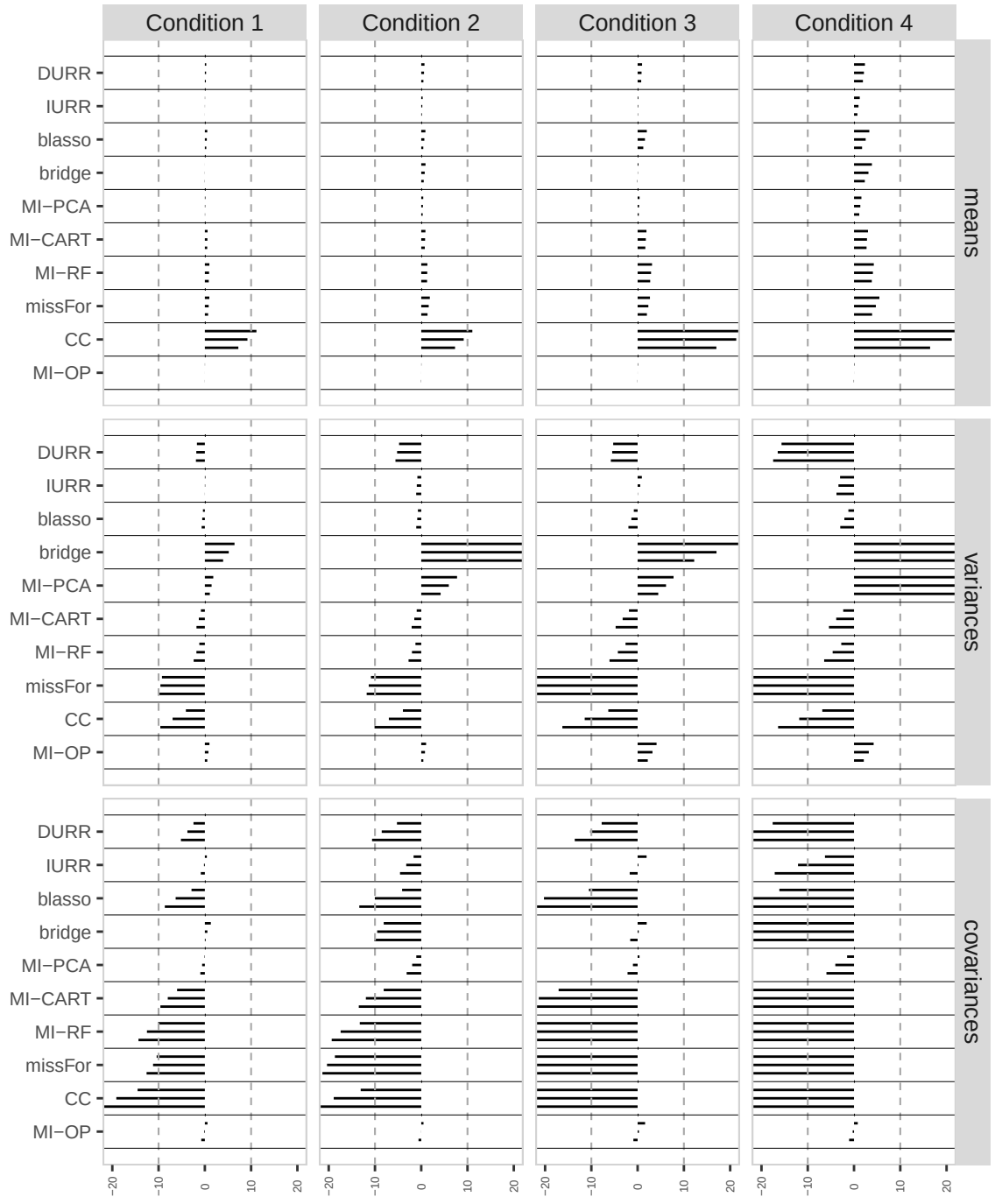

Figure 1. Percent Relative Bias (PRB) for item means, variances, and covariances. For every method, single horizontal lines, representing the PRB of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

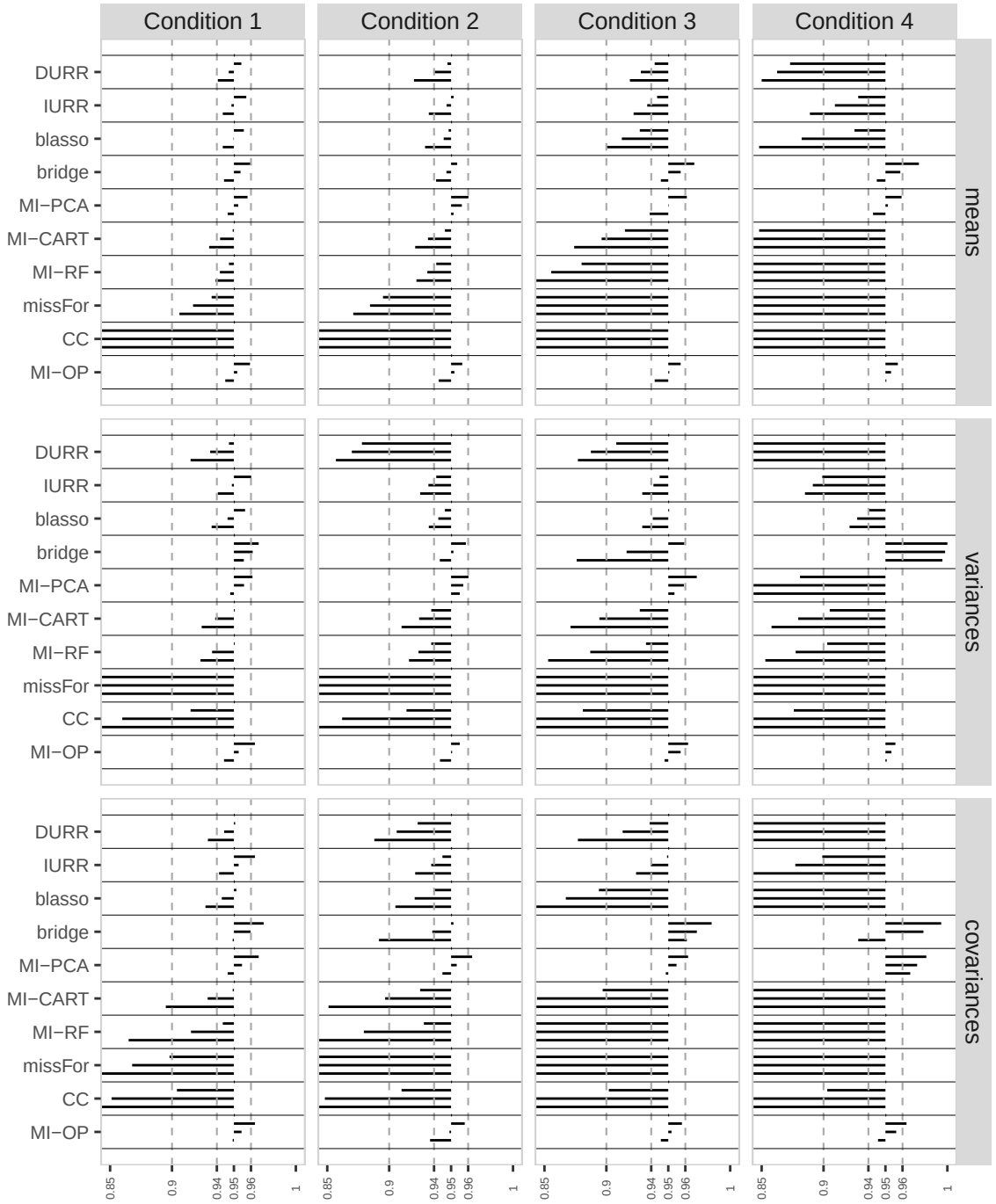

Figure 2. Confidence Interval Coverage (CIC) for item means, variances, and covariances. For every method, single horizontal lines, representing the CIC of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

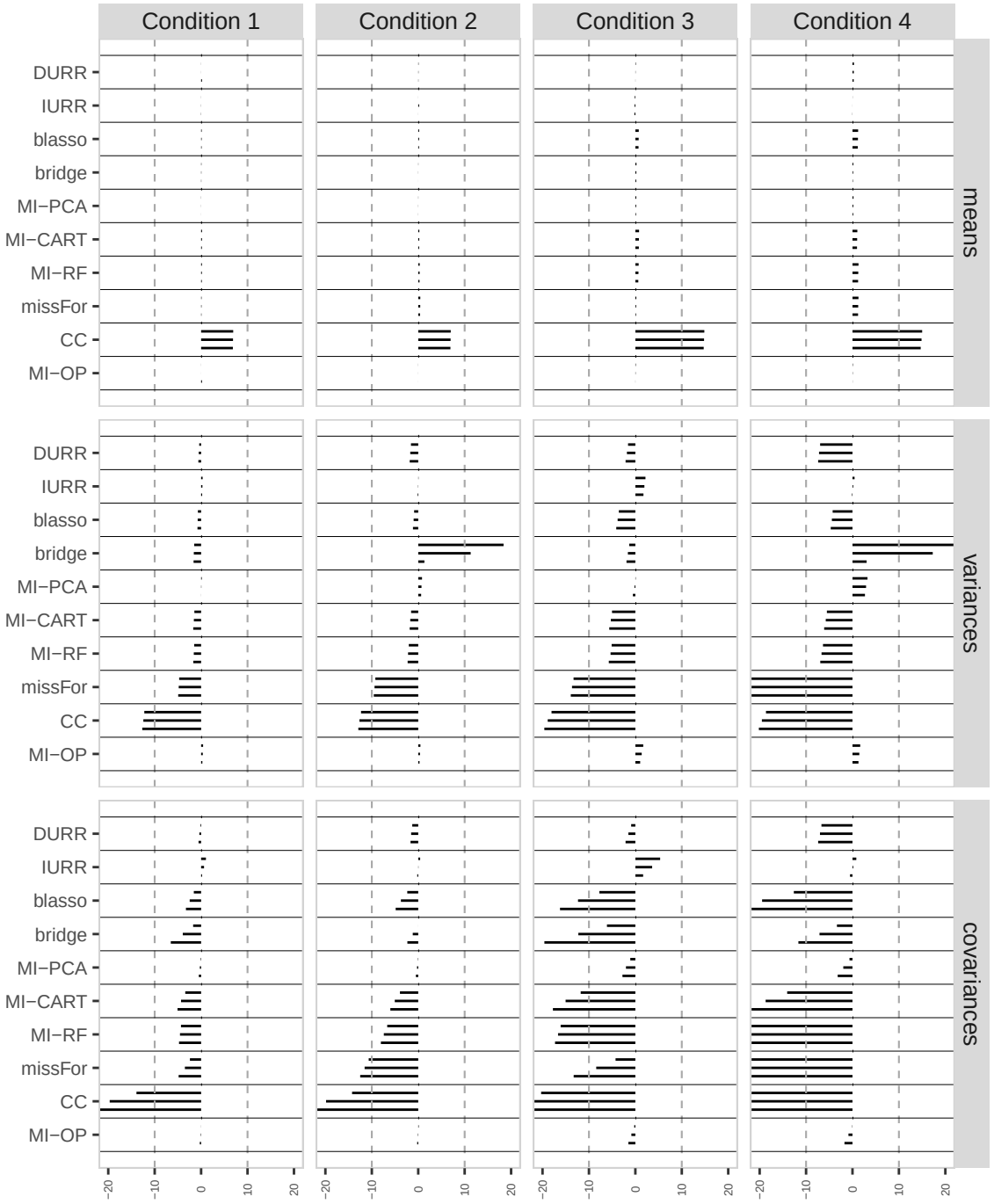

Figure 3. PRBs for the means, variances and covariances (PRB) for condition 1 to 4. For every method, single horizontal lines, representing the PRB of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

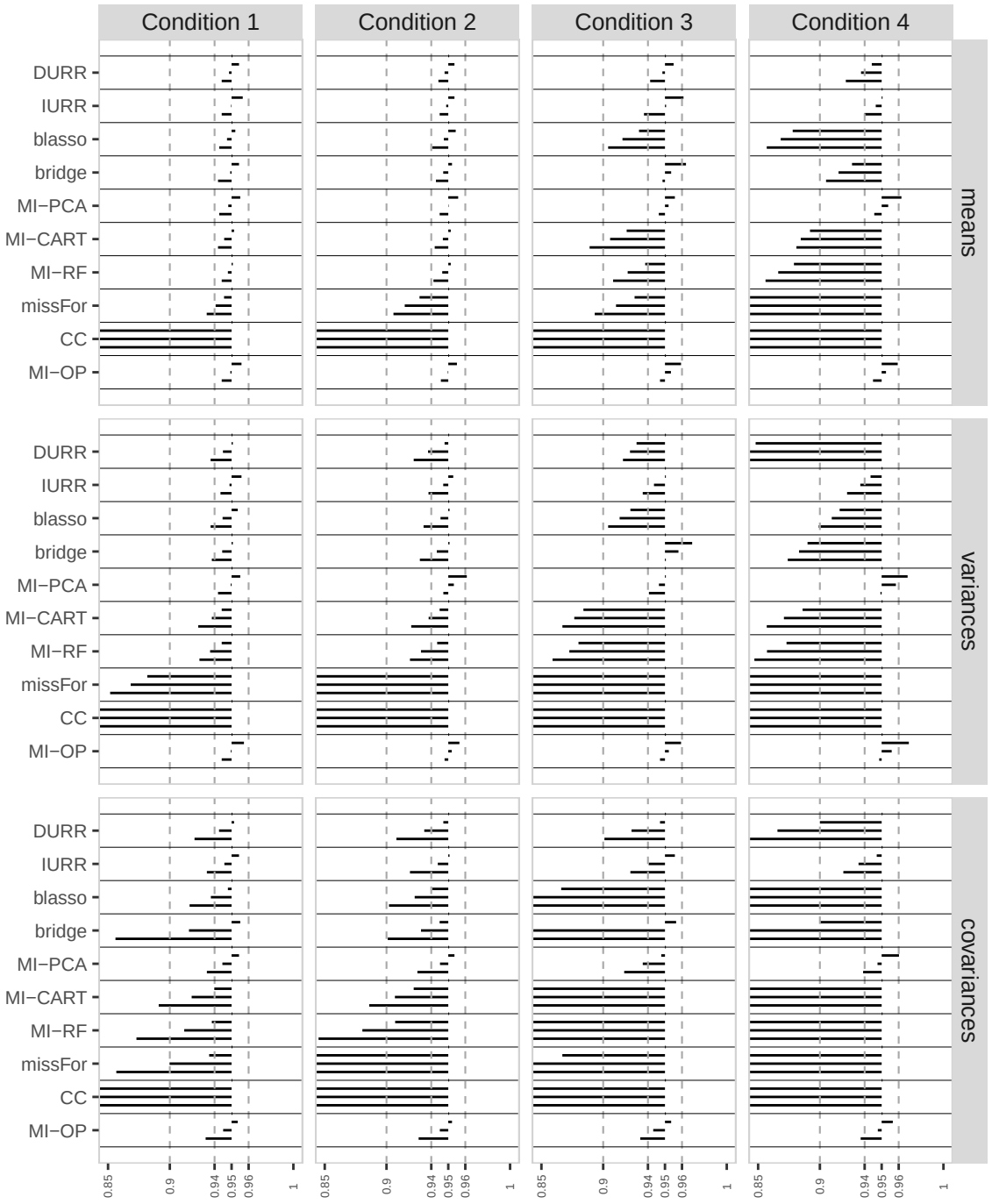

Figure 4. CIC for the means, variances, and covariances for condition 1 to 4. For every method, single horizontal lines, representing the CIC of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

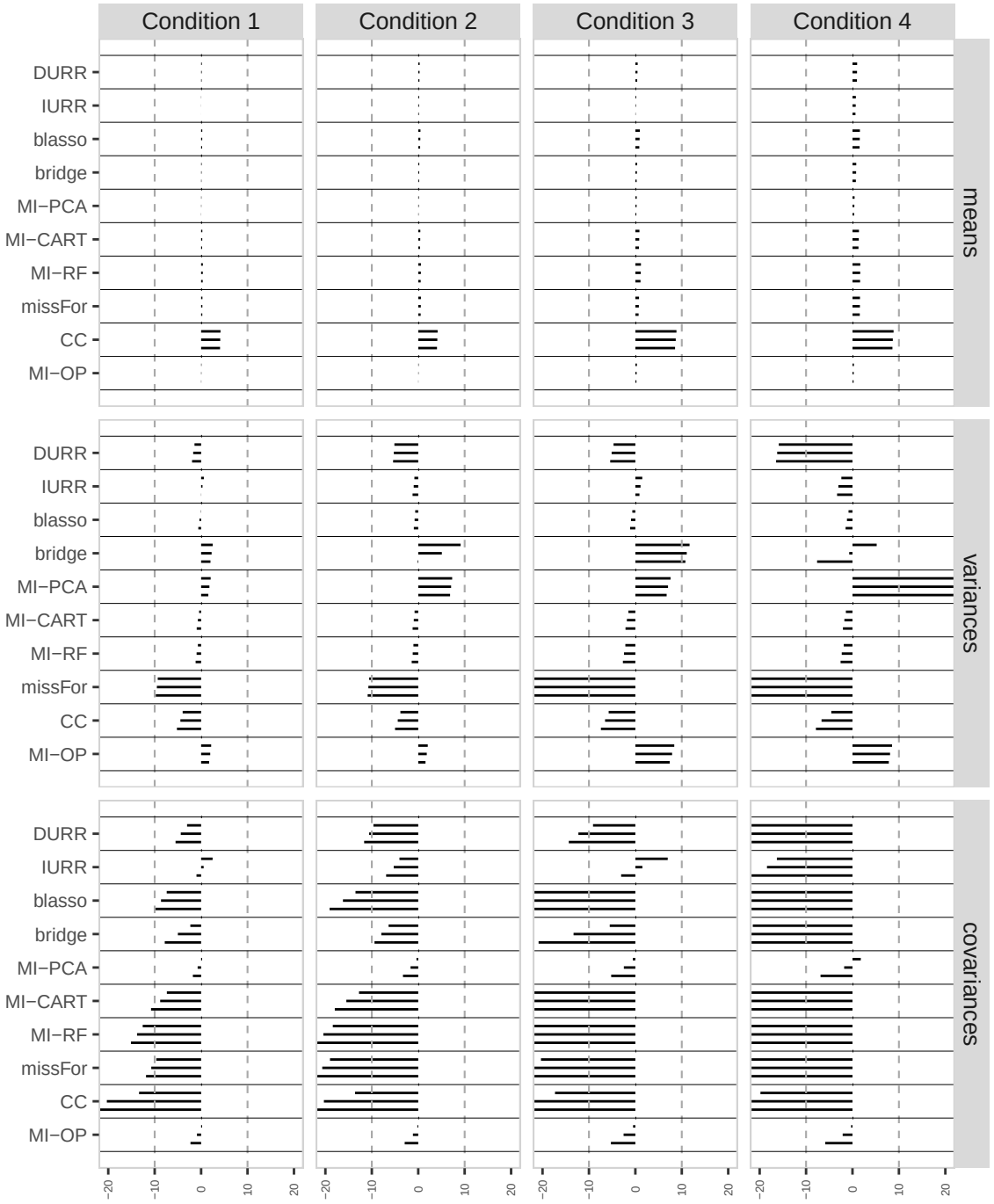

Figure 5. PRBs for the means, variances and covariances (PRB) for condition 1 to 4. For every method, single horizontal lines, representing the PRB of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

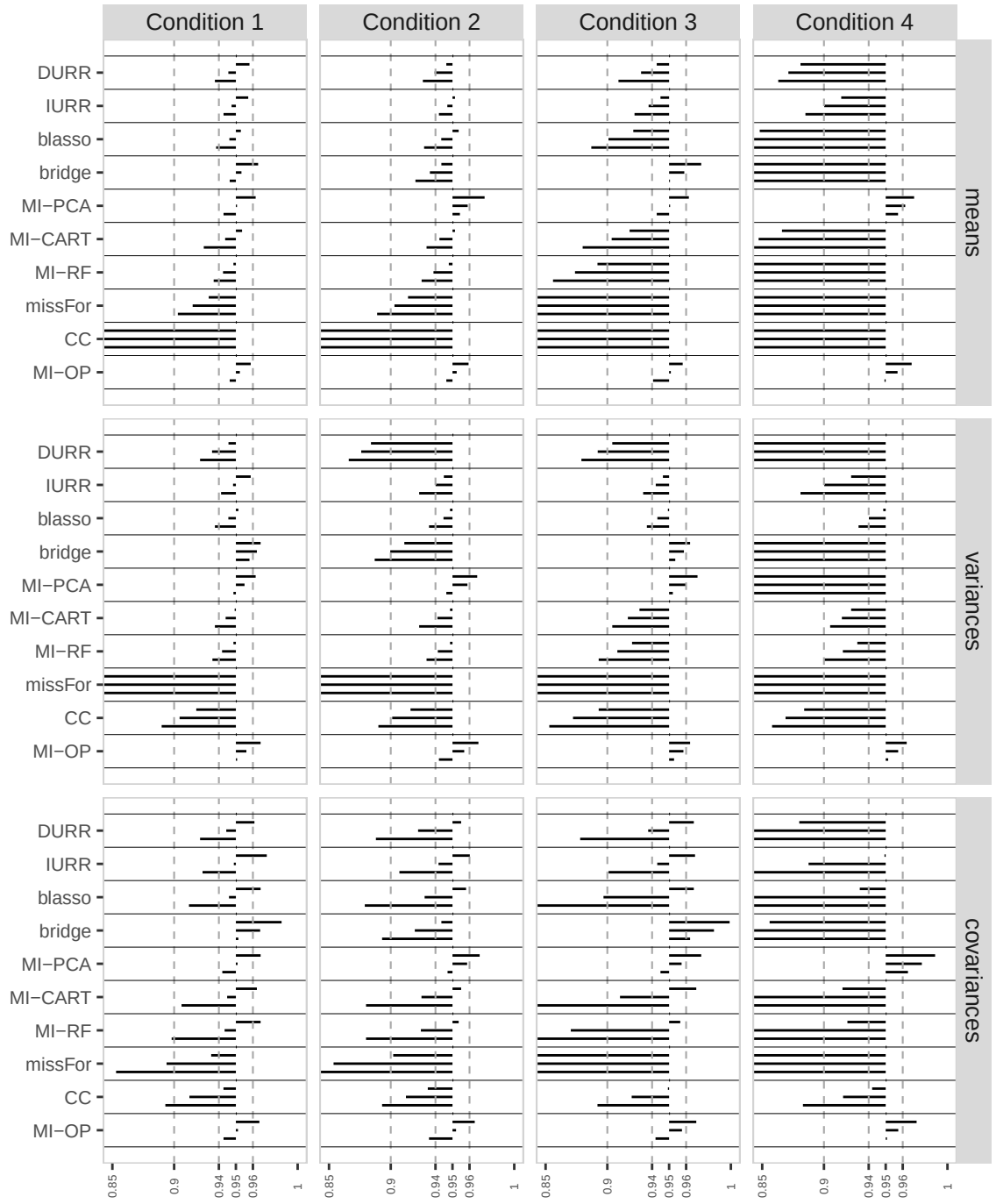

Figure 6. CIC for the means, variances, and covariances for condition 1 to 4. For every method, single horizontal lines, representing the CIC of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed across multiple variables with missing values.

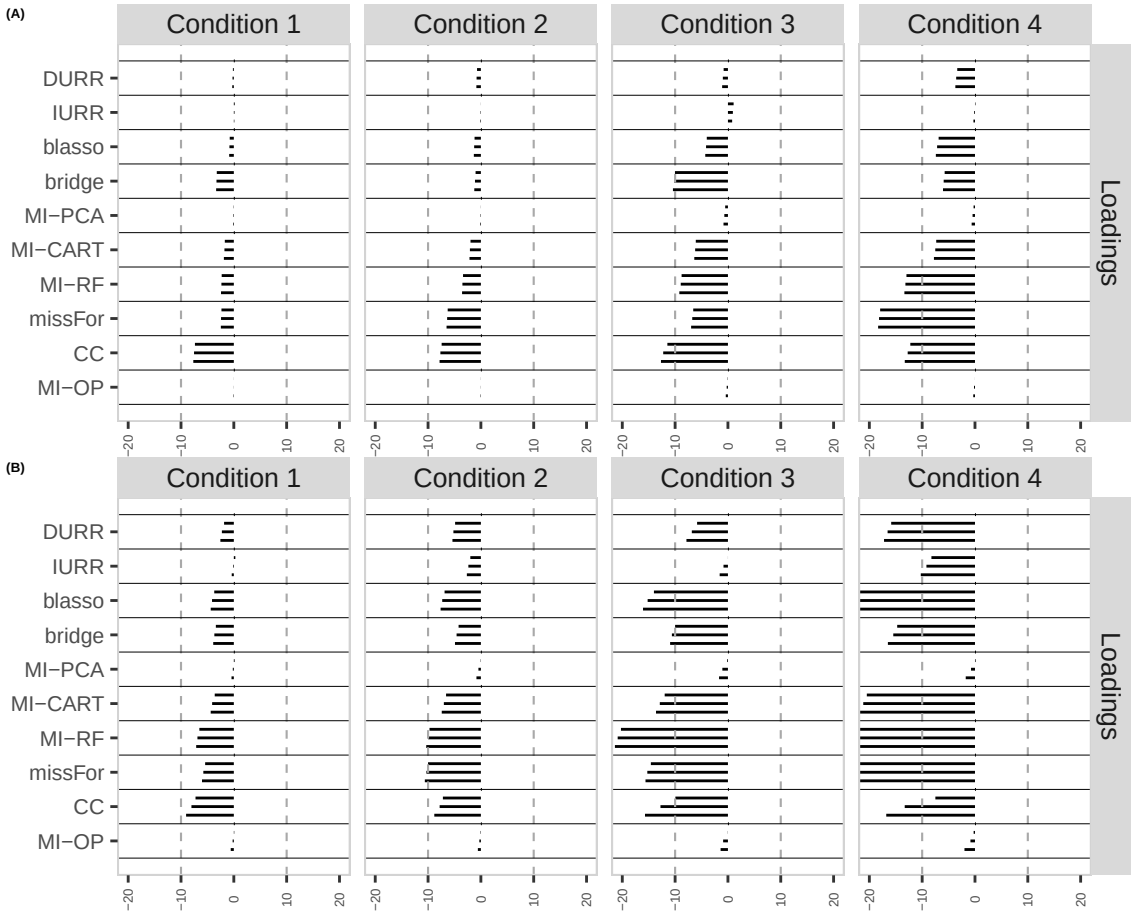

Figure 7. Percent Relative Bias (PRB) for the factor loadings in conditions 1 to 4 (panel A) and conditions 5 to 8 (panel B). Within each panel, for every method, single horizontal lines report the PRB of the factor loading estimation for each item with missing values.

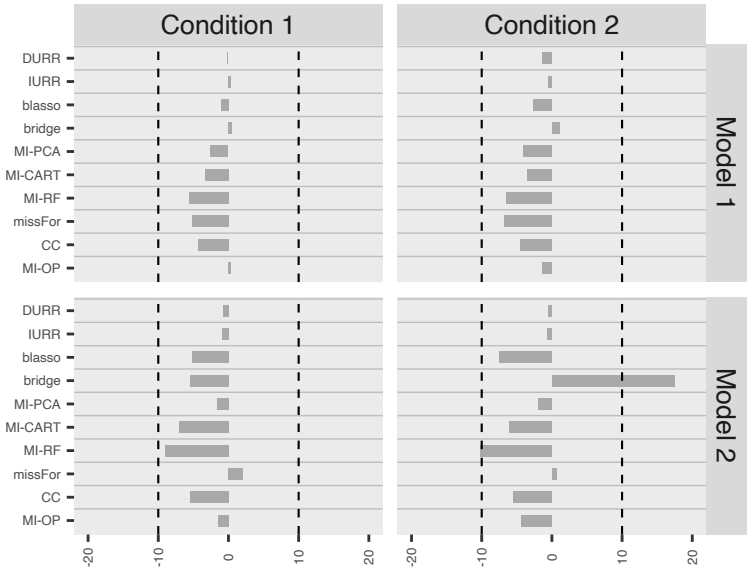

Figure 8. PRB in the estimation of parameters of interest $\beta_1^{(1)}$ and $\beta_1^{(2)}$, in model 1 and 2 respectively

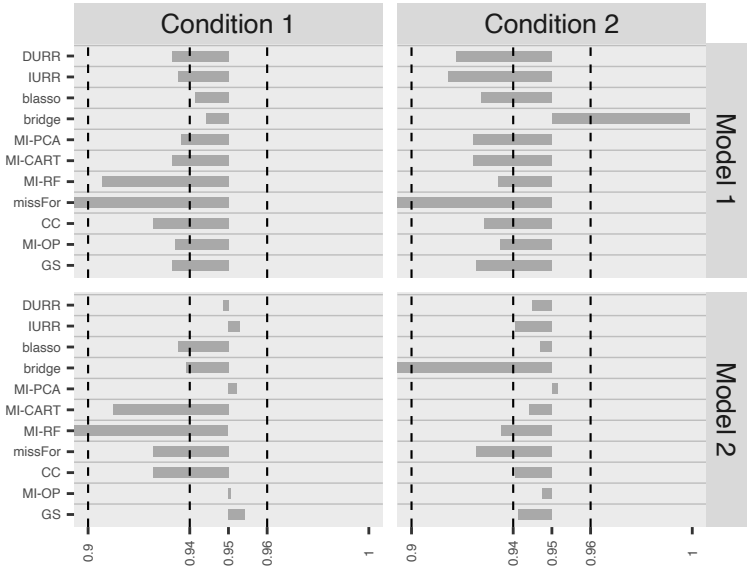

Figure 9. CIC in the estimation of parameters of interest $\beta_1^{(1)}$ and $\beta_1^{(2)}$, in model 1 and 2 respectively

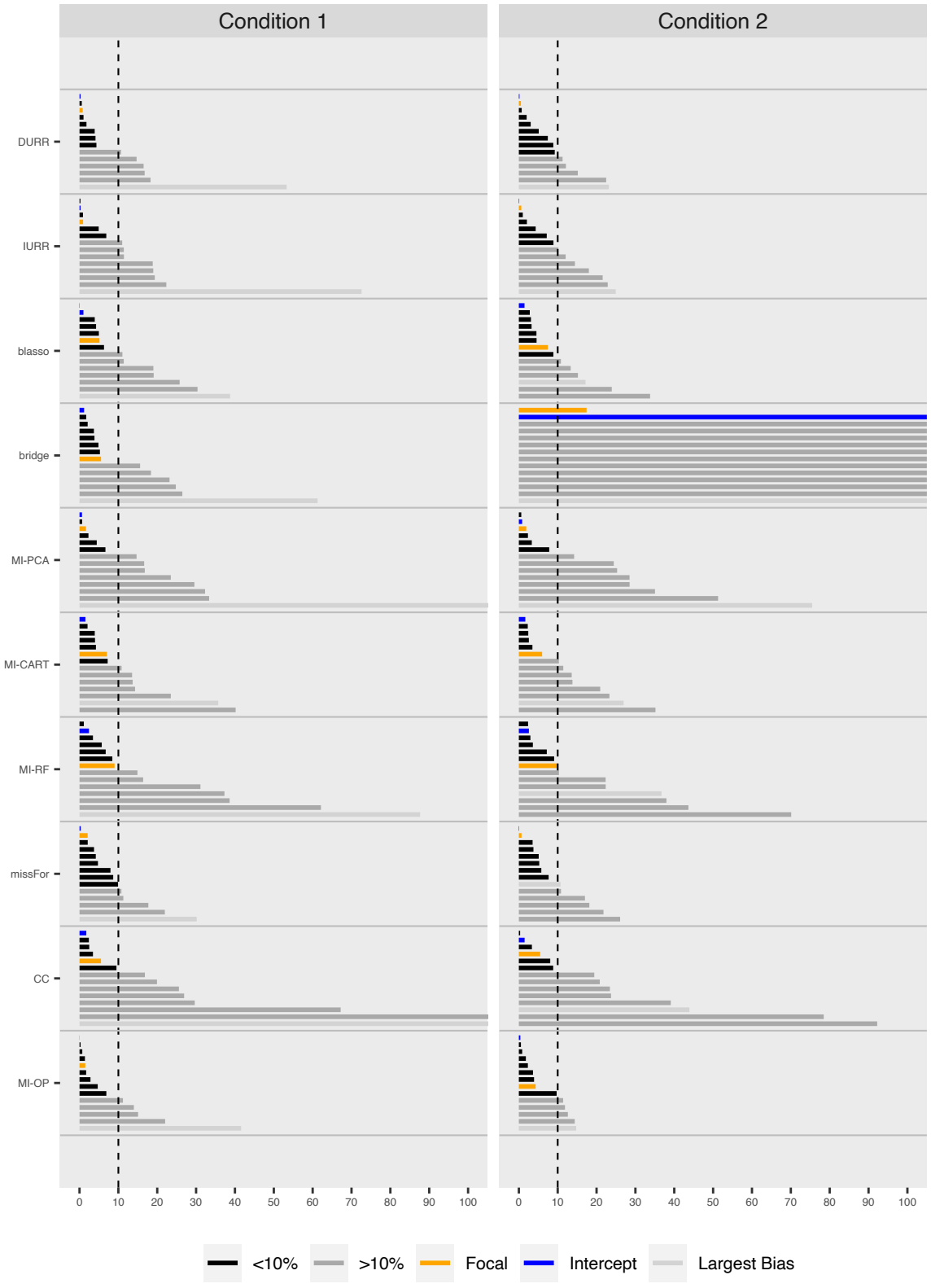

Figure 10. PRBs for all the model parameters in model 2. The order of the bars is based on the absolute value of the PRBs. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

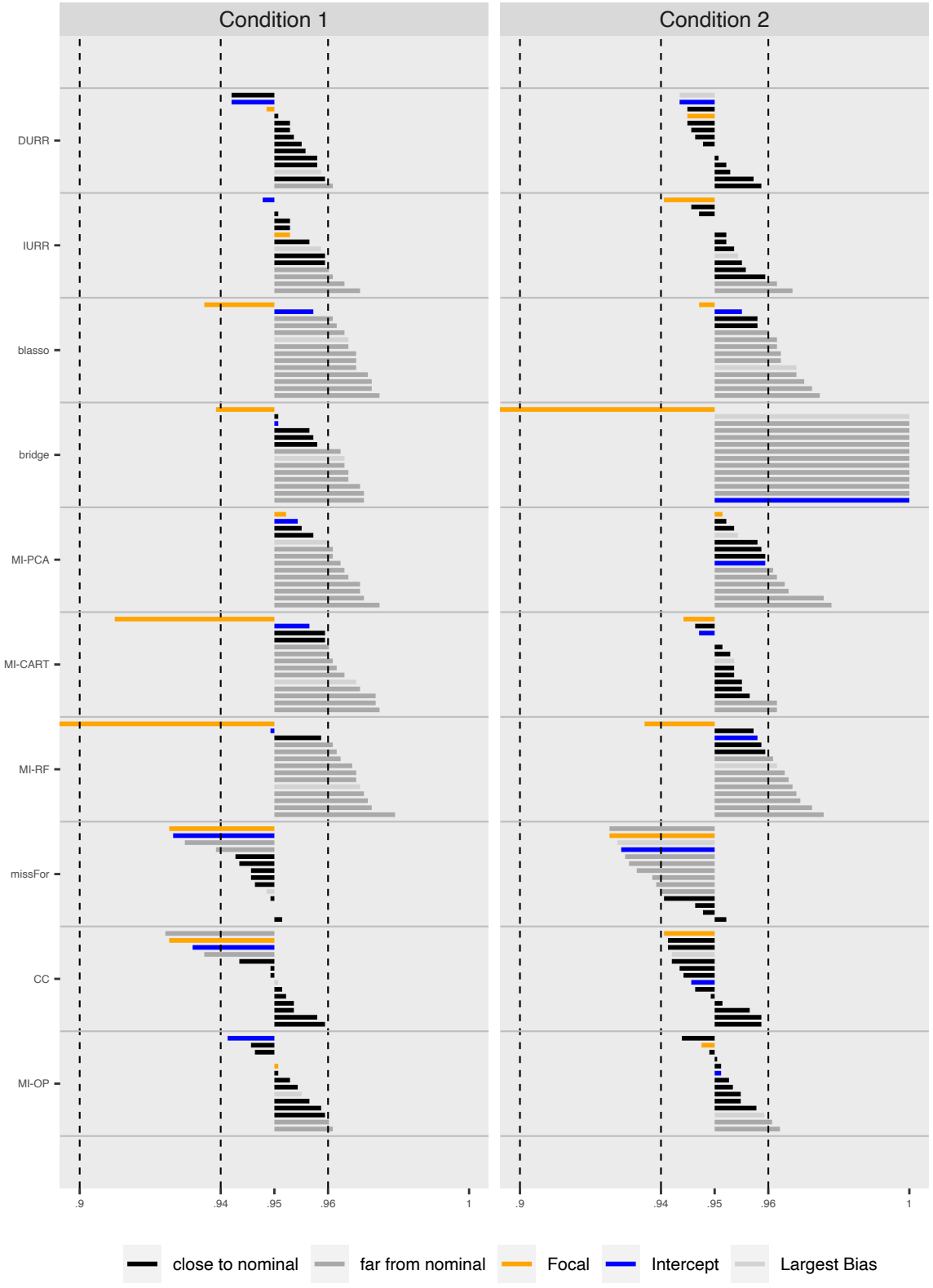

Figure 11. CIC for all model parameter in model 2. Bars are sorted in by ascending value. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted