SUPPLEMENTARY MATERIAL

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

ARTICLE HISTORY
Compiled May 11, 2021

1. Software and other computational details

The code to run the simulation was written in the R statistical programming language (version 4.0.3). All experiments were run using a 2.6 GHz Intel Xeon(R) Gold 6126 processor, 523.78 GB of Memory. The operating system was Windows Server 2012 R2. Computations were run in parallel across 30 cores. Parallel computing was implemented using the R package *parallel* and to ensure replicability of the findings seeds were set using the method by L'ecuyer, Simard, Chen, and Kelton (2002) implemented in the R package *rlecuyer*.

2. Convergence check details

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. The imputation algorithms were considered to have converged 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.

3. Ridge penalty cross-validation details

The ridge penalty used in the bridge algorithm was fixed across iterations . 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.

4. Additional Figures

4.1. Experiment 1: Simulated Data from Multivariate Normal Distribution

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

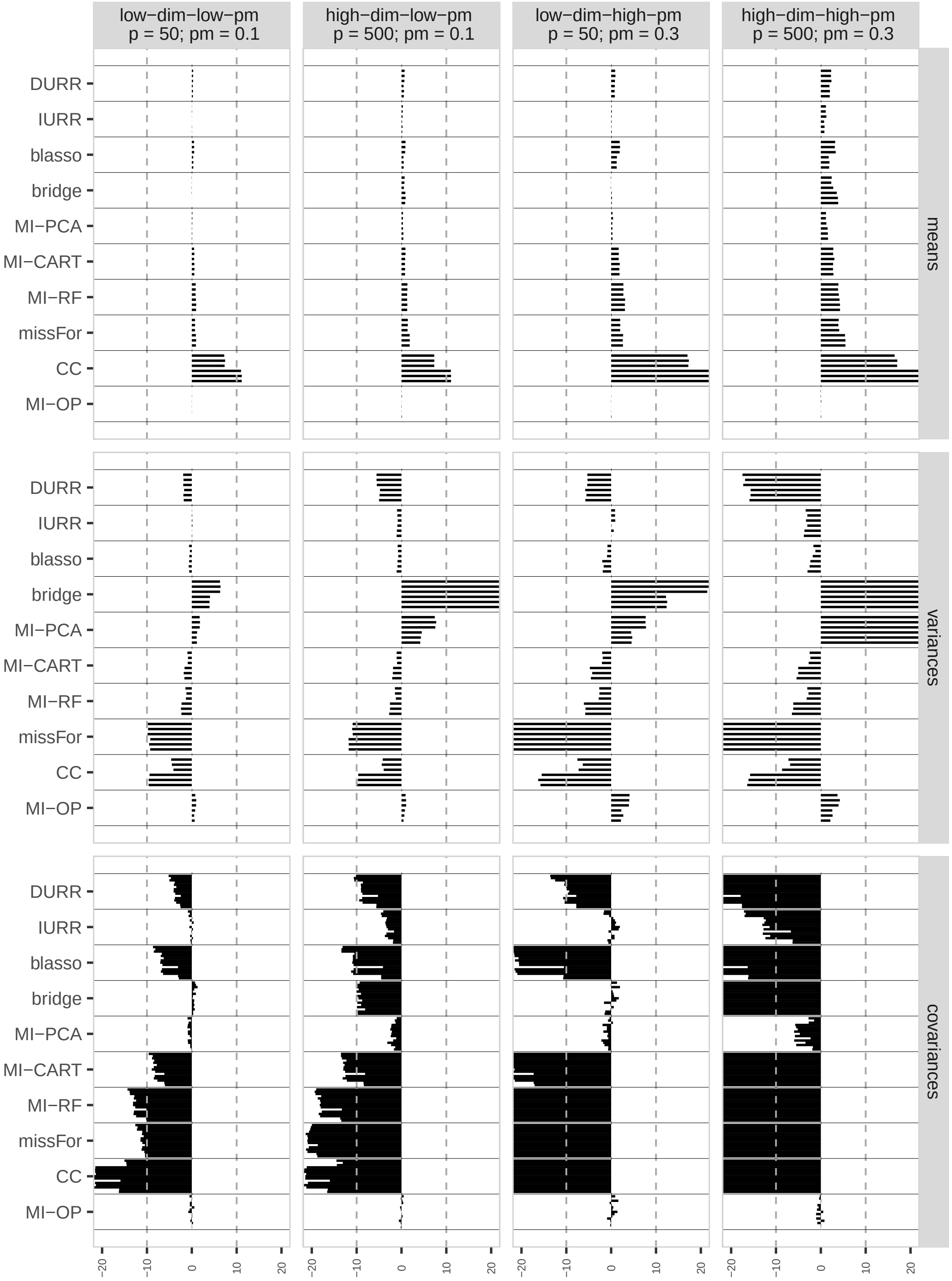

Figure 1. Percent Relative Bias (PRB) for item means, variances, and covariances. For every method, single horizontal lines, representing the PRB of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method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

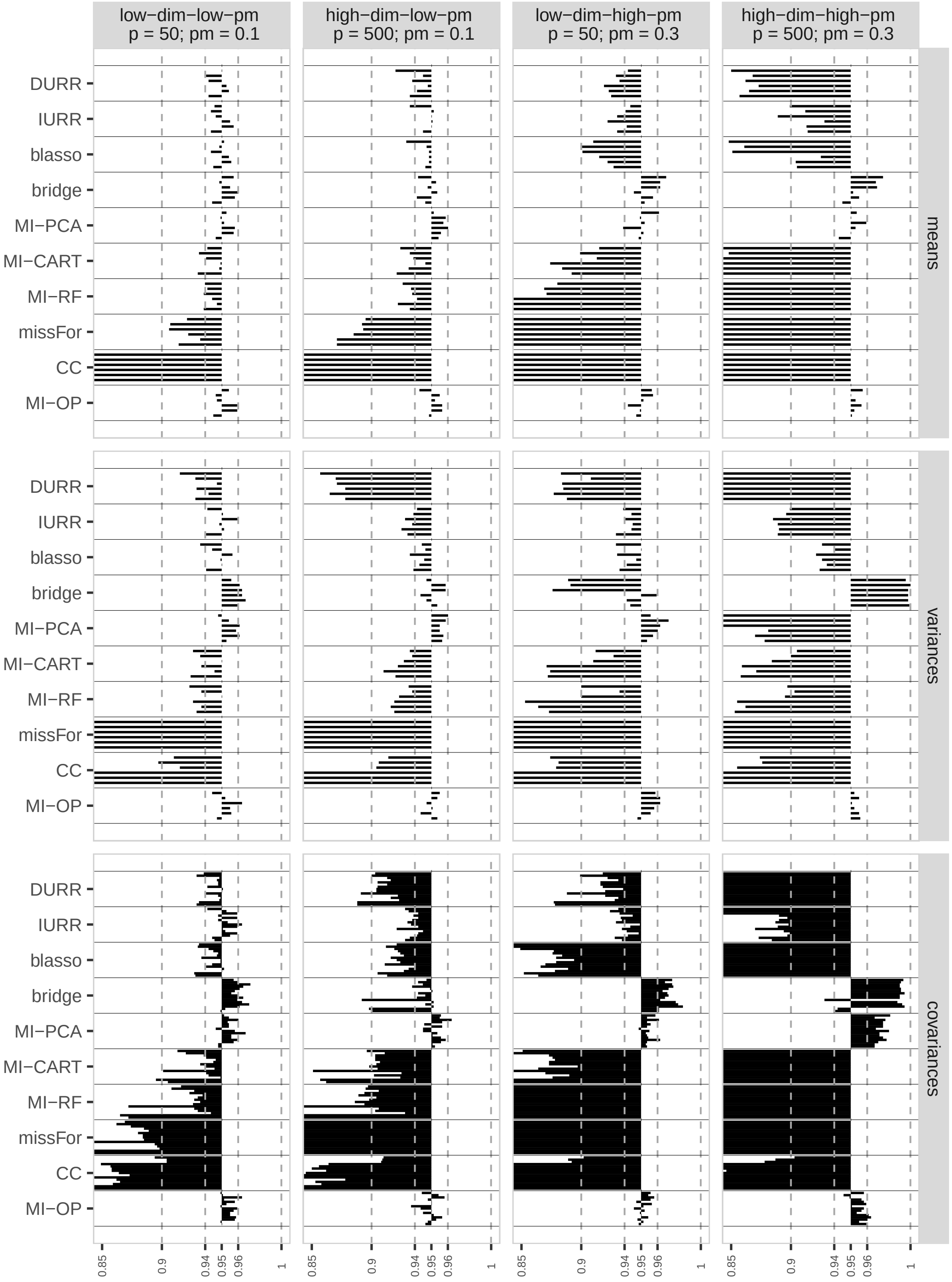

Figure 2. Confidence Interval Coverage (CIC) for item means, variances, and covariances. For every method, single horizontal lines, representing the CIC of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each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

4.2. Experiment 2: Simulated Data with Latent Structure

Figures 3 and 4 report the PRB and CIC of the estimated means, variances, and covariances of the 10 items with missing values in the first four conditions of Experiment 2, the ones characterized by high factor loadings (strong latent structure). Figures 5 and 6 report the same results for the conditions with low factor loadings. Figure **??** reports the PRB for all factor loadings in all conditions.

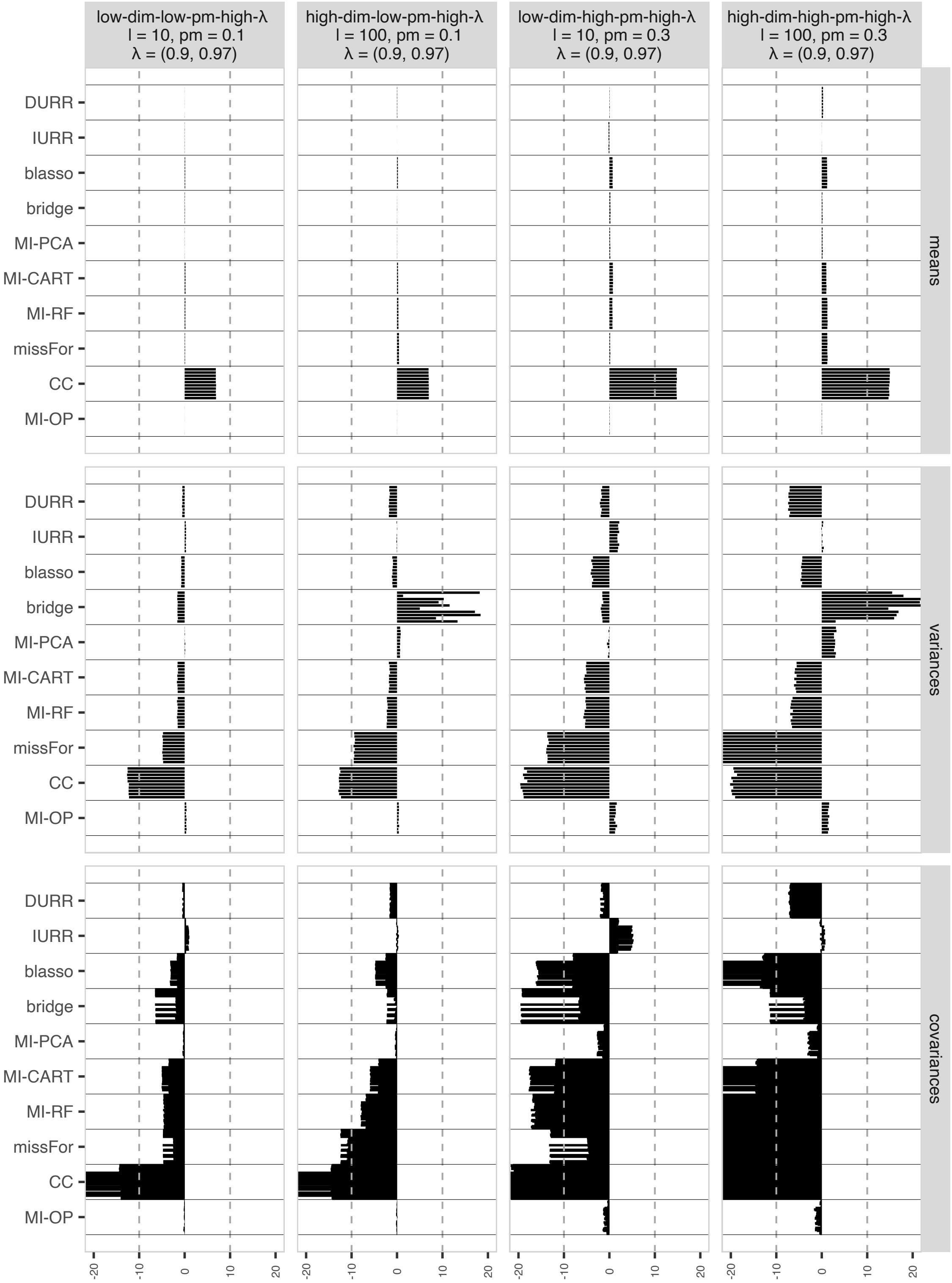

Figure 3. PRBs for the means, variances and covariances (PRB) for condition 1 to 4. For every method, single horizontal lines, representing the PRB of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method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

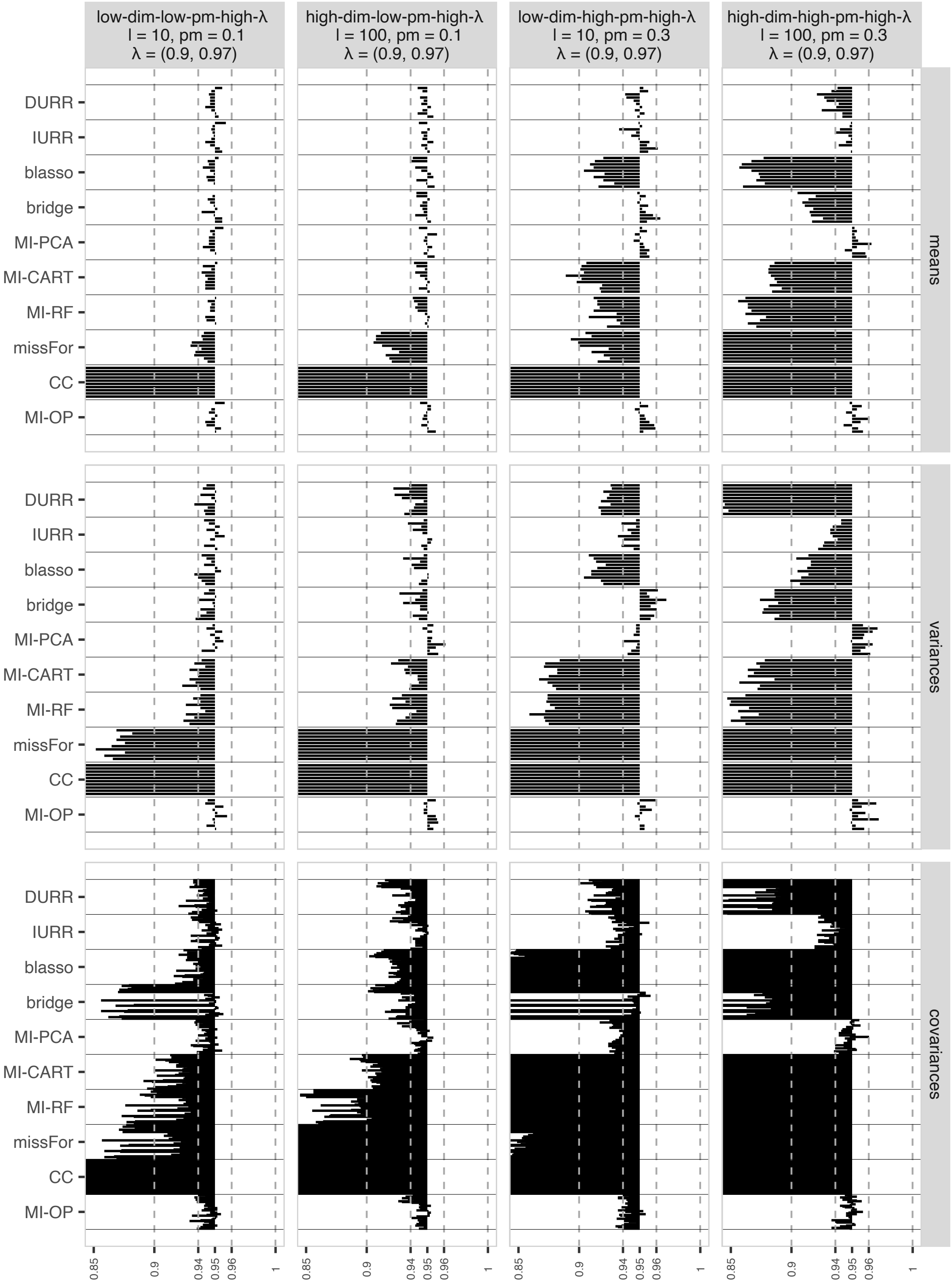

Figure 4. CIC for the means, variances, and covariances for condition 1 to 4. For every method, single horizontal lines, representing the CIC of a parameter estimate on a single variable (or pair of variables), combine to form larger horizontal bars giving an aggregate account of how each method performed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

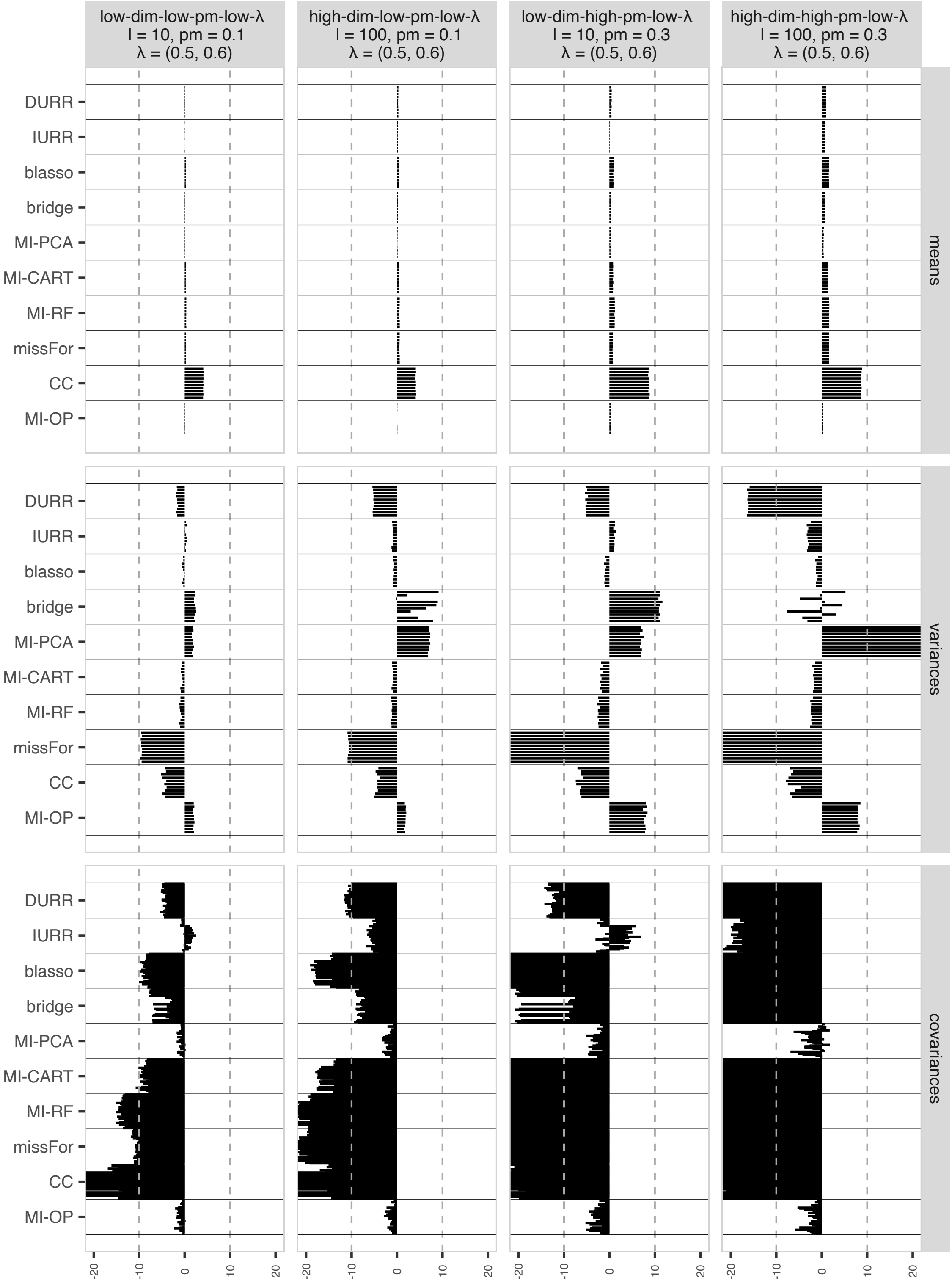

Figure 5. Bias estimation for the means (SB), variances and covariances (PRB) for condition 5 to 8.

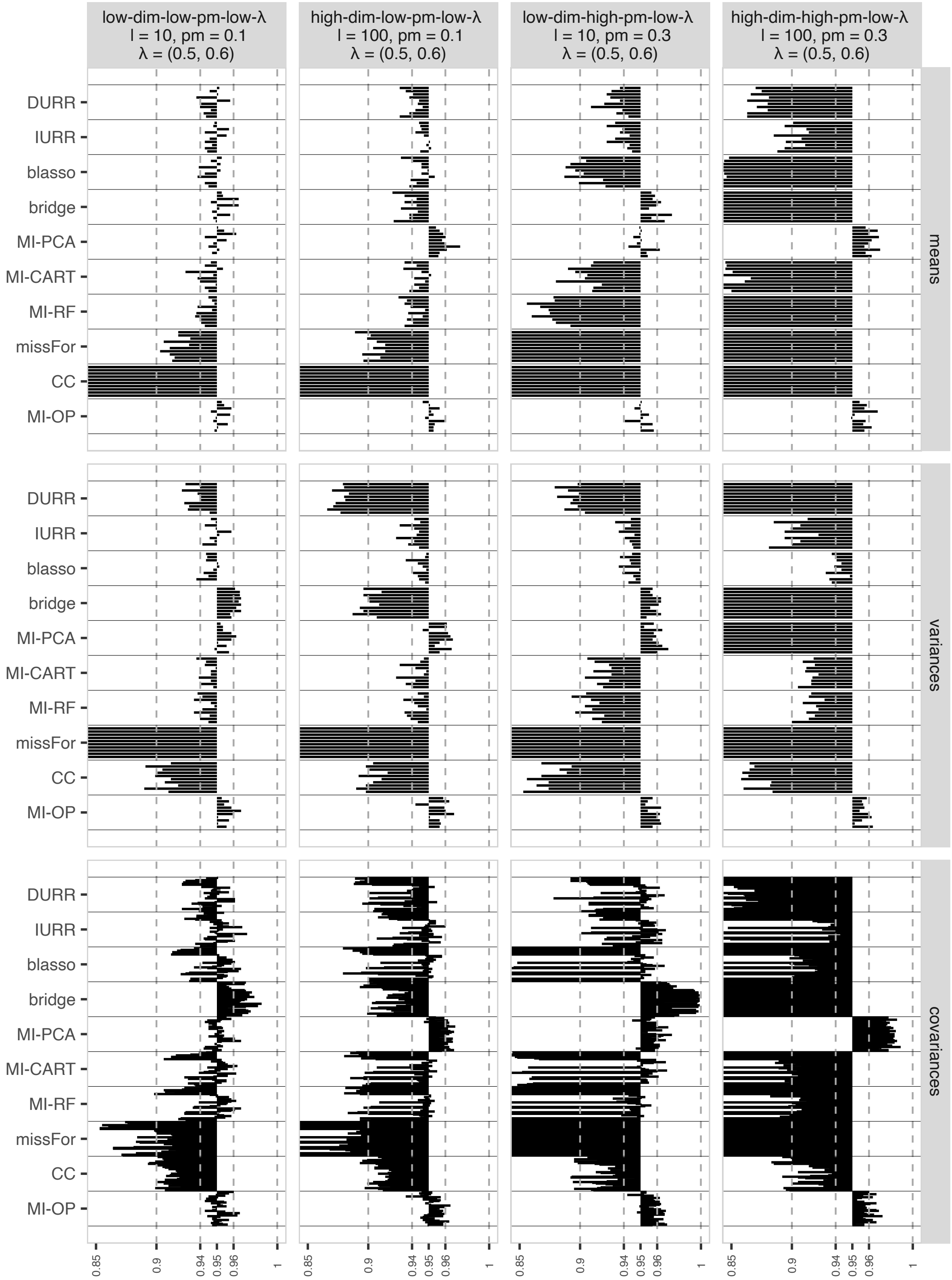

Figure 6. Confidence Interval Coverage (CIC) for the means, variances, and covariances for condition 5 to 8.

4.2.1. Factor Loadings

Figures 7 and 8 report PRBs and CICc for all factor loadings. IURR, MI-PCA, and MI-OP, outperformed all other methods by producing acceptable biases in all conditions. However, 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. DURR, blasso, bridge, and MI-CART produced acceptable to borderline-acceptable biases in all high-λ conditions but tended to produce unacceptable biases in the high-pm-low-λ conditions. MI-RF, missForest, and CC produced acceptable to borderline-acceptable biases with low pm but tended to produce unacceptable biases when pm was high.

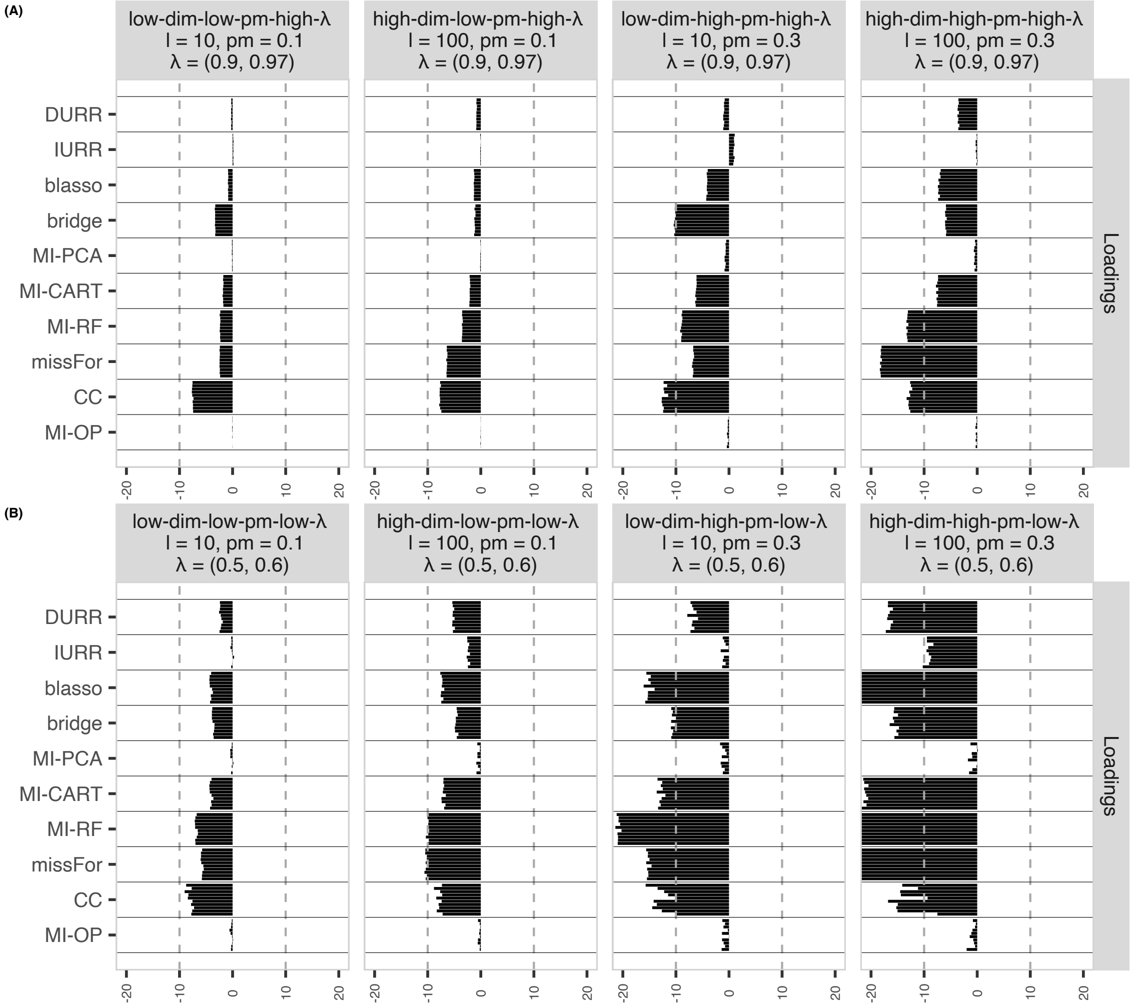

Figure 7. Percent Relative Bias (PRB) for the factor loadings. Within each panel, for every method, single horizontal lines report the PRB of the factor loading estimation for each item with missing values.

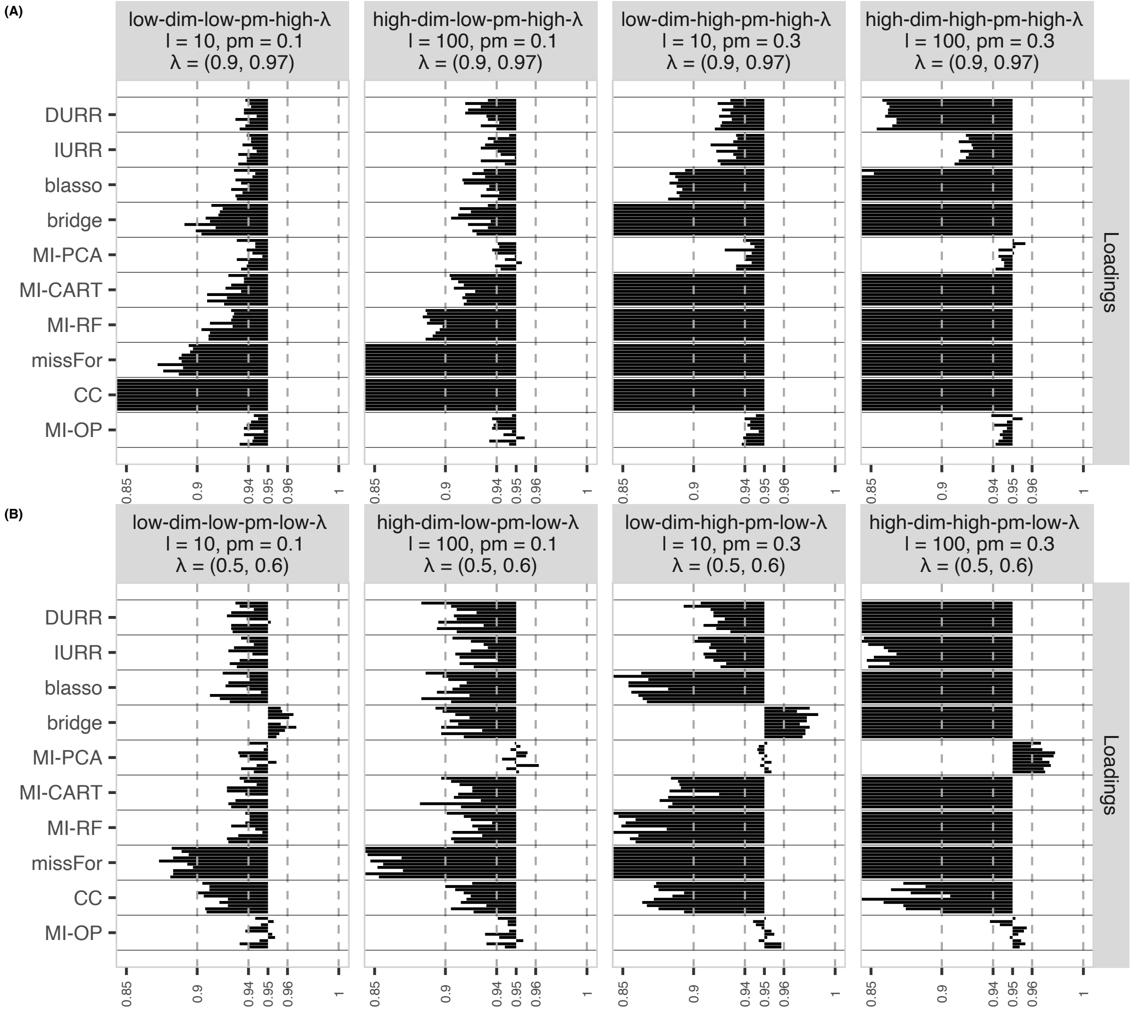

Figure 8. Confidence Interval Coverage (CIC) for the factor loading estimates.

4.3. Experiment 3: Resampling Study

Figure 9 reports the absolute values of the PRBs for the intercept and all the partial regression coefficients in Model 1, under the different imputation methods, for both the low- and high-dimensional conditions. Figure 10 reports CIC results in the same way.

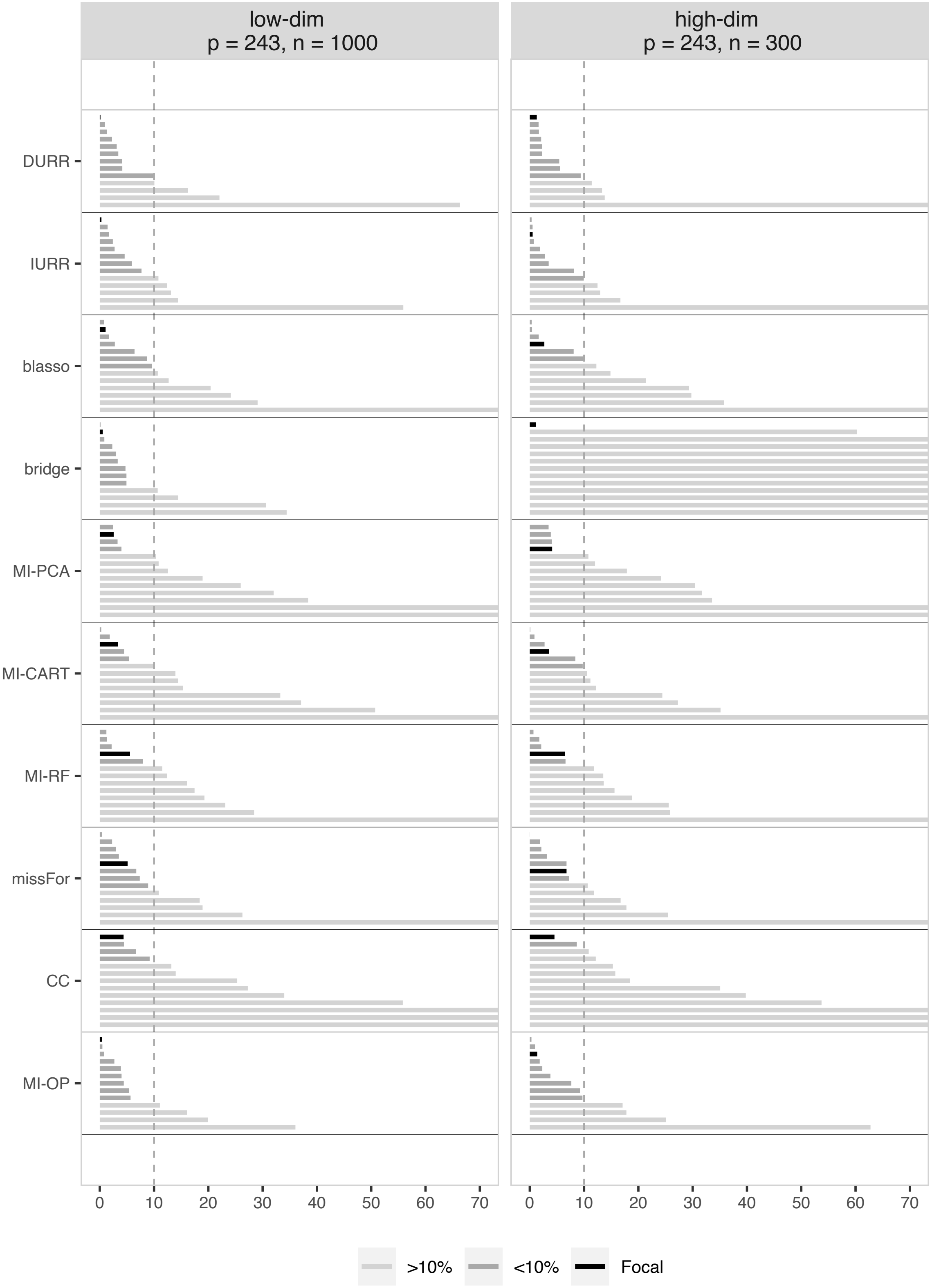

Figure 9. PRBs for all the model parameters in model 1. The order of the bars is based on the absolute value of the PRBs. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

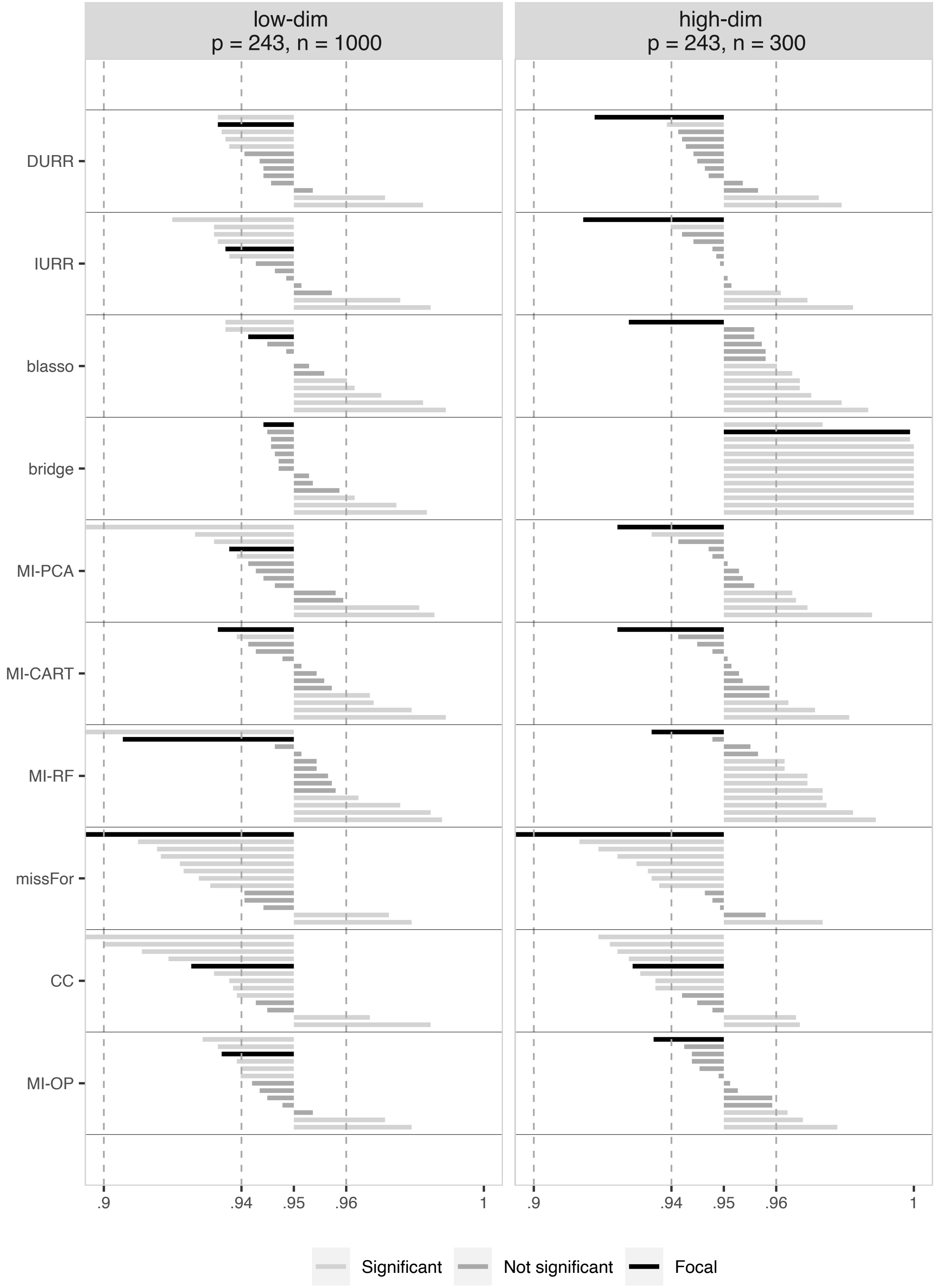

Figure 10. CIC for all model parameter in model 1. Bars are sorted in by ascending value. The values for the intercept, the focal regression coefficient, and the regression coefficient with which most methods struggle (Largest Bias) are highlighted

References

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

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.