Imputation for High Dimensional Data
A comprehensive review

Edoardo Costantini & Kyle M. Lang

July 2020

1 Introduction

Today's social and behavioral scientists are blessed with a wealth of large, high-quality and publicly available social scientific datasets such as the Longitudinal Internet Studies for the Social Sciences (LISS) Panel and the European Values Study (EVS), with initiatives being undertaken to link and extend these datasets into a full system of linked open data (LOD). Making use of the full potential of these data sets requires dealing with the crucial problem of missing data.

The tools researchers working with these data sets need to correct for the bias introduced by nonresponses require special attention. The large number of items recorded, coupled with the longitudinal nature of surveys and the necessity of preserving complex interactions and nonlinear relations, easily produces high-dimensional ($p > n$) imputation problems that impair a straightforward application of imputation algorithms such as MICE ([5]).

Furthermore, when employing Multiple Imputation to deal with missing values, data handlers tend to prefer including more predictors in the imputation models as to reduce chances of uncongenial imputation and analysis models [6]. High-dimensional data imputation settings represent both an obstacle and an opportunity in this sense: an obstacle, as in the presence of high-dimensional data it is simply not possible to include all variables in standard parametric imputation models; an opportunity, because the large amount of features available has the potential to reduces the chances of leaving out of the imputation models important predictors of missignenss.

Many solutions have been proposed to deal with missing values in high dimensional contexts, but most of them have focused on single imputations in an effort to improve the accuracy of individual imputations (see for example [7, 8, 9]). The main task of social scientists is to make inference about a general population based on a sample of observed data, and single imputation is simply an inadequate missing data handling technique for such purpose: it does not guarantee to find estimates that are unbiased and confidence valid ([1]).

Recent years have seen a plethora of studies proposing new Multiple Imputation methods for high dimensional missing data, which can be grouped as follows:

- MI through frequentist use of regularized regression - Using regularized regression to deal with the high-dimensionality of imputation models has been proposed and tested by [10, 11]. Their methods are referred to here as Direct and Indirect Use of Regularized Regression (DURR, and IURR, as abbreviated in their papers)

- MI through Bayesian regularized regression - In the R package 'mice', the implementation of Bayesian MI under the normal linear model allows to use a ridge penalty to estimate the imputation model regression coefficients in the presence of high collinearity and more features than observations (see [5] algorithm 3.1, p. 68). In the present work, we refer to such approach as 'bridge'. [10] have also proposed using [12] Bayesian Lasso regression to implement an alternative fully Bayesian high-dimensional imputation approach (blasso).

- non-parametric MI through regression trees - To this category belong the methods of [13, 14] who proposed, respectively, an integration of regression trees and random forest within the Multiple Imputation by Chained Equations algorithm. In this paper they are referred to as MI-CART and MI-RANF.

- MI through dimensionality reduction - [15] proposed using PCA to extract a few components from a set of auxiliary variables to be used as predictors in regular runs of a MICE algorithm bringing the imputation problem back to a low dimensional one. Such method is referred to here as MI-PCA.

Despite being so prolific, the field is in need of comprehensive studies that compare the performances of such state-of-the-art methods on fair ground for specific scientific endeavours.

With this study we set out to present a thorough review and comparison of MI approaches for high-dimensional datasets. The goal was assessing how they meet the requirements of statistical validity of the analysis performed on the treated data.

2 Methods

The performance of the selected methods was compared through a Monte Carlo simulation study. Performances were assessed in terms of how well the incomplete-data analysis remained statistically valid after missing data was treated. Assuming the complete-data analysis is statistically valid, MI should allow for unbiased and confidence valid analysis of the incomplete data.

2.1 Experimental Conditions

Two design parameters were varied in the simulation: the proportion of (per variable) missing cases (pm), with levels $\{.1, .3\}$; and the number of features of the dataset (p), with levels $\{50, 500\}$. In all conditions, 200 observations were

generated ($n = 200$) and the entire data set was considered when conducting imputations, so that the higher dimensionality of the data resulted in a higher number of potential auxiliary variables.

Table 1: Experiment 1 conditions ($n = 200$)

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

500 data sets were generated for each condition. After imputation, the MLE estimates and standard errors of the means, variances and covariances, of the variables originally with missing values, were estimated and pooled across multiply imputed datasets according to Rubin's rules [citation].

2.2 Data Generation

A dataset X, with dimensionality $n \times p$ (with n = number of observations and p = number of features), was generated according to the standard normal multivariate model:

$$X = MVN(\mu_0 = \mathbf{0}, \mathbf{\Sigma_0}) \tag{1}$$

where μ_0 is a $p \times 1$ vector of 0s, and Σ_0 is a $p \times p$ correlation matrix. Variables were divided in three correlation blocks: high, mid and low correlation. Five variables belonged to block 1 and were correlated among themselves with a correlation coefficient $\rho_1 = .6$. Another five variables belonged to block 2 and were correlated among themselves, and with variables in block 1, with $\rho_2 = .3$. All remaining variables belonged to block 3 and were correlated among themselves, and with variables in block 1 and 2, with $\rho_3 = .01$.

	$Block1$	$Block2$	$Block3$
$Block1$	ρ_1	ρ_2	ρ_3
$Block2$	ρ_2	ρ_2	ρ_3
$Block3$	ρ_3	ρ_3	ρ_3

After sampling the values of X from equation 1, all columns were rescaled to match means, variances, and covariances of continuously treated items in EVS waves. This was done to facilitate the interpretation of the findings in terms of real data applications.

2.3 Missing Data Imposition

Missing values were imposed on six variables, three in block 1 and three in block 2, using the cumulative logistic distribution to define the probability of missingness based on a linear combination of four scaled columns of X.

$$P(y_t = MISS|X) = G(\tilde{X}\theta) \quad (2)$$

where y_t is a variable target of missing values imposition ($t = 1, ..., T$, with $T = 6$), G is the standard cumulative logistic distribution (with location and scale parameters equal to 0 and 1 respectively), $\tilde{X}$ is a $n \times 4$ standardized subset of X, including only the determinants of missingness, and θ is a vector of regression coefficients.

$\tilde{X}$ was composed of 2 variables from block 1 and 2 from block 2. Of the two variables from both blocks, one was selected as target of missing values itself, and the other was fully observed. All variables have same weight in the linear combination $\tilde{X}\theta$.

Overall, the missing data imposition procedure allowed to: (1) specify a general missing data pattern; (2) work with a MAR missingness set up; (3) impose a desired proportion of missing values on each target variable; (4) induce substantive bias for parameters estimates of complete case analysis (between 20 and 30 percent of the size of the reference "true" value of the parameter).

2.4 Imputation Methods and Analyses Model

For each simulated data set, imputation was performed according to all the methods referenced to in the introduction.

As traditional parametric MI could not be applied to cases with $p > n$, we ran, for reference, a standard mice imputation routine that used all variables in block 1 and 2 (10 in total), which included all predictors of missingness and no auxiliary junk variables. Results from this "oracle" run are referred to as MI Optimal run (MI-OP). We have also considered two single dataset missing data handling approaches: MissForest ([8]), as implemented in the R package 'misForest' (here referred to as 'missFor'), and Complete Case analysis (CC), performed using only complete rows of the data.

Convergence of the imputations was checked through visual examination of trace plots showing the mean imputed values for each variable at each iteration. In the most complex condition, ($p = 500, pm = .3$), all methods converged after after approximately 20 iterations. In the simulation study we run a single chain of the MI algorithms for 50 iterations, considering the first 20 as burn-in, and selecting 10 imputed datasets through thinning.

Maximum Likelihood Estimates of the means, variances, and covariances of the six variables with missing values were obtained by fitting a saturated model to the treated data, and pooling the multiple estimates when necessary. Finally, "Gold Standard" (GS) MLEs of the same parameters were obtained from the fully observed data sets (before missing data imposition).

2.5 Evaluation Criteria

Estimation bias introduced by the missing data treatment was quantified as Percent Relative Bias (PBR):

$$PBR = \frac{\bar{Q}^k - R^k}{R^k} * 100 \tag{3}$$

where $\bar{Q}^k$ is the mean estimate of parameter k across the Monte Carlo simulations, and R^k is the reference value corresponding to that parameter. The reference ("true") values of the parameters of interest were obtained by averaging the 500 MLEs obtained on the fully observed datasets.

Furthermore, the euclidean distance d between vectors of raw parameter estimates of the same type of statistic ($\boldsymbol{Q}^K$) and a reference $\boldsymbol{R}^K$ vector was considered to provide a more aggregate quantification of bias:

$$d(\boldsymbol{R}^K, \boldsymbol{Q}^K) = \sqrt{(R_1^K - Q_1^K)^2 + (R_2^K - Q_2^K)^2 + ... + (R_T^K - Q_T^K)^2} \tag{4}$$

where $\boldsymbol{Q}^K$ and $\boldsymbol{R}^K$ are vectors of parameters estimates of statistic type K (i.e., means, variances, covariances). In particular, $\boldsymbol{R}^K$ is the vector or reference values, and $\boldsymbol{Q}^K$ is a vector of Monte Carlo parameters estimates after missing data treatment, and T is the number of variables with missing values.

Finally, to assess the integrity of hypothesis tests conducted under the various imputation approches, the 95% confidence interval coverage rates were computed as:

$$CI_{cov} = \frac{\sum_{s=1}^{S} I(Q \in \hat{CI}_s)}{S^k} * 100 \tag{5}$$

3 Results

3.1 Bias

Simulation results show that all multiple imputation methods perform well, in terms of relative bias in percent, in condition 1, the low-dimensional benchmark setting: the size of the bias is negligible for all methods ($PBR < 10\%$). MI-CART and MI-RANF are the only exceptions, exhibiting bias in percent around and above the 10% threshold.

As dimensionality increases (condition 2), deterioration of performances is found for most approaches. However, IURR and MI-PCA maintain great performances with PBR well below 10%, for all parameters. Blasso and DURR show slightly worst performances in terms of covariance bias (up to 10% in PBR), while maintaining negligible bias for all means and variances. Bridge exhibits a drastic reduction in performance, in particular with substantially biased variances, on par with complete case analysis results.

In condition 4, the larger proportion of missing values does not dramatically change relative performances. IURR and MI-PCA are still exhibiting the lower biases, Blasso and DURR slack behind a little, and tree-based methods are still underperforming. However, MI-PCA manifests substantial bias of the variance estimates, while PBR for covariances and means remains well below 10%. At the

same time, IURR starts to show considerable bias in the covariance estimates. Similarly, Multiple Imputation through Bayesian Blasso maintains extremely low bias for both means and variances, while displaying substantial covariances bias.

Table 2: Euclidean Distances between vectors of reference and estimated parameters grouped by type of statistic ($n = 200$)

Cond	DURR	IURR	bridge	blasso	MI PCA	MI CART	MI RF	MI OP	missFor	CC
All parameters										
p = 50, pm = .1	0.37	0.05	0.62	0.44	0.18	0.63	0.99	0.10	1.50	2.82
p = 50, pm = .3	1.10	0.21	2.78	1.42	0.78	1.73	2.58	0.39	3.96	5.08
p = 500, pm = .1	0.86	0.24	5.02	0.65	0.78	0.88	1.28	0.10	2.10	2.75
p = 500, pm = .3	2.74	1.00	6.22	2.11	3.64	2.42	3.34	0.36	5.40	5.12
Means										
p = 50, pm = .1	0.03	0.01	0.01	0.05	0.02	0.07	0.11	0.01	0.11	1.40
p = 50, pm = .3	0.14	0.06	0.05	0.21	0.10	0.25	0.39	0.05	0.32	3.22
p = 500, pm = .1	0.07	0.03	0.11	0.08	0.03	0.10	0.16	0.01	0.21	1.39
p = 500, pm = .3	0.28	0.14	0.44	0.31	0.20	0.35	0.53	0.06	0.62	3.24
Variances										
p = 50, pm = .1	0.24	0.02	0.62	0.08	0.17	0.20	0.27	0.09	1.20	1.51
p = 50, pm = .3	0.74	0.12	2.78	0.26	0.72	0.54	0.67	0.35	3.37	2.42
p = 500, pm = .1	0.65	0.13	4.96	0.12	0.77	0.24	0.31	0.09	1.41	1.48
p = 500, pm = .3	2.07	0.51	5.69	0.33	3.61	0.61	0.71	0.32	3.98	2.42
Covariances										
p = 50, pm = .1	0.28	0.04	0.02	0.43	0.06	0.59	0.95	0.04	0.88	1.93
p = 50, pm = .3	0.81	0.17	0.12	1.38	0.27	1.63	2.46	0.15	2.04	3.09
p = 500, pm = .1	0.57	0.20	0.77	0.63	0.11	0.84	1.24	0.04	1.54	1.86
p = 500, pm = .3	1.77	0.85	2.48	2.06	0.38	2.31	3.22	0.16	3.60	3.13

Euclidean distances reported in table 2 supplement the results based on the Percent Relative Bias:

- the vector of estimated parameters using high dimensional single imputation methods ('missFor') is many times larger than that of any other imputation method, and, apart from the means, the results match the poor inferential performances of Complete Case analysis.
- IURR and MI-PCA clearly outperform all other methods, providing the vectors of parameter estimates closest to the reference values, for all sets of parameters. However, the deteriorated performances of MI-PCA and IURR in condition 4 are quite evident in the estimation of variances and covariances, respectively.
- blasso performances are overall comparable to MI-PCA and IURR but show greater deterioration of performances due to larger proportion of missing values for variances and covariances.
- the use of regression trees remains quite unsatisfactory, especially when it comes to the bias of covariances.

3.2 Confidence Intervals

As for the confidence interval coverage of the reference values, the performance pattern of the approaches is similar to that of bias, with IURR and MI-PCA maintaining coverage rates closer to nominal levels than all the other methods.

All of the high-dimensional multiple imputation methods considered perform equally well in the low and high dimensional context. In both condition 1 and 2, CIs cover the reference values in approximately 90-95% of the simulated runs, and, for most methods, the coverages do not differ at all between the two conditions.

It is only in condition 3 and 4 that under-coverage of the 95% CI becomes a real concern. Keeping constant the dimensionality of the data (comparing condition 1 and 3, and 2 and 4), a larger pm results in CI coverage from the 90-95% range to 65-80% range for all approaches, except IURR and MI-PCA. However, in condition 4, IURR and MI-PCA start showing sings of under-coverage (CI_{cov} between 85-90 %) and over-coverage (CI_{cov} between 95-98%), respectively.

4 Discussion

Some of the most popular solutions currently implemented in the R package 'mice' to deal with high dimensionality of the data (i.e. bridge and MI-RANF) proved to be quite unsatisfactory in dealing with high-dimensional imputations compared to the other approaches considered. In particular, IURR and MI-PCA result to be clear winners, with Blasso being a worthy challenger.

When deciding which high-dimensional imputation approach to employ, the type of statistics a researcher cares most about should be taken into consideration. Indeed, while IURR, MI-PCA, and Blasso show overall the best performances, they revealed some statistics- specific weaknesses: in the most challenging condition, IURR and Blasso showed substantially biased covariances, while MI-PCA showed poor performance in terms of bias for the variances.

Unreported results from the analysis of linear regressions, involving the variables with imputed values, confirms the importance of considering the type of analysis and statistics, as blasso was even more competitive in the recovery of regression coefficients.

Confidence intervals coverage analysis lead to the conclusion that the reduction in confidence validity for all the high-dimensional imputation methods is a function of the proportion of missing cases, not the dimensionality of the data. While this was perhaps to be expected, it clearly shows the usefulness of high-dimensional imputation methods. Of particular interest was that MI-PCA outperformed all other methods showing consistently good coverage even in the conditions with a high proportion of missing values.

This simulation study is but one part of a larger endeavour that includes other simulation experiments that monitor the performances of high dimensional imputation methods analysed as a latent structure is added to the data generation mechanism, and as interactions come into play within the analysis

model and the missing data imposition.

References

[1] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 473–489, 1996.

[2] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. Hoboken, NJ: Wiley-Interscience, 2 ed., 2002.

[3] J. L. Schafer, *Analysis of incomplete multivariate data*, vol. 72. Boca Raton, FL: Chapman & Hall/CRC, 1997.

[4] C. K. Enders, *Applied missing data analysis*. New York, NY: The Guilford Press, 2010.

[5] S. van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press, 2012.

[6] X.-L. Meng, "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, pp. 538–558, 1994.

[7] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol. 21, no. 2, pp. 187–198, 2005.

[8] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011.

[9] A. D'Ambrosio, M. Aria, and R. Siciliano, "Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm," *Journal of Classification*, vol. 29, no. 2, pp. 227–258, 2012.

[10] Y. Zhao and Q. Long, "Multiple imputation in the presence of high-dimensional data," *Statistical Methods in Medical Research*, vol. 25, no. 5, pp. 2021–2035, 2016.

[11] Y. Deng, C. Chang, M. S. Ido, and Q. Long, "Multiple imputation for general missing data patterns in the presence of high-dimensional data," *Scientific reports*, vol. 6, p. 21689, 2016.

[12] C. Hans, "Bayesian lasso regression," *Biometrika*, vol. 96, no. 4, pp. 835–845, 2009.

[13] L. F. Burgette and J. P. Reiter, "Multiple imputation for missing data via sequential regression trees," *American Journal of Epidemiology*, vol. 172, no. 9, pp. 1070–1076, 2010.

[14] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study," *American journal of epidemiology*, vol. 179, no. 6, pp. 764–774, 2014.

[15] W. J. Howard, M. Rhemtulla, and T. D. Little, "Using principal components as auxiliary variables in missing data estimation," *Multivariate Behavioral Research*, vol. 50, no. 3, pp. 285–299, 2015.