Imputation for High Dimensional Data: A comparison of state-of-the-art methods

Edoardo Costantini[1], Kyle M. Lang[1],
[1]Department of Methodology and Statistics, Tilburg University, Netherlands

Abstract

One of the most popular principled missing data treatments, multiple imputation (MI), is challenged by computational limitations when applied to large social survey datasets. The large number of items recorded—coupled with the longitudinal nature of these surveys and the necessity of preserving complex interactions and non-linear relations—easily produces high-dimensional (p>n) imputation problems.

We performed a thorough review of the high-dimensional prediction literature to find the most promising, extant MI methods for this type for data. We found that principal component regression, classification and regression trees, ensemble learning, and regularized regression, have all been used for high-dimensional MI—both in their frequentist and Bayesian versions.

We compare these methods' performances through a Monte Carlo simulation study, and find that currently recommended approaches, such as mice-random forest and mice with ridge penalty, underperform compared to the other state-of-the-art high-dimensional imputation techniques. Ultimately, we provide practical guidelines for social scientists working with incomplete high-dimensional data.

Keywords: Multiple Imputation; high-dimensionality.

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 (van Buuren, 2012).

Furthermore, when employing Multiple Imputation (MI) 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 (Meng, 1994). 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 reduce the chances of leaving out of the imputation models important predictors of missingness.

Recent years have seen a plethora of studies proposing new Multiple Imputation methods for high dimensional data. The most promising approaches we have found 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 Zhao and Long (2016) and Deng et al. (2016). 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 MI under the Bayesian 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 (van Buuren, 2012, p. 68, algorithm 3.1). In the present work, we refer to such approach as 'bridge'. Zhao and Long (2016) have also proposed using Hans (2009)'s Bayesian Lasso regression to implement an alternative fully Bayesian high-dimensional imputation approach (blasso), and found promising results in a univariate missingness imputation set up.

- non-parametric MI through regression trees - To this category belong the methods of Burgette and Reiter (2010) and Shah et al. (2014) who proposed, respectively, an integration of regression trees and random forest within the Multiple Imputation by Chained Equations (MICE) algorithm. In this paper they are referred to as MI-CART and MI-RANF.
- MI through dimensionality reduction - Howard et al. (2015) proposed using PCA to extract principal 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. With this study, we set out to present a thorough review and comparison of MI approaches for high-dimensional datasets. Assuming that the complete-data analysis is statistically valid, MI should allow to obtain unbiased and confidence valid estimates of the analysis model parameters from incomplete-data (Rubin, 1996). The degree to which each method exhibits these properties was assessed through a Monte Carlo simulation study.

2. Method

2.1. Experimental Conditions

Two design factors 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}. Table 1 summarizes the resulting four conditions. In each of them, 200 observations were generated ($n = 200$) and the entire data set was considered when conducting imputations, so that the higher number of potential auxiliary variables in conditions 2 and 4, resulted in high dimensional imputation problems. Results were averaged over 500 data replications in each condition ($S = 500$).

Table 1: Conditions summary

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

2.2. Data Generation

A dataset X, with dimensionality $n \times p$, was generated according to the normal multivariate model: $\mathrm{X} \sim \mathrm{MVN}(\mu_0,\ \Omega_0)$, where μ_0 is a $p \times 1$ vector of 0s, and Ω_0 is a $p \times p$ correlation matrix.

Variables were divided in three correlation blocks: high, mid and low correlation. Variables 1 to 5 belonged to block 1, and were correlated among themselves with coefficient $\rho_1 = .6$; variables 6 to 10 belonged to block 2, and were correlated among themselves, and with variables in block 1, with $\rho_2 = .3$; variables 11 to p belonged to block 3, and were correlated among themselves, and with variables in block 1 and 2, with $\rho_3 = .01$.

In order to facilitate the interpretation of the findings in terms of real data applications, after sampling the values of X, all columns were rescaled to approximately match means, variances, and covariances of agreement items, on a scale from 1 to 10, in the 2017 wave of EVS.

2.3. Missing Data Imposition

Missing values were imposed on six variables (multivariate missing), 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:

$$P(y_t = \mathrm{NA} \mid \mathrm{X}) = \mathrm{G}(\theta_0 - \dot{X}\theta) \tag{2}$$

where y_t is a variable target of missing values imposition ($t = \{1, \ldots, T\}$ with $T = 6$), G is the standard cumulative logistic distribution (with location and scale parameters equal to 0 and 1 respectively), $\dot{X}$ is a $n \times 4$ standardized subset of X, including only the determinants of missingness, and θ is a vector of regression coefficients. θ_0 is an 'offsetting' intercept term that allowed to specify where, in the distribution of the linear combination $\dot{X}\theta$, the majority of missing values occurred. The specification of equation 2 used here allowed to define condition-specific proportions of missing values (pm) in the lower part of the weighted sum of $\dot{X}$.

$\dot{X}$ was composed of two variables from each of block 1 and 2. Of the two variables from each block, one was selected as target of missing values itself, and the other was fully observed. All predictors had the same weight in the linear combination $\dot{X}\theta$. To avoid MNAR, a target variable was never a predictor in its own response model.

2.4. Imputation Methods and Analysis Model

For each simulated data set, imputation was performed according to all the methods referenced to in the introduction. Ridge penalty was set to the 'mice' R-package default

value of 1e-5 and the priors specification for blasso followed Hans (2010). In the MI-PCA approach, enough components were extracted so that the cumulative proportion of variance explained was .5.

After imputation, Maximum Likelihood Estimates of the means, variances, and covariances of the six variables with missing values were estimated, and pooled across multiply imputed datasets according to Rubin (1987)'s rules, when necessary.

2.5. Evaluation Criteria

The bias of the parameter estimates, introduced by the missing data treatment, was quantified as Percentage Relative Bias (PRB):

$$PRB = \frac{\bar{Q}_k - R_k}{R_k} \times 100 \quad (3)$$

where $\bar{Q}_k$ is the mean estimate of parameter k across the S Monte Carlo replications, 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.

To assess the integrity of hypothesis tests conducted under the various imputation approaches, the 95% Confidence Interval Coverage rates were computed as:

$$CIC = S^{-1} \times \sum_{s=1}^{S} I(R_k \in \widehat{CI}_s) \times 100 \quad (5)$$

where $\widehat{CI}_s$ is an estimated 95% confidence interval for one parameter estimate in the s-th repetition ($S = 500$, the total number of Monte Carlo repetitions), and $I(\cdot)$ is the indicator function that returns 1 when the argument is true, 0 otherwise.

3. Results

Table 2 shows results for selected parameter estimates: the means and variances of variable 1 and 6, belonging to block 1 and 2, respectively, and covariances between variable 1 and 2, 1 and 6, and 6 and 7. These point estimates and CIC summarize the general patterns found for all the parameters of their type. The "OP" analysis is performed on an Optimal imputation model run of mice, that included all the predictors of missingness in the true response model, together with all the other variables in block 1 and 2.

3.1. Bias

In condition 1, the low-dimensional benchmark setting, the size of the bias was negligible for all methods (PRB < 10%). MI-CART and MI-RANF were the only exceptions, exhibiting bias in percent around and above the 10% threshold for the covariance estimates.

As dimensionality increased (condition 2), IURR and MI-PCA maintained great performances with PRB well below 10%, for all parameters. Blasso and DURR showed slightly worst performances in terms of covariances bias (up to 10% in PRB), while maintaining negligible bias for all means and variances. Bridge exhibited a drastic reduction in performance, in particular with substantially biased variances.

In condition 4, the larger proportion of missing values did not dramatically change relative performances. However, MI-PCA manifested substantial bias of the variance estimates, while PRB for covariances and means remained well below 10%. At the same time, IURR showed non-negligible bias in the covariance estimates (PRBs between 10 and 20%). Similarly, blasso maintained extremely low bias for both means and variances, while displaying substantial covariances bias (PRBs between 30 and 40%).

3.2. Confidence Intervals

As for the confidence interval coverage of the reference values, the performance pattern of the approaches was 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 performed equally well in the low and high dimensional context. In both condition 1 and 2, CIs covered the reference values in approximately 90-95% of the simulated runs, and, for most methods, the coverages did not differ at all between the two conditions.

It is only in condition 3 and 4 that coverage of the 95% CI became a real concern. Keeping constant the dimensionality of the data (comparing condition 1 and 3, and 2 and 4), a larger *pm* resulted in many CI coverages to fall below 90%, especially for variances and covariances. IURR and MI-PCA maintained the best performances even though, in condition 4, they started showing signs of under-coverage and over-coverage for variances and covariances. Bridge, MI-CART and MI-RANF showed the worst performances of all multiple imputation approaches with some CIC as low as 30%.

Table 2 - Percentage Relavtive Bias (PB) and Confidence Interval Coverage (CR) for selected parameters

		μ_1		μ_6		σ^2_1		σ^2_6		σ_{12}		σ_{16}		σ_{67}	
		PB	**CR**	**PB**	**CR**	**PB**	**CR**	**PB**	**CR**	**PB**	**CR**	**PB**	**CR**	**PB**	**CR**
	CC	11	19	8	53	9	81	5	89	17	74	23	86	17	90
	p = 50														
pm = .1	**misFor**	1	93	1	93	9	74	10	73	11	79	11	88	13	88
	DURR	0	96	0	96	2	93	2	93	3	91	4	95	6	96
	IURR	0	97	0	97	0	94	0	93	1	93	1	95	1	96
	bridge	0	97	0	98	4	97	6	96	0	95	0	96	1	98
	blasso	0	96	1	96	1	94	0	94	3	92	7	93	9	95
	PCA	0	96	0	97	1	95	2	94	0	94	1	95	0	97
	CART	1	95	1	96	2	92	1	93	6	89	9	94	9	95

	RANF	1	95	1	96	2	92	1	93	10	85	13	93	14	93
	OP	0	96	0	98	1	96	1	94	0	93	1	94	1	96
	p = 500														
	misFor	2	89	1	88	11	69	11	69	19	61	21	83	21	78
	DURR	1	95	1	92	5	89	6	87	6	90	9	93	12	91
	IURR	0	96	0	93	1	94	1	94	2	93	4	96	6	94
	bridge	1	96	1	93	33	92	61	94	10	87	9	95	11	93
	blasso	1	96	1	94	1	95	1	96	5	91	11	94	15	91
	PCA	0	96	0	95	5	97	7	97	1	95	1	97	2	96
	CART	1	95	1	94	2	93	1	93	8	88	13	93	15	87
	RANF	1	95	1	92	2	93	2	94	13	83	18	92	20	90
	OP	0	96	0	94	1	95	1	96	0	95	0	97	1	94
	CC	26	2	17	25	17	75	7	87	29	66	35	84	29	85
	p = 50														
	misFor	3	77	2	80	26	15	28	12	26	35	26	61	30	57
	DURR	1	92	1	93	6	87	6	89	8	89	11	94	14	89
	IURR	0	94	0	94	0	93	0	95	1	93	2	95	3	91
	bridge	0	96	0	98	17	90	27	84	0	96	0	97	3	97
	blasso	1	92	2	91	2	95	1	94	11	85	22	89	27	83
	PCA	0	95	0	96	4	97	7	97	1	96	3	96	2	95
	CART	2	89	2	91	4	88	2	94	18	74	23	87	25	83
	RANF	3	83	3	87	6	85	3	92	29	51	34	83	37	79
pm = .3	**OP**	0	95	0	95	2	96	4	96	1	94	2	95	1	94
	p = 500														
	misFor	6	45	4	56	33	4	31	6	45	2	49	21	48	22
	DURR	2	87	2	86	16	53	17	54	18	66	27	75	30	67
	IURR	1	93	1	89	5	89	3	92	7	86	14	88	16	85
	bridge	4	73	3	80	30	57	80	67	33	35	31	73	31	69
	blasso	2	92	3	83	3	93	1	95	16	76	33	79	40	73
	PCA	2	96	1	95	23	87	36	76	2	98	7	97	1	97
	CART	3	86	3	82	6	86	2	91	23	59	34	74	36	69
	RANF	4	78	4	79	6	85	2	92	37	30	46	65	47	65
	OP	0	96	0	95	1	97	4	96	1	94	3	95	1	97

4. Conclusions

Some of the most popular solutions currently implemented in the R package ‘mice’ to deal with a large number of predictors (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 close 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 (and coverage) for the variances.

This simulation study is but one part of a larger endeavor that includes other simulation experiments that monitor the performances of the high dimensional imputation methods analyzed, 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

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070– 1076.

Deng, Y., Chang, C., Ido, M. S., and Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6:21689.

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

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. Statistics and Computing, 20(2), 221-229.

Howard, W. J., Rhemtulla, M., and Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, 50(3):285–299.

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

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

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

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774.

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

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.

Zhao, Y. and Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, 25(5):2021–2035.