Imputation for High Dimensional Data
A comprehensive review

Edoardo Costantini & Kyle M. Lang

July 2020

1 Introduction

Frame the problem 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 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 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 (Kim et al., 2005; Stekhoven and Bühlmann, 2011; D'Ambrosio et al., 2012). 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 (Rubin, 1996).

Discuss background literature What are the most relevant works in the field

What is the main focus? What is the reason to write this paper? Frame the social sciences focus of the project.

These are the sections coming up. Describe in words the set up of the article (expected 40 manuscript styled pages):

- Introduction (done) Frame problem. Discuss background literature. Main focus: what is the reason to write paper With these sections coming up.
- Algorithms and Imputation methods - DURR - IURR - blasso etc - last one Focus on minimal possible description to give reader sense of what the method is without going to the source paper. (Deng et al max). Do not feel obligated to use eq, prefer words: use it if it adds specificity and clarity. One option can be: having extensive definition in appendix and very streamlined here. In appendix definitely all the "adaptation" that were required.
- Simulation Studies
  - Methods Study 1 + Study 2
  - Data generation
  - Missing data imposition
  - Analysis models
  - Criteria
  - procedure: summary of crossed conditions: describe sequentially what happens each replication
  - Results
- Resampling Study (EVS)
  - Methods
  - Data preparation (giving correct documentation for the data, what it is, why collected cleaning, with general demographics originally + went to systematic cleaning process with general purpose for the cleaning + large western European countries, details are in the appendix if interested)
  - Missing data imposition
  - Analysis models
  - Criteria
  - Procedure (summary with the number of observations kept and so on)

  - Results/Discussion again divide by type. with some implication but not comparison

- Discussion Synthesize findings: here make parallels and comparisons.
- Conclusions One or two paragraphs with take home, limitations, future directions (hint at MY future work)
- Appendices - methods details - EVS quirks

2 Imputation methods and Algorithms

I will provide a minimal description of the imputation algorithms.

2.1 Multiple Imputation Strategies

Consider a dataset $\boldsymbol{Z}$ that has p variables (columns) and n observations (rows). Assume that there are T variables with missing cases in at least one row (i.e. imputation target variables).

Direct Use of Regularized Regression (DURR) For a target variable $\boldsymbol{z}_j$, the DURR algorithm follows these directions:

- Sample with replacement n rows of $\boldsymbol{Z}$ and keep the current values of all columns. This yields $\boldsymbol{z}_j^*$ and $\boldsymbol{Z}_j^{*(m)}$, the bootstrap data set considered at iteration m.
- Use any regularized regression method (such as Lasso regression) to fit a linear model with $\boldsymbol{z}_{j,obs}$, the observed values of $\boldsymbol{z}_j$, as outcome and $\boldsymbol{Z}_{j,obs}^{*(m)}$, the values of $\boldsymbol{Z}_j^{*(m)}$ corresponding to the observed values of $\boldsymbol{z}_j$, as set of predictors. This produces a set of parameter estimates (regression coefficients and error variance) $\hat{\boldsymbol{\theta}}_j^{(m)}$ that can be considered as samples from the parameters' posterior distribution conditioned on the observed part of the data.
- Predict $\boldsymbol{z}_{j,mis}$, the missing values on target variable j, based on $\boldsymbol{Z}_{j,mis}^{*(m)}$, the values of and $\hat{\boldsymbol{\theta}}_j^{(m)}$, to obtain draws from the posterior predictive distribution of the missing data.

At iteration m, these steps are repeated to for each j-th variable in the set of T target variables. After convergences, M different sets of imputations are kept to form M differently imputed data sets. Any substantive model(s) can then be fit to each data and estimates can be pooled appropriately.

Indirect Use of Regularized Regression (IURR) While DURR performs simultaneously model trimming and parameter estimation, another approach is to use regularized regression exclusively for model trimming, and to follow it up by standard multiple imputation procedure. At iteration m, the IURR algorithm performs the following steps for each target variable:

- Fit a multiple linear regression using a regularized regression method with $\boldsymbol{z}_{j,obs}$ as dependent variable and the corresponding current values of all the other variables ($\boldsymbol{Z}^{(m)}_{j,obs}$) as predictors. 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 $\boldsymbol{Z}^{(m)}_{j,obs}$ and draw a new value of these coefficients by sampling from a multivariate normal distribution centered around these MLEs.

$$(\hat{\boldsymbol{\theta}}^m_j, \hat{\sigma}^m_j) \sim N(\hat{\boldsymbol{\theta}}^m_{MLE}, \hat{\boldsymbol{\Sigma}}^m_{MLE}) \tag{1}$$

- Predict $\boldsymbol{z}_{j,mis}$, the missing values on target variable j, by sampling from the posterior predictive distribution based on $\boldsymbol{Z}^{(m)}_{j,mis}$, and the parameters posterior draws $(\hat{\boldsymbol{\theta}}^m_j, \hat{\sigma}^m_j)$ to obtain draws from the posterior predictive distribution of the missing data.

After convergence is reached, M differently imputed data sets are kept and used for the substantive analysis.

MICE with Bayesian lasso (blasso) A Bayesian hierarchical BLasso linear model is a regular Bayesian multiple regression with a prior specification for the regression coefficients that induces some form of shrinkage toward 0 of the sampled parameters values.

The Bayesian Lasso imputation algorithm is a standard Multiple Imputation MCMC sampler that uses the shrinkage priors defined by **?** to compute the posterior distributions of the regression coefficients. For a given target variable, parameter's values sampled from a full conditional posterior distribution are used to sample plausible values from its predictive posterior distribution. The algorithm can then be summarized as an iterative repetition of the following sampling steps:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(m)}_1 &\sim p(\boldsymbol{\theta}_1 | \boldsymbol{z}_{1,obs}, \boldsymbol{Z}^{m-1}_{1,obs}) \\
\boldsymbol{z}^{(m)}_{1,mis} &\sim p(\boldsymbol{z}_{1,mis} | \boldsymbol{Z}^{m-1}_{1,mis}, \hat{\boldsymbol{\theta}}^{(m)}_1) \\
&... \\
\hat{\boldsymbol{\theta}}^{(m)}_T &\sim p(\boldsymbol{\theta}_1 | \boldsymbol{z}_{T,obs}, \boldsymbol{Z}^{m-1}_{T,obs}) \\
\boldsymbol{z}^{(m)}_{T,mis} &\sim p(\boldsymbol{z}_{T,mis} | \boldsymbol{Z}^{m-1}_{T,mis}, \hat{\boldsymbol{\theta}}^{(m)}_T)
\end{aligned} \tag{2}$$

where $\hat{\boldsymbol{\theta}}_j^{(m)}$ are draws from the posterior defined with shrinkage priors at the m-th iteration. The superscript $(m-1)$ implies that the missing values in $\boldsymbol{Z}_{obs,j}$ and $\boldsymbol{Z}_{mis,j}$ are filled in with the imputations drawn at the previous iteration.

MICE with Bayesian Ridge (bridge) As "blasso", the "bridge" imputation procedure closely follows a standard MICE algorithm for imputation of multivariate missing data (van Buuren, 2012, p. 120, algorithm 4.3): for each target variable, at each iteration, plausible values of the imputation model parameters are drawn from their posterior distribution, and imputations are drawn from the posterior predictive distribution.

The sampling of imputation model parameters values is done as in the standard *Bayesian imputation under normal linear model algorithm* described by (van Buuren, 2012, p. 68, algorithm 3.1) and implemented in the mice package with the *impute.mice.norm()* function. The algorithm uses a ridge penalty to avoid problems of singular matrices. By doing so, it allows to perform Bayesian Multiple Imputation even with data affected by high collinearity and/or with a higher number of columns than rows ($p > n$).

MICE with PCA (MICE-PCA) Considering a data analysis task on a dataset $\boldsymbol{Z}$ with missing values, a set of T imputation target variables can be identified. These variables are afflicted by non-response and are part of some substantive model of interest, or in other words, there is a desire to obtain inferential conclusions based on their inclusion in said model.

The remaining variables in the dataset constitute a pool of possible *auxiliary* variables that could be used to improve the imputation procedure. In particular, one or more predictors of missingness could be among the auxiliary variables. By extracting Principal Components out of them, it is possible to summarise the information contained in this set with just a few components. This allows to include in a standard MICE imputation algorithm all the information contained in a large number of auxiliary variables without incrementing the dimensionality of the imputation models.

Note that if missing values are present in the set of auxiliary variables, one can fill them in with a stochastic single imputation algorithm of choice as the goal of said imputation would be to simply allow PCs extraction and not inferential.

The imputation procedure can be summarized as follows:

- Extract Principal Components from all variables in $\boldsymbol{Z}$ that are not part of set T
- Create a new data matrix $\boldsymbol{Z}^{'}$ by combining the target variables with the first principal components that explain 50% of the variance in the auxiliary variables.
- Use a standard MICE algorithm for imputation of multivariate missing data to obtain multiply imputed datasets from the low dimensional $\boldsymbol{Z}^{'}$.

Describe your implementation of the method proposed by Howard et al. (2015) and developed in the PcAux package.

MICE with regression trees (MI-CART and -RANF) Describe your implementation of the methods proposed by Burgette and Reiter (2010); Shah et al. (2014) and implemented in the mice R package.

MICE optimal model Describe how (van Buuren, 2012, p. ??) recommends we deal with the selection of the imputation model predictors. This method follows that approach but has also some oracle property: the variables responsible for missingness are always included in the imputation models.

2.2 Single data strategies

Single Imputation Describe missForest apporach.

Mean Imputation and Complete Case analysis Describe imputation using the mean of observed values. Describe deletion of rows with missing values.

3 Simulation Studies

3.1 Methods: two simulation studies

Data generations Describe normal multivairate distribution used for the first simulation study.

Describe latent variable data generation for the second simulation study.

Missing data imposition Describes how missing data are imposed on the generated data (target variables, response models)

Analysis model(s) Describe the saturated model (MLE estiamtes of means and variances), linear models,

Confirmatory Factor Analysis.

Criteria Description and formulas for bias and confidence interval coverage.

Description of multivairate measure of distance for groups of parameters

Procedure Summary description of crossed conditions.

Description of 1 data replication steps (data gen, imputation, anlaysis, pooling, averaging results)

3.2 Results

Bias Report results of comparison in terms of estimates bias for relevant parameters

Given concise idea of implications. Avoid higher level comparisons.

Confidence Interval Coverage Report results of comparison in terms of confidence interval coverage of the "true" values of parameters

Given concise idea of implications. Avoid higher level comparisons.

4 Resampling Study

4.1 Methods

Data preparation Give correct documentation and references on original EVS data.

Give description of EVS data used: why collected; what general demographics.

Give (concise) description of systematic cleaning process: general purpose of cleaning; large western european countries focus; reference appendix with detailed description.

Missing data imposition Describes how missing data are imposed on the generated data (target variables, response models)

Analysis model(s) Describe model 1: effect of dimensions trust on euthanasia acceptance

Describe model 2: effect of gender on left/right voting behaviour

Criteria Reference description in simulation study section and add specific details if needed.

Procedure Summary description of crossed conditions.

Description of 1 data replication steps (data gen, imputation, anlaysis, pooling, averaging results)

4.2 Results

Bias Report results of comparison in terms of estimates bias for relevant parameters.

Given concise idea of implications. Avoid higher level comparisons.

Confidence Interval Coverage Report results of comparison in terms of confidence interval coverage of the "true" values of parameters

Given concise idea of implications. Avoid higher level comparisons.

5 Discussion

Make parallels and comparisons between methods. Try to portray the general pattern that comes out of the combined results from the three studies.

6 Conclusions

Take-home message Give the take-home message in one or two paragraphs

Limitations and future directions Describe limitation with specific focus on what are your planned next steps in this line of research.

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.

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

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.

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.

Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198.

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

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 Bühlmann, 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.