OUTLINE OF PAPER 1
High Dimensional Imputation for the Social Sciences: A Review of State-of-the-Art Methods

This is the outline planned for the paper. It will not be included in the final work.

- **Introduction:** Frame problem; Discuss background literature; Focus/Reason to write paper; Content Summary.
- **Algorithms and Imputation methods:** Describe bridge, blasso, DURR, IURR, MI-PCA, etc.; Focus on minimal possible description to give reader sense of what the method is (max Deng et al. (2016); Reference papers for details.
- **Simulation Studies**
  - Methods for Study 1 (MVN) + Study 2 (Latent Structure)
    - Data generation
    - Missing data imposition
    - Analysis models
    - Criteria
    - Procedure: Summary of crossed conditions, describe sequentially what happens during each replication
  - Results: distinguish by type of performance measure
- **Resampling Study (EVS)**
  - Methods
    - Data preparation: documentation for the data; what is it; why collected; general original demographics of cases; selected demographics (e.g. western European Countries); systematic cleaning process with general purpose; reference to appendix for details.
    - Missing data imposition
    - Analysis models
    - Criteria
    - Procedure: Summary of crossed conditions, describe sequentially what happens during each replication
  - Results: again divide by type.
- **Discussion:** Synthesize findings, make parallels and comparisons.
- **Conclusions:** Short take home message, limitations, future directions (hint at MY future work)
- Appendices - methods details - EVS quirks

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

Edoardo Costantini

October 21, 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 multivariate missing data.

The tools researchers working with these data sets need to correct for the bias introduced by nonresponses require special attention. In general, when performing Multiple Imputation, data handlers tend to prefer including more, rather than less, predictors in the imputation models. This reduces the chances of uncongenial imputation and analysis models (Meng, 1994) and of leaving out important predictors of missingness. On top of this standard source of dimensionality, the large number of items recorded in surveys, coupled with their longitudinal nature, and the necessity of preserving complex interactions and nonlinear relations, easily produces high-dimensional ($p > n$) imputation problems.

When data is sparse (n not substantially larger than p) or afflicted by high collinearity (correlation among certain variables is so high that some of their linear combinations have no variance) the data covariance matrix is singular. Singular matrices are not invertible, an operation that is fundamental in the estimation of imputation models in any parametric Multiple Imputation procedure. As a result, high dimensionality of the data matrix prevents a straightforward application of imputation algorithms, such as MICE (van Buuren, 2012).

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 available 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.

(Discuss background literature) Many solutions have been proposed to deal with missing values in high dimensional contexts. Some researchers 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). However, the main task of social scientists is to make inference about a population based on a sample of observed data and single imputation is simply inadequate for this purpose: it does not guarantee unbiased and confidence valid estimates of the parameters of interest (Rubin, 1996).

Multiple Imputation is more suitable for the task. Its application to high dimensional data has been directly tackled by specific algorithms using either shrinkage or dimensionality reduction methods (Song and Belin, 2004; Zhao and Long, 2016; Deng et al., 2016). Furthermore, there are other methods, that could potentially suit well the purpose, but have been tested only in low-dimensional settings (Burgette and Reiter, 2010; Doove et al., 2014; Howard et al., 2015).

(Focus/Reason to write paper) With this article we set out to provide a comparison of these state-of-the-art imputation algorithms in high-dimensional scenarios. We compare methods based on their ability to allow inferential statements that are as valid as if they were made on a dataset without missing data. The comparison is developed both through simulation studies and a real survey data application.

(Content Summary) This paper is organized as follows. Section 2 discusses the imputation methods compared. Section 3 presents two simulation studies, their design and the result of the comparison. Section 4 presents a resampling study performed on the 2017 wave of the EVS. Section 5 discusses the implication of the combined results of the simulation and resampling studies. Finally, section 6 provides concluding remarks, description of the limitations of the study, and future directions we want to take.

2 Imputation methods and Algorithms

Consider a dataset $\boldsymbol{Z}$ of dimensionality $n \times p$, with n observations (rows) and p variables (columns). Assume there are $T < p$ variables with missing cases in at least one row that are also part of the substantive model of interest. An imputation procedure targeting these T variables could be used to allow fitting a substantive model (e.g. some linear or logistic regression) without discarding data units (rows). The $p-T$ variables in the dataset constitute a pool of possible *auxiliary* variables that could be used to improve the imputation procedure.

Most of the methods described in this section iteratively impute each target variable with imputation models that use as predictors the other target variables and the information contained in the auxiliary data.

2.1 Multiple Imputation Strategies

2.1.1 MICE with Bayesian Ridge (bridge)

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

After initialization of the missing values, at each m-th iteration, performs the following sampling steps for each target variable:

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

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

where $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ are draws from the parameters posterior distribution (1) and posterior predictive distribution (2), respectively, for the j-th target variable at the m-th iteration. The superscript $((m))$ implies that the missing values in $\boldsymbol{Z}_{obs,j}^{m}$ and $\boldsymbol{Z}_{mis,j}^{m}$ are different at every iteration as they are filled in with the previous iteration draws.

The sampling of each $\hat{\boldsymbol{\theta}}_j^{(m)}$ and $z_{j,mis}^{(m)}$ 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 as in the *impute.mice.norm()* function of the *mice* R package. The algorithm uses a ridge penalty to avoid problems of singular matrices. When the sample covariance matrix is singular, it is not invertible, an operation that is key to the sampling of parameters in (1) (Schafer, 1997). By adding a biasing ridge penalty, singularity is circumvented and the sampling scheme described above is possible even on data affected by high collinearity and/or with a higher number of columns than rows ($p > n$).

2.1.2 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 (Park and Casella, 2008; Hans, 2009) effectively performing a form of Bayesian model selection.

The Bayesian Lasso imputation algorithm (blasso) used here is a standard Multiple Imputation MCMC sampler that uses the shrinkage priors defined by Hans (2010) to compute the posterior distributions of the regression coefficients (which are used in (1)). Posterior parameters draws are then used to sample plausible values from the predictive distributions of the missing data. For a detailed description of the algorithm for Bayesian Lasso Multiple Imputation (blasso) in a univariate missing data context we recommend reading Zhao and Long (2016). The R code to perform blasso imputation is heavily based on the Bayesian Lasso R Package *blasso* developed by Hans (2010).

2.1.3 Direct Use of Regularized Regression (DURR)

As proposed by Zhao and Long (2016) and Deng et al. (2016), Regularized Regression can be directly used in a MICE algorithm to perform multiple imputation of high dimensional data. For a target variable $\boldsymbol{z}_j$, the DURR algorithm follows these directions:

- Generate a bootstrap sample $\boldsymbol{Z}^*$ by sampling with replacement rows of $\boldsymbol{Z}$. Denote $\boldsymbol{z}^*_{j,obs}$ and $\boldsymbol{Z}^{*(m)}_{j,obs}$ as the observed part of z^*_j and the corresponding values on the other variables in $\boldsymbol{Z}^*$, respectively. Suffix m is used to clarify that at each iteration $\boldsymbol{Z}^{*(m)}_{j,obs}$ is different as it includes values previously imputed on the other target variables.

- Use any regularized regression method (such as Lasso regression) to fit a linear model with $\boldsymbol{z}_{j,obs}$ as outcome and $\boldsymbol{Z}^{*(m)}_{j,obs}$ as set of predictors. This produces a set of parameter estimates (regression coefficients and error variance) $\hat{\boldsymbol{\theta}}^{(m)}_j$ that can be considered as sampled from the parameters' posterior distribution conditioned on the observed part of the data (1).

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

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

2.1.4 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 with a standard multiple imputation procedure (Zhao and Long, 2016; Deng et al., 2016). 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 $\boldsymbol{Z}^{(m)}_{j,obs}$ as predictors (compared to DURR, there is no asterisk in the notation as the original data is used, not a bootstrap version). 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.

$$(\hat{\boldsymbol{\theta}}_j^{(m)}, \hat{\sigma}_j^{(m)}) \sim N(\hat{\boldsymbol{\theta}}_{MLE}^{(m)}, \hat{\boldsymbol{\Sigma}}_{MLE}^{(m)}) \tag{3}$$

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

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

2.1.5 MICE with PCA (MICE-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 perform a standard MICE algorithm in a well-behaved low dimensional setting. The MICE-PCA 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 cumulative explain at most 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 $\boldsymbol{Z}^{'}$ and the set of target variables.

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. This method is inspired by Howard et al. (2015) and the *PcAux* package that implements and developed its ideas.

2.1.6 MICE with regression trees (MI-CART and -RANF)

A variety of Multiple Imputation methods using regression and classification trees have been proposed (Reiter, 2005; Burgette and Reiter, 2010; Shah et al., 2014) They all share the following core steps:

- For a given variable z_j, target of imputation, a CART algorithm partitions $\boldsymbol{Z}_{j,obs}^{(m)}$ to identify a collection of leafs with homogeneous $z_{j,obs}$ values. Each leaf contains a subset of the observed z_j, called donors.
- Each unit with a missing value on the target variable is placed in one of the leafs based on its $\boldsymbol{Z}_{j,mis}^{(m)}$ values.
- Each missing value on z_j is sampled from the pool of corresponding leaf donors.

At iteration m, these steps are followed for all of the T target variables. After convergence, the last M datasets are kept as multiply imputed datasets that can be used for the analysis and pooling phases.

The implementation of MI-CART used in this paper corresponds to the one presented in (Doove et al., 2014, p. 95, algorithm 1) and the *impute.mice.cart()* R function from the mice package.

The Multiple Imputation with Random Forest algorithm (MI-RANF) used in this paper is an adaptation of the one described for MI-CART. To impute z_j at iteration m, MI-RANF first draws K bootstrap samples from the rows of the data with observed z_j. One tree is fitted to every bootstrap sample, with random features selection, and donors are identified. Imputations are then drawn from a pool of donors combined from the K trees that have been fitted to Z_{obs}. Imputations are not sampled from donor values averaged across trees as this procedure would reduce the uncertainty incorporated in the imputation model.

For greater details on the algorithms, the reader may consult algorithm A.1 in (Doove et al., 2014, p. 103, appendix B). The programming of the algorithm was heavily inspired by the *impute.mice.rf()* function in the R package *mice*.

2.1.7 MICE optimal model

We have also used an ideal standard MICE with Bayesian Linear Regression approach (MI-OP) that considered, for each target variable imputation model, the following groups of predictors:

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

Following these criteria is one of the most recommended strategies to deal with a large number of possible imputation model predictors (van Buuren, 2012, p. 168). In this sense, it represents an *ideal* strategy that could be used to deal with high-dimensional data, in the absence of alternatives. In practice, researchers can never be sure requirement 2 is fulfilled, as there is no way to know exactly which variables are responsible for missingness. The MI-OP approach used here remains *ideal* in the sense that it is not applicable in practice, but it does offer an interesting benchmark case.

2.2 Single data strategies

2.2.1 Single Imputation

We consider the MissForest imputation method proposed by Stekhoven and Bühlmann (2011). Being a non-parametric imputation approach it does not suffer from the problem of unidentified imputation models and it can accommodate for mixed data type of the missing variables. However, as a single

imputation method we do not expect it will allow to perform statistically valid inference on the treated data.

2.2.2 Mean Imputation and Complete Case analysis

In the social sciences, and especially in the analysis of social surveys, imputing the mean of the observed values on a variable is still a quite popular choice in dealing with missing data. Therefore, we include this method to portray a picture of the possible improvements the different high-dimensional imputation algorithms can achieve.

Finally, for the sake of comparison, two additional approaches are considered that do not involve imputation: list-wise deletion (or CC, complete case analysis), which entails fitting the analysis models exclusively on the complete rows of the data; and a gold standard analysis (GS) which consists of fitting the substantive models on the underlying fully observed data and represents the counterfactual analysis that would have been performed if there had been no missing data.

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.

Doove, L. L., Van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

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.

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.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 21(3):7–30.

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

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

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.

Song, J. and Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, 23(18):2827–2843.

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.