Missing Data Analysis
High-Dimensional Imputation

Compiled by Kyle M. Lang

2019-10-28

The sources listed below represent an overview of the work on high-dimensional missing data imputation. This list is certainly not exhaustive—and may not be especially representative—but it should provide a good starting point for readers interested in learning about the current state-of-the-art in missing data imputation for high-dimensional problems.

Although most of these sources discuss/evaluate several different imputation methods, I have attempted to classify each paper by its primary focus.

Overviews/Comparisons

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, *55*(12), 3232–3243. doi: 10.1016/j.csda.2011.06.006

> The authors compare four Machine Learning methods to create synthetic datasets: (sequential/repeated) CART, implemented in a similar fashion to Burgette and Reiter (2010); CART applied to bootstrapped subsamples of the original dataset (random forests and bagging); and Support Vector Machines. These prediction methods are expected to generate synthetic datasets that preserve the original complex data structure, and hence grant statistical validity of any secondary analyses carried out on them.
>
> Generating a synthetic dataset can be thought of as imputing a dataset with missing data. The issue of granting statistical validity of the analyses performed on a partially synthetic dataset is the same as that of obtaining statistically valid analyses on an imputed dataset. Therefore, these findings are directly applicable to the high-dimensional missing data handling problem. The study shows sequential/MI CART and SVM synthetic dataset generation (imputation) procedures outperform both random forests and bagging approaches.

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing & Applications*, *19*(2), 263–282. doi: 10.1007/s00521-009-0295-6

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907. doi: 10.1016/j.atmosenv.2004.02.026

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), 259–275. doi: 10.1023/A:1008334909089

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, *8*, 1217–1250.

Single Imputation using Classification/Regression Trees

Borgoni, R., & Berrington, A. (2013). Evaluating a sequential tree-based procedure for multivariate imputation of complex missing data structures. *Quality & Quantity*, *47*(4), 1991–2008. doi: 10.1007/s11135-011-9638-3

The authors integrate a sequential (one-variable-at-the-time + data augmentation) tree-based imputation algorithm and non-parametric bootstrap. Specifically, given a set of fully observed inputs and a set of target variables with missing values, their approach grows a tree on a complete subset of the instances for each target variable sequentially (starting with the target variable with fewest missing cases, uses the imputed values to augment the original data, and improve imputation of variables with more missing values), and iterates this procedure until some criteria is reached. To account for the added uncertainty due to imputation, the authors apply this method to a large number of bootstrap samples from the original dataset.

Considering only categorical variables, the authors compare mode, sequential regressions, a non-iterative tree-based imputations to their bootstrap tree-based imputation algorithm. Most notably, the results of the simulation study point out that non-iterative regression trees are worst than simple mode imputation in terms of estimation accuracy (bias and efficiency of a true model parameters estimates), and that their bootstrap version of an imputing decision tree compensates well for the extra variability added by the imputation process.

Conversano, C., & Siciliano, R. (2009). Incremental tree-based missing data imputation with lexicographic ordering. *Journal of Classification*, *26*(3), 361–379. doi: 10.1007/s00357-009-9038-8

The authors introduce the Incremental NonParametric Imputation (INPI) algorithm for imputing large datasets with missing values on multiple categorical and/or continuous variables. This approach reorganises the data according to a lexicographic order (arranging the columns and rows of a data matrix so as to concentrate the missing values in a corner), grows a decision tree with a FAST algorithm (selecting first the best splitting predictor and then the best splitting cut off value), imputes a missing value that minimises a given error/decision rule, and augments the complete data before repeating the process for all the other missing values.

The INPI algorithm exploits all of the advantages of decision tree methods (i.e. nonparametric nature, flexibility to variable types) and grants an effective imputation method for large datasets. Another advantage of the is that the algorithm can be easily extended to include imputation model uncertainty (Multiple Imputation). However, it must be noticed that this algorithm assumes a MAR mechanisms and that there is at least one completely observed variable in the original dataset.

D'Ambrosio, A., Aria, M., & Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of Classification*, *29*(2), 227–258. doi: 10.1007/s00357-012-9108-1

The imputation algorithm proposed in this article improves on the INPI algorithm proposed by Conversano and Siciliano (2009) in two ways: by incrementally imputing one variable at the time, it reduces the computational effort compared to the imputation of single missing data points one at the time; by using the ensemble learning classifier AdaBoost, instead of single decision tree, the new algorithm is more accurate, in terms of prediction error.

This article also integrates this imputation approach in a "data fusion" problem. This could be interesting for imputation problems with datasets gathered according to planned missing data designs.

The article interestingly includes the MICE approach as one of the imputation methods compared. However, it does not provide a valid measure of comparison between MICE and BINPI. The authors stress how the performance of a BINPI algorithm should be evaluated based on its ability to recover the missing values (according to the Statistical Learning Theory framework). However, MICE is an imputation strategy grounded in the stochastic framework proposed by Rubin. Hence, MICE should be evaluated not in terms of the ability to recover missing values (prediction error), but in terms of how statistically valid the analysis that it allows to perform are.

Iacus, S. M., & Porro, G. (2007). Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics & Data Analysis*, *52*(2), 773–789. doi: 10.1016/j.csda.2006.12.036

Nanni, L., Lumini, A., & Brahnam, S. (2012). A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, *55*(1), 37–50. doi: 10.1016/j.artmed.2011.11.006

Single Imputation using K-Nearest Neighbors

Brás, L. P., & Menezes, J. C. (2007). Improving cluster-based missing value estimation of dna microarray data. *Biomolecular engineering*, *24*(2), 273–282.

de Andrade Silva, J., & Hruschka, E. R. (2009). Eacimpute: an evolutionary algorithm for clustering-based imputation. In *2009 ninth international conference on intelligent systems design and applications* (pp. 1400–1406).

de Andrade Silva, J., & Hruschka, E. R. (2013). An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data & Knowledge Engineering*, *84*, 47–58. doi: 10.1016/j.datak.2012.12.006

The article compares 5 popular Nearest-Neighbour algorithms based on both their prediction accuracy and the classification bias. The results show that in a MCAR context, the Iterative KNNImpute algorithm proposed by Bràs and Menezes (2007) outperforms all other methods, according to both the Normalised Root Mean Squared Error (NRMSE) and the Average Correct Classification Rate (ACCR) criterion. However, in a MAR scenario, the KNNI proposed by Troyanskaya et al 2001, the Sequential KNNI proposed by Kim et al (2004), and the EACI proposed by de Andrade Silva and Hruschka (2009) perform best.

One of the clearest contribution of the article is showing that better prediction accuracy does not necessarily imply a better modelling performance (i.e. an algorithm might "recover" the missing values better, but another algorithm might provide a better estimation of the relationship between attributes). This result motivates a departure from the exclusive use of the Normalised Root Mean Squared Error measure of prediction accuracy as the criteria to evaluate an imputation algorithm. It must be noted that this article is concerned exclusively with a classification task.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2009). *K* nearest neighbours with mutual information for simultaneous classification and missing data imputation. *Neurocomputing*, *72*(7–9), 1483–1493. doi: 10.1016/j.neucom.2008.11.026

The authors introduce an improved version of the KNNImpute algorithm proposed by Troyanskaya *et al.* 2001. Their proposed strategy is called MI-KNNImpute, a somewhat confusing label: the "MI" portion does not refer to "multiple imputation" but to the concept of Mutual Information (i.e. the reduction of the uncertainty of a variable when another one is known).

The authors show with different incomplete datasets that the performance of the classification task, performed by a KNN algorithm that uses a distance measure of Euclidian form and one that uses MI, is improved by first imputing the datasets with the MI-KNNImpute algorithm compared to the standard KNNImpute.

The improved performance is achieved by including some information regarding the classification task in the imputation phase, through the use of MI as a measure of distance between the target class variable and the other input attributes. The use of MI in the algorithm effectively weights the importance of each attribute for the imputation of the target class variable.

Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, *21*(2), 187–198. doi: 10.1093/bioinformatics/bth499

The auhtors developed the Local Least Square imputation (LLSImputation)

method by which similar k-neighbour genes are selected, then used to estimate a prediction model which is finally employed to predict the missing values.

The article compares LSSImputation with the KNNImputation and SVDimputation proposed by Troyanskaya et al (2001), and the Bayesian PCA proposed by Oba *et al.* (2003). As the BPCA approach improves on the SVDImputation by incorporating Bayesian optimisation in a PC based method, LLSI improves on KNNI by combining the local similarity structure and the optimisation procedure of least squares.

Finally, the article directly confronts the issue of choosing the optimal number of k-nearest neighbours. In absence of clear theory, the authors propose to empirically identify on a case-by-case basis the value of k by repeatedly predicting some artificially imposed missing values and selecting the one that best recover the known fabricated missing values.

Kim, K.-Y., Kim, B.-J., & Yi, G.-S. (2004). Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC bioinformatics*, *5*(160). doi: 10.1186/1471-2105-5-160

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. doi: 10.1093/bioinformatics/17.6.520

In the context of DNA microarrey studies with missing data, the authors show the better performance of a k-nearest neighbour imputation method (KNNImpute), over an SVD-based regression imputation method (SVDImpute), and mean and zero imputation.

The KNNImpute algorithm outperforms all the other methods granting (1) less deterioration in performance with increasing percentage of missingness, (2) robustness to the type of data considered (time-series or not, noisy or not), (3) less sensitivity to the number of parameters used (the choice of k, the number of nearest neighbours considered for KNNImpute and the number of most significant eigengenes selected for SVDImpute).

The generalisability of these results is somewhat hindered by two methodological choices:

- the missing data mechanisms considered is MCAR;
- the metric used to identify the better method is the Root Mean Squared error, a standardised difference between the true data points values and the imputed ones. As many contributions have highlighted (see Rubin, 1996; de Andrade Silva *et al.*, 2013), the goal of missing data handling procedures is not necessarly recovering the true missing values but rather granting the statistical validity of the analyses performed on a dataset afflicted by missing data.

Wasito, I., & Mirkin, B. (2005). Nearest neighbour approach in the least-squares data imputation algorithms. *Information Sciences*, *169*(1–2), 1–25. doi: 10.1016/j.ins.2004.02.014

The authors introduce the INImpute approach, an algorithm that combines an iterative SVD-based least-square imputation with a nearest neighbour approach. Missing values are imputed first globally (i.e. considering the entire

dataset) through an Iterative Majorization Least Square algorithm (an SVD/PCR-based imputation method with 4 principal components, p=4). Then, a kNN algorithm is used to select k-nearest-neighbours to an instance that had a missing value for a particular variable, and replaces the previously (globally) imputed value with a new one found with anoter IMLS run only among the nearest-neighbours (and with p=1).

The theoretical properties of this approach are not discussed but the authors do show the superiority of INImpute to regular kNN and other ILS approaches in a variety of scenarios.

Wasito, I., & Mirkin, B. (2006). Nearest neighbour approach in the least-squares data imputation algorithms with different missing patterns. *Computational Statistics & Data Analysis*, *50*(4), 926–949. doi: 10.1016/j.csda.2004.11.009

The authors extend the comparison of the different least-squares imputation techniques performed by Wasito and Mirkin (2005) to accommodate for different missing data mechanisms (MCAR, MNAR and a merged data missingness).

The results mainly show that all versions of Nearest-Neighbour-based least-squares imputations presented outperform the global versions. Furthermore, the global-local INI (IMLS-NN-IMLS) approach wins in almost all contexts except when only the local version (the NN-IMLS) wins. The article also includes a detailed description of a data generating process according to the Neural Network NetLab framework, and of the missing data patterns generation. Hence, it is a good reference for anyone planning a simulation study to test imputation methods.

Zhang, S. (2012). Nearest neighbor selection for iteratively *k*NN imputation. *The Journal of Systems and Software*, *85*(11), 2541–2552. doi: 10.1016/j.jss.2012.05.073

The author introduces the Gray k-Nearest-Neighbour (GkNN) algorithm as an improvement to the traditional kNNImpute algorithm (see Troyanskaya *et al.* 2001). This new version uses a Gray Relation Grade measure of similarity, from Gray Relational Analysis, which easily accommodates for both numerical and categorical input variables.

Furthermore, GkNN is an iterative algorithm rooted in the EM framework, and the author claims that such feature makes the algorithm able to account for the uncertainty related to the imputation procedure. However, it is not explicit how this algorithm takes into account the additional uncertainty regarding parameters estimates.

Finally, the GkNN algorithm uses all the information included in the dataset by including in the imputation of an instance i, the observed and imputed values of other instances that have missing values.

Multiple Imputation

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. doi: 10.1093/aje/kwq260

The authors propose a multiple imputation via sequential decision trees to deal with large datasets with missing data on many predictors. The approach is rooted in the MICE algorithm. The main point of departure is the way predictive distribution draws are done. For each target variable with missing values a tree is grown on the remaining complete (augmented) dataset. The imputations are done by locating each case with a missing value on the current target variable in a leaf of such a tree and attributing a randomly sampled target variable value among the observed ones in that leaf.

In the simulation setup, the approach proposed works fine in terms of providing low bias of the parameter estimates of the data generating model fitted to the imputed data. However, the coverage rates are alarmingly low.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, *6*, 21689. doi: 10.1038/srep21689

The authors carried out a comparison study similar to that of Zhao and Long (2013) in order to assess the efficiency of regularised regression for multiple imputation of high-dimensional data. One key contribution is generalising the univariate missing case to multivariate missing patterns. However, the study considers at most three variables afflicted by missing values (at least in the simulation study).

The study shows a clear dominance of MICE-IURR (MICE with indirect use of regularised regression) in terms of bias compared to all other viable methods. In terms of coverage rate, there is no clear winner. Notably, the Random Forest MICE approach proposed by Shah et al 2014 does not live up to the promising findings of this previous study (the unsatisfactory performance of MICE-RF is corroborated by Drechsler and Reiter, 2011). Finally, the result shown for MICE-DURR are also in disagreement with what was found by Zhao and Long (2013).

He, R., & Belin, T. (2014). Multiple imputation for high-dimensional mixed incomplete continuous and binary data. *Statistics in Medicine*, *33*(13), 2251–2262. doi: 10.1002/sim.6107

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

The author proposes the use of CART to generate synthetic data sets (i.e. substituting sensitive observed data points with multiply imputed values to avoid disclosure of information). For each of many subsamples of the original dataset, the algorithm sequentially grows a tree for each of the k-th sensitive variables using all the other variables (X, Y_{-k}) as inputs, and then imputes the

values to be replaces (the "missing data points") though bootstrap sampling of the Y_k values in the leaf where each observation falls.

The literature on using CART and other synthetic data generator methods is closely related to that of missing data imputation. However, a few remarks are due. The first body of literature is concerned with balancing between guarantying statistical validity of the secondary analyses and reducing the risk of sensitive information disclosure, while missing data-handling literature is exclusively concerned with the statistical validity issue. This is important because the different goal influences the tree pruning strategy. Furthermore, the concept of missing data mechanism is not relevant for the synthetic data literature and replacement (missingness) is forced on either the entire variable or just specific ranges of it (e.g. replace (impute) values of income which are greater then a threshold). The counterpart scenario in a missing data context is imputing a fully missing variable with especially good initial guesses.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & 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., & Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, *23*(18), 2827–2843. doi: 10.1002/sim.1867

Wallace, M. L., Anderson, S. J., & Mazumdar, S. (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in Medicine*, *29*(29), 3004–3016. doi: 10.1002/sim.4079

The authors propose a multiple imputation decision tree method that includes uncertainty regarding the imputed values. This is achieved by slightly modifying the single imputation tree-based algorithm proposed by Conversano and Siciliano (2003): the tree is grown M times, and, after the first iteration, the imputed variables are considered as possible candidates for the split as well as the complete ones; once the best splitting variable and values are chosen, the algorithm classifies the cases with missing values on the target variable, based on their observed values on the other features, in a given node, and imputes that node mean value of the target feature. Each time a value is imputed with the node's mean, a random error is added to it. As a result each missing data point is filled with M different values, allowing the algorithm account for uncertainty of the imputed value.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, *25*(5), 2021–2035. doi: 10.1177/0962280213511027

The authors discuss the used of chained equations regularised regression imputation that enables missing data handling within the Multiple Imputation framework with high-dimensional data ($p > n$). Three main approaches are proposed: a direct use of regularised regression on multiple bootstrapped datasets (DURR), an inderect use of it, and finally a bayesian lasso approach.

The method BLasso method seems to outperform the other ones. Furthermore, it can easily be extended to general missing data patters (the

study only shows results for univariate missing data), while the other methods do not share such feature. However, the BLasso approach becomes computationally extremely intensive (even infeasible) when many variables are afflicted by missing values.