

# Missing Data Analysis High-Dimensional Imputation

Compiled by Kyle M. Lang

2019-10-15

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
- 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
- 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

- 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
- 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

- 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
- 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
- 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 authors 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 Troyan et al (2001), and the Bayesian PCA proposed by Oba *et al.* (2003). As the BPCA approach improves on the SVDImputation by incorporating Bayesian optimization in a PC based method, LLSI improves on KNNI by combining the local similarity structure and the optimization 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 recovers the known fabricated missing values.

- 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 microarray 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), the goal of missing data handling procedures is not recovering the true missing values but granting the statistical validity of the analyses performed on a dataset afflicted by missing data. It is also true that such arguments usually apply in the context of multiple and not single imputation.

- 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
- 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
- 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

## 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
- 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
- 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.
- 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
- 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