Advanced Introductory Reading List: Missing Data Analysis

Compiled by Kyle M. Lang

September 16, 2019

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, *17*(1), 71–103.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*(278), 200–203.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy *c*-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25–35.

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

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

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

Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Pscyhological Methods*, *6*(4), 330–351. doi: 10.1037//1082-989X.6.4.330

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

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38. doi: 10.2307/2984875

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

Fessant, F., & Midenet, S. (2002). Self-organising map for data imputation and correction in surveys. *Neural Computing and Applications*, *10*(4), 300–310. doi: 10.1007/s005210200002

Gabrys, B. (2002). Neuro-fuzzy approach to processing inputs with missing values in pattern recognition problems. *International Journal of Approximate Reasoning*, *30*(3), 149–179.

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

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2013). Classifying patterns with missing values using multi-task learning perceptrons. *Expert Systems with Applications*, *40*(4), 1333–1341. doi: 10.1016/j.eswa.2012.08.057

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

Gheyas, I. A., & Smith, L. S. (2010). A neural network-based framework for the reconstruction of incomplete data sets. *Neurocomputing*, *73*(16–18), 3039–3065. doi: 10.1016/j.neucom.2010.06.021

Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. doi: 10.1207/S15328007SEM1001_4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–576.

Gupta, A., & Lam, M. S. (1996). Estimating missing values using neural networks. *The Journal of the Operational Research Society*, *47*(2), 229–238. doi: 10.2307/2584344

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. doi: 10.1111/j.1540-5907.2010.00447.x

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

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

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. doi: 10.1016/j.artmed.2010.05.002

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

Kang, H. M., & Yusof, F. (2012). Application of self-organizing map (SOM) in missing daily rainfall data in malaysia. *International Journal of Computer Applications*, *48*(5), 23–28.

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.

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

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

Lang, K. M., & Wu, W. (2017). A comparison of methods for creating multiple imputations of nominal variables. *Multivariate Behavioral Research*, *52*(3), 290-304. doi: 10.1080/00273171.2017.1289360

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical variables with missing values. *Biometrika*, *72*, 497–512.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2013). On the joys of missing data. *Journal of Pediatric Psychology*, 1–12. doi: 10.1093/jpepsy/jsto48

Luengo, J., García, S., & Herrara, F. (2010). A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method. *Neural Networks*, *23*(3), 406–418. doi: 10.1016/j.neunet.2009.11.014

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*(302), 415–434.

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

Nordbotten, S. (1995). Editing statistical records by neural networks. *Journal of Official Statistics*, *11*(4), 391–411.

Nordbotten, S. (1996). Neural network imputation applied to the norwegian 1990 population census data. *Journal of Official Statistics*, *12*(4), 385–401.

Piela, P. (2002). Introduction to self-organizing maps modelling for imputation—techniques & technology. *Research in Official Statistics*, *2*, 5–19.

Polikar, R., DePasquale, J., Mohammed, H. S., Brown, G., & Kuncheva, L. I. (2010). Learn++.MF: A random subspace approach for the missing feature problem. *Pattern Recognition*, *43*(11), 3817–3832. doi: 10.1016/j.patcog.2010.05.028

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

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

Rey-del-Castillo, P., & Cardeñosa, J. (2012). Fuzzy min–max neural networks for categorical data: application to missing data imputation. *Neural Computing & Applications*, *21*(6), 1349–1362. doi: 10.1007/s00521-011-0574-x

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical*

Association, *91*(434), 473–489.

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

Samad, T., & Harp, S. A. (1992). Self-organization with partial data. *Network: Computation in Neural Systems*, *3*(2), 205–212. doi: 10.1088/0954-898X/3/2/008

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of state of the art. *Psychological Methods*, *7*(2), 147–177. doi: 10.1037//1082-989X.7.2.147

Sharpe, P. K., & Solly, R. J. (1995). Dealing with missing values in neural network-based diagnostic systems. *Neural Computing & Applications*, *3*(2), 73–77. doi: 10.1007/BF01421959

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la Vega, M.-D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, *24*(1), 121–129. doi: 10.1016/j.neunet.2010.09.008

Song, Q., Shepperd, M., Chen, X., & Liu, J. (2008). Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation. *The Journal of Systems and Software*, *81*(12), 2361–2370. doi: 10.1016/j.jss.2008.05.008

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

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

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivarite imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117.

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*(1), 265–291.

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

Wang, X., Li, A., Jiang, Z., & Feng, H. (2006). Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, *7*(32), 1–10. doi: 10.1186/1471-2105-7-32

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

Yoon, S.-Y., & Lee, S.-Y. (1999). Training algorithm with incomplete data for feed-forward neural networks. *Neural Processing Letters*, *10*(3), 171–179. doi: 10.1023/A:1018772122605

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

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

Zhu, R., & Kosorok, M. R. (2012). Recursively imputed survival trees. *Journal of the American Statistical Association*, *107*(497), 331–340. doi: 10.1080/01621459.2011.637468