PhD Research Proposal

About the Evaluation of Missing Data Methodologies

Author: Rianne Schouten
Supervisors: prof.dr. Stef van Buuren, dr. Gerko Vink
Supported by: Fred Kroon, DPA Data Science

September 18, 2017

1 Introduction

The occurence of missing data may be problematic anywhere data is being stored, processsed or analyzed. This applies to scientific research where datasets are incomplete due to non-response, drop-out and other forms of missing data. But the available datasets in an application-oriented field like data science suffer from incompleteness as well. In order to make sensible policy decisions based on reliable analysis results, a proper handling of the unobserved data is essential. Therefore, my research will focus on the evaluation of methods dealing with these missing data.

Of all possible methods for dealing with missing data, the most advantageous missing data method for a specific missing data problem is determined by - among others - the scientist's aim with the data. It is important to realize this aim could differ per discipline or per field. For instance, data scientists generally predict the value (or category) of an output variable. In this *prediction* domain, the goal of data analysis is to obtain the model with the most accurate predictions (James, Witten, Hastie, & Tibshirani, 2014). In contrast, scientific researchers aim at drawing valid statistical inferences about a population. More specifically, they intent to investigate how an output variable changes as a function of certain input variables. Here, the aim of data analysis is to explore the relation between variables. We call this statistical *inference* making (James et al., 2014).

It depends on the research goal of a given situation whether a missing data method is appropriate to deal with the missing values. For data scientists, fast and easy methods - such as deleting incomplete rows from the dataset (i.e. drop) or imputing the missing values with the mean or median of a feature - are appealing approaches. In addition, a data scientist may choose to use crossvalidation techniques to compare the performance of more than one missing data methods. Eventually, the appropriateness of a missing data methodology is determined by its capability to accurately predict the output variable (Hastie, Tibshirani, & Friedman, 2009). In the world of inference making, on the other hand, missing data methods are evaluated by their ability to enable the finding of valid statistical inferences. Suitable missing data methods are methods that result in unbiased population estimates without disturbing the covariance structure of the data. In general, multiple imputation is considered to be a useful technique (Van Buuren, 2012).

Great part of missing data theory is written in the context of statistical inference making (Rubin, 1976, 1987; Schafer & Graham, 2002). As a consequence, we know a lot about subtle differences between missing data methodologies, what to do in special cases such as surveys with small sample sizes and how to deal with missing data in particular types of analyses such as longitudinal data analysis and multilevel analysis. In addition, several authors have provided us with systematic comparisons of missing data methods (e.g. Peeters, Zondervan-Zwijnenburg, Vink, & Van de Schoot, 2015) and their software implementations (e.g. Horton & Kleinman, 2007). However, the translation of all this knowledge to situations where prediction is the main purpose of an analysis seems to occur only sporadically. For example, data science books use only 1 or 2 pages to describe missing data methods (Hastie et al., 2009; James et al., 2014). Remarkably, the few missing data methods that are discussed, are particularly those methods that scientific researchers consider inappropriate for most analyses (i.e. listwise deletion, mean imputation). Other places where data scientists may search for the knowledge they need, such as blogposts or data science journals, also rarely cover the topic of missing data.

With my research, I intent to form a bridge between two fundamentally different worlds that both have to deal with missing data. Data scientists could benefit from the large collection of scientific literature, provided that a sufficient translation and application of this knowledge is available. On the other hand, missing data methodologists may use the far-reaching experience of data scientists with machine learning techniques to further develop and implement missing data methods. In addition, data scientists have a very practical and standardized way of working. A standardized approach of missing data problems may be valuable for scientific researchers as well and the data science practice could serve as an example for the development of clear and logical algorithms.

2 Aims

As the title of this document points out, my research focuses on the evaluation of missing data methodologies. A concept that will be key in my work is perfectly formulated by Schafer and Graham (2002, p. 149): "A missing value treatment cannot be properly evaluated apart from the modeling, estimation, or testing procedure in which it is embedded." With this in mind, I started my project with the development and implementation of a multivariate amputation procedure (`ampute`: Schouten, Lugtig, Brand, & Vink, 2017). The availibility of `ampute` provides a way to systematically explore the effects of missing data problems on statistical analyses. Furthermore, it has now become straightforward to set up a simulation study and to evaluate the performance of missing data methods. Especially `ampute`'s ease to generate different missing data problems is valuable for testing realistic missing data situations. With the strong fundament of `ampute`, I will continue to build four pillars of missing data knowledge: accuracy, robustness, standardization and fusion.

2.1 Accuracy

The few data science documents that discuss the use of missing data methods evaluate the performance of these methods with so-called accuracy measures. Examples of these mesures are the confusion matrix, precision and recall measures and the area under the ROC curve for classification tasks, and mean squared error (MSE) and explained variance (R^2) for regression tasks. It is easy to obtain accuracy measures with the `.metrics` module from Scikit-learn in Python (Pedregosa et al., 2011). Since the goal of data science is to predict the value or class of an output variable - thus, to obtain the best accuracy - researchers evaluate the performance of missing data methods with accuracy measures as well. For instance, Acuña and Rodriguez (2004, p. 6) used "the 10-fold crossvalidation estimates of the misclassification error." Apart from factors such as efficiency and computation time, the missing data method that gives the best accuracy is considered to be the best way to deal with the missing values.

Existing research about the effect of missing data methods on accuracy measures could and should be extended. For instance, the effect of missing data on continuous predictions needs evaluations. Comparable to the way Garcia-Laencina, Sancho-Gomez, and Figuiras-Vidal (2010) and Acuña and Rodriguez (2004) showed the performance of missing data methods in classification tasks, I will set up simulation studies to evaluate the effect of missing data methods on continuous measures. Examples are the MSE and R^2. Furthermore, missing data methods other than complete case analysis (i.e. drop) or mean imputation need evaluation, such as regression imputation, stochastic regression imputation and multiple imputation. In scientific research, stochastic regression imputation and multiple imputation are favored over the other methods, because they do not disturb the covariance structure of the data. Multiple imputation, in particular, is known to give valid statistical inferences in many circumstances (Van Buuren, 2012). However, due to the Bayesian nature of the method, implementation of multiple imputation in the crossvalidation framework is not so straightforward. More work on this has to be done.

Two other aspects of dealing with missing data in the data science framework require consideration: 1) which variables contain missing values? and 2) are test and validation datasets complete or incomplete? Because a data scientist aims to predict an output variable, it is quite common to drop records from the dataset with a missing value on this output variable. This specific record in the data is considered to have *no label* (Hastie et al., 2009; James et al., 2014). However, what happens if the unlabeled records are different from the labeled records? What if you want to predict these unlabeled cases in real testsets but the model did not account for the missingness? An approach known as semi-supervised learning could be a solution. Here, only labeled records are selected for the validation sets, but unlabeled records are allowed to be part of the trainingset (Hastie et al., 2009; James et al., 2014). A downside of this approach is the selection effect that may occur for some, common, types of missingness. It would be useful to know for which types of missingness this is the case and if so, what other crossvalidation techniques can be used instead?

While performing all these studies, an important question will arise: "What value is considered to be the truth?" According to Garcia-Laencina et al. (2010, p. 280), "in classification tasks with missing values, the main objective of an imputation method must be to help to enhance the classification accuracy." However, does this statement indicate that any thought about the origin of the missing data, the type of missingness and its effect on the distribution of the data should simply be ignored? If that is true, then imputation of missing values with the same model as the prediction model will always give the best results. For instance, if a linear regression model is performed and MSE is taken to be the evaluation measure, then imputation of the missing values on a prediction line will definitely decrease the MSE. But does that mean that regression imputation is the appropriate missing method? According to Van Buuren (2012, p. 46), "we cannot evaluate imputation methods (...) by their ability to optimize classification accuracy." When is this statement true? And when is it not? Shah, Bartlett, Carpenter, Nicholas, and Hemingway (2014, p. 772) mention that "better predictions do not mean better coverage of confidence intervals; it is important that imputation methods incorporate the correct amount of variation in order to produce unbiased estimates with correct coverage of confidence intervals." But does this apply to data science use cases as well? In other words, while evaluating missing data methodologies, what is our reference and why? Could it be that the truth changes when the aim of an analysis changes? All these questions need answers if we wish to exchange missing data knowledge between scientific researchers and data scientists.

2.2 Robustness

The difference between (completely) random and non-random missing data problems - MCAR, MAR and MNAR respectively - is a widely discussed topic in missing data theory (Rubin, 1976, 1987). In general, a missing data method such as multiple imputation will yield valid statistical estimates under MCAR and MAR missingness, and has the potential to deal with MNAR data (Rubin, 1987; Van Buuren, 2012). The presence of the latter in practical situations is quite realistic, as shows an example at Philips Research Europe. Scientist Mariana Simons-Nikolova explained to me that as part of the development of medical technology, Philips invites patients to repeatedly visit a test clinic to measure certain health qualities such as blood pressure and heart rate. With these data, Philips is able to estimate and improve the performance of their technology. However, patients with acute or increased health issues may temporarily be prevented from visiting these test clinics and as a result, Philips' observed data contains the medical results from patients who are, on average, healthier than the total patient population. This type of non-response can be seen as a form of MNAR data: the missing values themselves determine whether they are missing are not.

The robustness pillar of my research will focus on the question: Under which circumstances are missing data methods robust against MNAR? Specifically, I will investigate what the correlation between variables should be to justify the MAR assumption under MNAR? In the example above, Philips actively calls medical centers to obtain extra knowledge about the missing patients. What kind of illness do they have? How long will they need to stay in the hospital? What medications do they use? The answers to these questions do not fill-in the gaps in the data, but they give the researchers extra information. This information can be used to improve the quality of the analysis. In missing data terms, you could say these extra data enhances the covariance structure of the total dataset, and as such makes the MAR assumption more reliable. Collins, Schafer, and Kam (2001) concluded something similar by showing that the inclusion of a variable that correlates either with the incomplete variable or with the missingness improves parameter estimation in the situation of MNAR. My research will build on this by showing the behaviour of the three missingness mechanisms for an estimate of the mean (Schouten & Vink, 2017). Furthermore, I am working on simulations showing the effect of data correlations and missingness mechanisms on other type of statistical estimates.

Based on the observed data alone, it is not possible to differentiate between MAR and MNAR. Molenberghs, Beunckens, Sotto, and Kenward (2008) have shown that "Every missingness not at random model has a missingness at random counterpart with equal fit." Therefore, it is common to perform some sort of sensitivity analysis to test the stability of statistical estimates (Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke, 2015). A relatively easy but reliable way to perform such an analysis is by adding a constant δ to the imputed values. Generally, several values of δ are tested and when the analysis outcomes are stable for different δ, we can be confident that the MAR assumption holds (Van Buuren, 2012). Shahab (2015) developed a method to estimate the amount of δ with a so-called random indicator. By iteratively determining this indicator, the imputation procedure can be adapted while running, and as such result in good imputations even under MNAR. Shahab's method is tested for one type of MNAR missingness and in the situation of normally distributed variables. Extension to other situations needs to be done, such as cases with the presence of categorical variables or with poisson and exponential distributions. In addition, it would be interesting to know whether it is possible to test H_0: $\hat{\delta} = 0$ against H_a: $\hat{\delta} \neq 0$. If so, we could have a beginning of what might be a test to discriminate between MAR and MNAR.

2.3 Standardization

The development and implementation of standardized data analysis methods is extremely attractive, especially in the data science framework. An example of this is the `.Pipeline` module from Scikit-learn in Python (Pedregosa et al., 2011). Here, the user specifies a list of possible data transformations and a second list with some machine learning methods. The `.Pipeline` module will then automatically try every combination of data transformation and analysis method, including the k-fold crossvalidation procedure and the desired evaluation metrics. The `.Preprocessing` module of Scikit-learn contains a function called `Imputer` which can be implemented in the same pipeline and automatically generates mean, median or most frequent imputations of the missing values. In `R`, comparable trends emerge such as the development of `tidyverse` (Wickham & RStudio, 2017). Additionally, Van der Loo (2017) recently published the `simputation` package which "offers a number of commonly used single imputation methods, each with a similar and hopefully simple interface."

The extent to which imputation strategies can be 'pipelined' needs examination. There may be circumstances where scientific research could benefit from machine learning techniques and pipline practices. For example, Statistics Netherlands recently started research to investigate whether a standardized use of random forests models would give valid statistical estimates (Park, Pannekoek, & Van der Loo, 2017). Park et al. (2017) describe that random forests do not require a specific modelling or variable selection. As such, this machine learning technique is useful in situations where an imputation model needs to be generalized to other datasets, or when new variables are added to an existing, already imputed, dataset. Shah et al. (2014) describe that random forests are especially useful when interactions between variables need to be taken into account. For which situations do these results from Park et al. (2017) and Shah et al. (2014) hold? What guidelines do we need to standardize missing value treatment? And what are the implications for software implementations?

Furthermore, I wish to develop a standardized approach to investigate missing data. In general, it is acknowledged that a proper handling of missing data problems requires careful thinking about: the missingness percentage, missing data patterns, underlying mechanisms, possible predictor matrices and very important, the research or analysis question at hand (Rubin, 1987; Van Buuren, 2012). An `R`-package such as `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) therefore includes functions as `md.pattern` to inspect the missing data patterns. Other packages are created especially to assist with a thorough exploration of the missing values (e.g. `narniar`: Tierney, 2017). However, few researchers know how to actually perform such an exploration. Moreover, it is often hard for researchers to relate the results of such an exploration to their choice of missing data method. Therefore, I will develop an `rmarkdown` and Jupyter notebook that processes an incomplete dataset and returns valuable overviews of the missing data and their meaning. The availability of such a standardized exploration method might help all sorts of researchers to take logical decisions. Remark that Python does not yet contain exploration, amputation or imputation funections, and all of these need to be developed and published as well.

2.4 Fusion

Data fusion is the process of combining two or more datasets before doing an analysis (Leulescu & Agafitei, 2013). The method has many applications, such as matching of non-overlapping surveys or matching of surveys to business or administrative data. Data scientists are quite used to combining multiple source data as part of the feature engineering process (Hastie et al., 2009; James et al., 2014). For instance, I am currently involved in a specific, and quite innovative, type of data fusion: statistical matching based on location measures. Front-runners Almere and Amsterdam are using mobile phone data, public transport data and open source data such as Google data to gain insights into the number of passengers at a certain time on a certain place in the city center. Hence, a proper matching of these datasets is an important part of the data analysis.

The development of evaluation measures of fusion procedures is an ongoing process (e.g. De Waal, 2015). Theoretically, the evaluation of a fusion procedure should be done based on four distributional aspects (i.e. the individual, marginal, conditional and joint distribution, Rässler, 2004), but the absence of a true dataset complicates such evaluation. Therefore, in practice, evaluation measures such as the hit rate are used (Rässler, 2004). However, there are circumstances where the hit rate turns out extremely well but the matching procedure is in fact not sufficient. Simulation studies are needed to more deeply assess the reliability of fusion procedures.

3 Organization

This research project is carried out by me, Rianne Schouten, as a PhD student at Utrecht University under supervision of dr. Gerko Vink and prof. Stef van Buuren. Funding for this project is provided by DPA Professionals, Business Unit IT. At DPA Professionals, I am part of the Data Science Excellence Program and learn how to set up a data science pipeline and use machine learning techniques. You could say I work at the intersection of the two worlds I described earlier. In general, the planning for my research project is as follows:

Year 0: Development of R-function `ampute`, write an article draft to present an explanation and extensive test of the method.

Year 1: Send `ampute` article for review, send a second article on missingness assumptions for review (pillar: robustness), start simulations to translate missing data methodology to data science use cases, start communicating these results in blogposts and write an article about this (pillar: accuracy).

Year 2: Send accuracy paper for review. Start research on the random indicator (RI) method (pillar: robustness), develop amputation and imputation Python functions and generate `rmarkdown` and Jupyter notebook documents (pillar: standardization). Write blogposts to proof the practical use of these documents and communicate the notebooks with businesses.

Year 3: Write article about RI method and send for review. Start and finish study on the evaluation of fusion techniques (pillar: fusion).

References

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy. In D. Banks, F. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering and data mining applications* (pp. 639 – 647). Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330 – 351.

De Waal, T. (2015). *Statistical matching: Experimental results and future research questions* (Tech. Rep.). Statistics Netherlands, Discussion paper.

Garcia-Laencina, P., Sancho-Gomez, J., & Figuiras-Vidal, A. (2010). Pattern classification with missing data: a review. *Neural Computations & Applications*, *19*, 263 – 282.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer, New York.

Horton, N., & Kleinman, K. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Journal of the American Statistical Association*, *61*(1), 79 – 90.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning.* Springer, New York.

Leulescu, A., & Agafitei, M. (2013). *Statistical matching: a model based approach for data integration* (Tech. Rep.). Eurostat Methodologies and Working Papers.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B*, *70*, 371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., & Verbeke, G. (2015). *Handbook of missing data methodology.* Chapman & Hall/CRC Press: Boca Raton.

Park, S., Pannekoek, P., & Van der Loo, M. (2017). Random forests for official statistics imputation: towards a more efficient methodology.
(Under review)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peeters, M., Zondervan-Zwijnenburg, M., Vink, G., & Van de Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, *12*, 377 – 394.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, *33*(2), 153 – 172.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581 – 590.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New York.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147 – 177.

Schouten, R., Lugtig, P., Brand, J., & Vink, G. (2017). Generate missing values with ampute [Computer software manual]. Retrieved from https://github.com/RianneSchouten/Amputation_with_Ampute/tree/master/Vignette

Schouten, R., & Vink, G. (2017). The dance of the mechanisms: how observed information influences the validity of missingness assumptions.
(Under review)

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missig data using mice: a caliber study. *American Journal of Epidemiology*, *179*, 764 – 774.

Shahab, J. (2015). Imputation under a nonignorable missingness mechanism. In *Dual imputation strategies for analyzing incomplete data* (pp. 47 – 62). (Dissertation)

Tierney, N. (2017). Getting started with naniar [Computer software manual]. Retrieved from https://github.com/njtierney/naniar/blob/master/vignettes/getting-started-w-naniar.Rmd

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC.

Van Buuren, S., & Groothuis-Oudshoorn, C. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3).

Van der Loo, M. (2017). Getting started with simputation [Computer software manual]. Retrieved from https://github.com/markvanderloo/simputation/blob/master/pkg/vignettes

Wickham, H., & RStudio. (2017). Tidyverse [Computer software manual]. Retrieved from https://www.tidyverse.org/