No solution? Clustering to Evaluate Multiple Imputation

Anthony S. Chapman, Dr Steven Turner, Dr Wei Pang, Dr Lorna Aucott

Abstract

Blah

1. Introduction

Introduce the paper not the problem.

2. Background

In this section we introduce some key concepts and why they are needed.

2.1. Missing Data

Like many other scientists, from all disciplines, we encountered datasets with missing values within them[14]. The amount of missingness on some of these datasets was as high as 75% of all records having one or more values missing. Before any analysis can be carried out, we need to decide on what to do with the datasets. The first options is to ignore any missing values and carry out the analysis on the subset consisting of all records with no missing values (complete case analysis). The second option is to replace (impute) any missing values with what is likely to be there correct values.

Looking through some systematic reviews of cohort studies [8, 10, 16] we were able to see how hundreds of different studies dealt with missing data. The majority (circa 70%) seems to carry out complete case analysis, and some (circa 25%) don't mention whether there were any missing values or what they did if there were. (Theory about most data possible for better analysis) In our case, this would mean only being able to use 30% of the data available. Although ignoring missing values seems to be the norm, the systematic reviews also show studies which use imputation methods before their analysis to deal with the missinng values. Imputation has been shown to be useful[9] but it should not be taken lightly, [8, 10] have shown 9% and 6% (respectively) of the studies they reviewed used imputation methods that are known to produce biased results.

2.2. Imputation

Once it was decided that we would rather impute datasets instead of conducting complete case analyses, we set about finding the best imputation techniques. We decided to use the statistical computing software R [12] for any analysis and after plenty of searching settled with a package called MICE [17]. Although it seems to be a good imputation package, how are we sure that this imputation methods (or any) will enhance our data for better analytical results. Even if this method has been proven to work on another daatset, it is no indication that it will work on any other.

maybe explain multiple imputation...

In order to check whether the imputation method works on a given dataset, one can first apply the method to a benchmark dataset and analyse the effects of such an action. For this process to imply the effects of the method on the given dataset, the benchmark must behave as close as possible to the dataset. Unfortunately, it is often very difficult and time consuming to find such a benchmark. It would be almost impossible to find a benchmark which could also be used for any given dataset. Thus, we felt that benchmark datasets should also be specific to the given dataset. Once could create a benchmark by analysing the way data is missing in the dataset and applying it to a dataset which is complete. The best complete dataset to mimic the original behaviour could be a complete cases dataset extracted from the original dataset.

2.3. Clustering

The next challenge will be to interpret the effect of imputing a dataset with any given imputation method. Measure theory and clustering provide us with good starting points on achieving this, measure theory related to systematic ways in which to assign numbers to suitable members of a group, this way one can give a sense of distance in a less conventional manner (for example, what is the distance between record 5 in a dataset and the average record in the same dataset, note the dataset could include non-numerical variable). Clustering is an unsupervised computation

method which takes a group of items and puts them into smaller groups according to their characteristics. The idea being, all items that behave similarly will be put into one group and all other items will be in other groups.

3. The Problems

When working with raw routinely acquired data, one of the first problems a researcher will have to overcome is how to deal with missing values [4]. With so much data being collected daily [14], it is no wonder that some of the data may become corrupt within the process, from gathering it to it being stored. These data corruptions may happen due to human error, computational inefficiency or other unforeseen circumstances [3]. Missing values in a dataset create an array of questions that need to be answered, namely, will it affect the results from any analysis, what can be done with records with missing values, if a solution to another dataset with missing values is found, will it work on another and lastly, if more than one solution is found, which one is the best for a specific dataset. We now discuss some of these problem and provide possible solutions.

3.1. Incompleteness / Missing Values

When analysing data at it's most raw form, the amount of missing values in datasets are at their highest. Although there are ways to combat missing data, such as mean-value imputation or multiple imputation [11, 13, 2], many researchers whom may not be computationally or statistically confident would rather ignore any records with missing values [1, 8, 10, 16]. As an example, in [1], the authors decided to use 2,758 records for analysis out of the possible 44,261 mainly due to missing data, this is a mere 6.2% out of the records available. There must be a way for even non-computing or non-statistical researchers to benefit from the tools available.

3.1.1. Possible Solution: Imputation. *Imputation is the process of replacing missing values with substituted values, one has to be careful when imputing data as there are many techniques (default value, mean value and multiple imputation just to name a few [5]) and using them without care will lead to erroneous data analysis [6]. By creating a user-friendly system with clear guidelines on how to impute data and some explanation on how it works, we believe that researchers whom would normally ignore data with missing values will be more likely to use more of their data through imputation, thus improving the quality and credibly of their analysis.*

3.2. Will it work on my data./Does it work for all datasets.

After deciding that imputation is beneficial to the study, the next step will be to find an imputation method for the dataset. Some good starting points may be Muliple Imputation by Chained Equations (MICE [17]) using the computational language R [12] or the Impute Missing Values function in the statistical software SPSS [7]. The problem is, how does one know if the imputed values are representative to the truth, how does one know whether record 2,754 column 5 is "42" or not after applying the imputation method.

Even if the imputation method has been proven to work on someone else's dataset such as [15], it is no indication it will work for any other dataset. This is due to the complex nature that missing data is created, for example there might be a relationship between one missing value and another one.

In order to test whether an imputation method works on any specific dataset, one needs something to compare the results to a benchmark. Then an analysis can be carried out to assess the effects of the imputation. Unfortunately, it is very difficult to find a complete dataset which contains the same characteristics as any another dataset, there will always be differences.

3.2.1. Possible Solution: Testing your own data. *The proposed solution is to create clone datasets by analysing the missingness characteristics from a given dataset and applying them to a new dataset created by only keeping the complete records. We can therefore create artificial mini datasets which behave in the same manner as the original dataset. In this manner we will be left with an original dataset, a subset consisting of only the complete records (the benchmark) and artificially incomplete datasets created from the complete records with the characteristics from the original.*

We then have a benchmark and a testing dataset that is as close to your original dataset as any dataset can be. The idea being that if the imputation method works on the testing datasets, then it will work on the original, and we can test whether imputation is successful by comparing to the benchmark.

3.3. Which imputation is best for me

The following problem applies to researchers, even those computationally competent, who wish to know whether one imputation method is better than another.

There is nothing to easily compare results from different imputation methods or same imputation methods with slightly different parameters. The main problem arises when one tries to compare the outcomes from one method to another, here an adequate analogy would be to compare imputation method A to method B would be like comparing chocolate with a bicycle; the outcomes might not be comparable.

There should be a program that can take different imputation methods and output something that can be compared to the output from another imputation methods. Having to do this is hard enough if you have a computing degree, so it is essential that something can be created that everyone can use.

3.3.1. Possible Solution: Comparing Imputation. *In order for a researcher to be able to compare different imputation techniques on their own datasets, the outcomes of the techniques need to "talk the same language". Creating a program that takes in a dataset and any imputation method and outputs a standardised efficiency classification we will be able to compare different imputation methods on the same dataset using this standardised efficiency classification. Thus every researcher will be able to compare different imputation methods without having to understand the individual imputation technique outputs.*

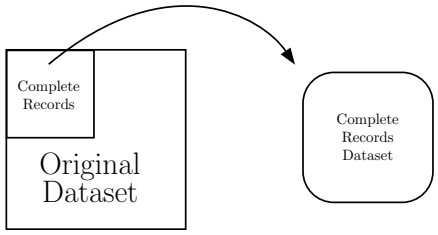
4. Proposed Framework

The idea: *The underlying concept is to first create a benchmark dataset from the complete records, then create artificially missing datasets, which represent the original dataset, out of the benchmark by mimicking the original dataset. We achieve this by analysing how data is missing and create testing datasets by imposing the same missingness into the benchmark. We would then impute the artificially missing datasets and analyse how far they have travelled from the benchmark. We can check how far imputation has taken the datasets from the benchmark by clustering the benchmark and the testing datasets. Like this we will be able to see the effects of any imputation technique on any dataset.*

Pre-requisite: *In order to evaluate an imputed dataset, one has to first have a dataset with missing values and an imputation method. At this stage, any imputation method can be used, this accommodates the ability to compare different imputation methods on the same dataset to evaluate the best one. Similarly, one can apply the same imputation methods to the dataset multiple times by changing the imputation parameters, thus finding the optimal imputation. We will caal the dataset O and imputation method $imp(x)$, were x is any incomplete dataset.*

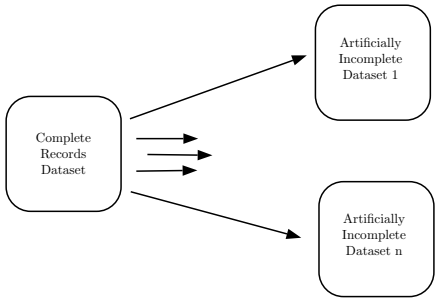
Stage 1: Extracting Benchmark *Firstly we need to create a benchmark dataset by extracting all the complete records from O, we call this benchmark dataset CC for Complete Cases. We then analysis O and find it's missingness characteristics, this will be used to create replicas later in the process. Notice that $CC \subset O$*

Figure 1. Stage 1



Stage 2: Create Dummy Datasets *Next, we create n artificially incomplete datasets, called artMiss.i where i is a number from 1 to n by applying the missingness characteristics from O to CC n times. It is important to apply the missingness in a manner that treats each artMiss.i separately, this way we can test the imputation more robustly. Thus we now have a benchmark dataset CC, and n artificial datasets with missing data which follow the same structure as the original dataset.*

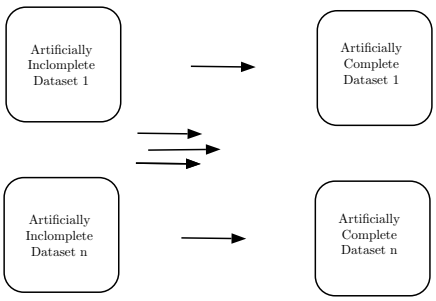
Figure 2. Stage 2



Stage 3: Impute Dummy Datasets *The next step will be to impute all artMiss.i using the chosen imputation method, it is very important to apply exactly the same procedure (same imputation with the same parameters)to all datasets in order to have reliable results, the imputation process must be deterministic. This will create n artificially complete (imputed) datasets, called artComp.i where i is a number from 1 to n.*

Stage 4: Cluster all Datasets *In order to evaluate the effect of an imputation method, we will use clustering techniques and measure theory to find the distance*

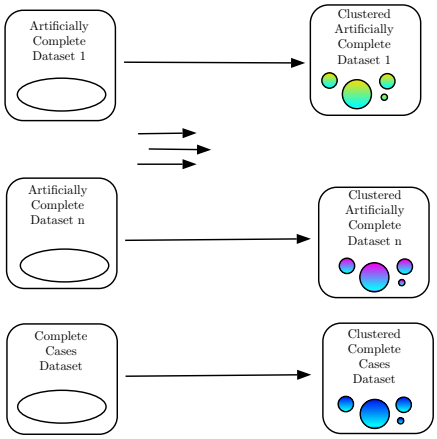
Figure 3. Stage 3



between all artComp.i and CC. By finding the distance between sets of cluster, we can see how close (or far) an imputation method has taken each artMiss.i from the true values (the benchmark CC). We will now be able to compare the effect of different imputation methods and evaluate each one.

Thus we need to cluster our benchmark dataset CC and all imputed datasets artComp.i, we will call clustCC the clustering of CC and clust.i will be all clustered artComp.i. Note that as with imputing all datasets we need to make sure the same clustering method with the same parameters in being used. This way we can accurately calculate the distances between the datasets. Similarly to the imputation process, for this to work, the clustering process must be deterministic.

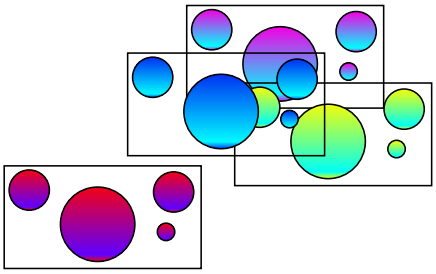
Figure 4. Stage 4



Stage 5: Distance Between Dummies and Benchmark *Finally, we will need to calculate the distance between all clust.i and clustCC. This will indicate the effect of imputing an incomplete dataset by finding the distance between the clustering of the imputed datasets (clust.i) and the clustering of our benchmark (clustCC). The system should output an efficiency indicator to show how far away all clust.i are from clustCC, this will be how we judge whether an imputation method gives correct values or not.*

Whether the final output describes a successful or efficient imputation method will be subjective. Each researcher will have to decide whether the imputed values are close enough to represent the truth or whether they are too far to provide fair results from any type of analysis.

Figure 5. Stage 5



5. Discussion

It would be beneficial for the output to have a scale in which the researcher will be able to judge the imputation methods (ie. 40% efficient of it travelled 15% from benchmark). These will be discussed at a later date

Ways to compute the distance between two clusterings (not individual clusters)

We need a reference point to judge whether the imputation is good or not, mean imputation as a reference point.

Distance measure - could also use modelling to create models of each dataset and compare the results

6. Conclusion

References

[1] Amy M. Branum, Jennifer D. Parker, Keim Sarah A., and Schempf Ashley H. Prepregnancy body mass index and gestational weight gain in relation to child body mass index among siblings. *American Journal of Epidemiology*, 174(10):1159–1165, 2011.

[2] Alan C. Cock. Working with missing values. *Journal of Marriage and Family*, 174(67):10121028, 2005.

[3] The Analysis Factor. Missing data mechanisms: A primer. 2013 (accessed: February 22, 2016). `http://www.theanalysisfactor.com/causes-of-missing-data/`.

[4] Ana S. Fernandes, Ian H. Jarman, Terence A. Etchells, José Manuel Fonseca, Elia Biganzoli, Chris Bajdik, and

Paulo J. G. Lisboa. Stratification methodologies for neural networks models of survival. In *Bio-Inspired Systems: Computational and Ambient Intelligence, 10th International Work-Conference on Artificial Neural Networks, IWANN 2009, Salamanca, Spain, June 10-12, 2009. Proceedings, Part I*, pages 989–996, 2009.

[5] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[6] Mustansar Ali Ghazanfar and Adam Prugel-Bennett. The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations. *Informatica*, 37(37):6192, (2013).

[7] SPSS Inc. *SPSS Statistics for Windows, Version 17.0.* Chicago: SPSS Inc, 2008. `http://www-01.ibm.com/software/uk/analytics/spss`.

[8] Amalia Karahalios, Laura Baglietto, John B Carlin, Dallas R English, and Julie A Simpson. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*, 2012.

[9] Amalia Karahalios, Laura Baglietto, Katherine J Lee, Dallas R English, John B Simpson, Carlin, and Julie A. The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study. *Emerging Themes in Epidemiology*, 2013.

[10] Katya L Masconi, Tandi E Matsha, Justin B Echouffo-Tcheugui, Rajiv T Erasmus, and Andre P Kengnecorresponding. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. *The EPMA Journal*, 2015.

[11] Therese D. Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):. 353–383, 2001.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0 `http://www.R-project.org`.

[13] Donald B. Rubin. An overview of multiple imputation.

[14] ScienceDaily. Big data, for better or worse. 2013 (accessed: January 18, 2016). `http://www.sciencedaily.com/releases/2013/05/130522085217.htm`.

[15] Anoop D. Shah and Jonathan W. Bartlett. Comparison of parametric and random forest mice in imputation of missing data in survival analysis. 2014.

[16] Leigh Tooth, Robert Ware, Chris Bain, David M. Purdie, and Annette Dobson. Quality of reporting of observational longitudinal research. *PRACTICE OF EPIDEMIOLOGY*, 2005.

[17] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. `http://www.jstatsoft.org/v45/i03/`.