No solution? Clustering to Evaluate Multiple Imputation

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

Abstract

Blah

1. Introduction

Data collection has been increasing Missing data is inevitable (human and computing reasons, i.e. people not putting it in or computer corrupting it,) Non-computing people either imputate willy-neely or ignore missing data - need to use as much data as you can. (Ask Graham about theory about using as much data as possible for better analysis) Many imputation algorithms out there with many parameters, which is best? Need

2. Background

Talk about something [1] [9]

3. The problems

One of the most common set backs when analysing routinely collected data is the sheer amount of missing values. This problems leads to a domino effect of further problems, namely what can you do with data with missing values, if a solution to another dataset with missing values is found, will it work on your own and lastly, if more than one solution is found, which one is the best for a specific dataset.

3.1. Incompleteness

With so much new data being collected daily [9], it is inevitable that data contains large amounts of missing values [6], whether they be through human error or computational inefficiency. Although there are ways to combat missing data, such as mean-value imputation or multiple imputation [6, 8, 2], many researchers whom are not very computationally or statistically confident would rather ignore any records with missing values [1, **?***,* **?***,* **?***,* **?***]. 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 will create values where there were non before, 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 [3]) as using them without care will lead to erroneous data [4]. By creating a user-friendly program with clear guidelines on how to use it 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*

3.2. Will it work on my data.

This next problem arises when a researchers does decide to use the records with missing values but does not have sufficient knowledge to apply the available methods efficiently, methods such as Muliple Imputation by Chained Equations (MICE [11]) using the computational language R [7] or the Impute Missing Values function in the statistical software SPSS [5]. 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 male or not after you apply the imputation method.

Even if the imputation method has been proven to work on someone else's dataset such as [10], there is no indication it will work for yours. This is due to the many reasons and ways 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 your dataset, one needs something to compare the results to, a benchmark, like this one would be able to analyse what effect of the methods. Unfortunately, it is very difficult to find a complete dataset which contains the same characteristics of your own dataset, there will always be differences.

3.2.1. Possible Solution: Testing your own data. *In order to test how well an imputation technique, such* *as MICE, one needs to be able to compare the effects of the imputation method to a benchmark, we propose creating a benchmark from the users own dataset. By analysing the missing data characteristics, extracting the subset of complete data from the dataset and then replicating the same dataset onto the subset, we are able to create artificial mini datasets which behave as the original one except now we have a benchmark to compare the effects of imputation.*

3.3. Which imputation is best for me

The following problem applies to researchers, even those computationally competent, who wish to find out 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 that compare imputation method A to method B would be like comparing chocolate with a bicycle; the outcomes might not be comparable.

There should a way to compare different methods without having to create your own computer software in the process. Although

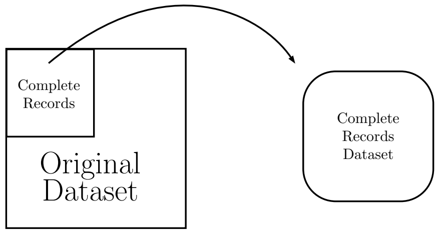
3.3.1. Possible Solution: Comparing Imputation. *In order for a researcher to be able to compare different imputation techniques on their own datasets, the outcomes of the techniques need to "talk the same language". By having a program that takes a dataset and imputation pair and then outputs the efficiency of the imputation, one is able to compare these outputs with ease and without having to understand the individual imputation technique outputs.*

4. Proposed Framework

The idea: *The underlying concept is to create a benchmark out of all the records with no missing values and then create replicas out of the benchmark to mimic the original dataset by analysing how data is missing and create testing datasets by imposing the same missingness into the benchmark. We would then impute the testing 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.*

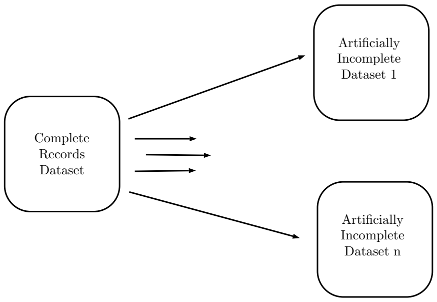
Stage 1: *Firstly we need to create a benchmark dataset by extracting all the complete records, we call this dataset, CC for Complete Cases. We then analysis the original dataset and find the the characteristics of missing data.*

Figure 1. Stage 1



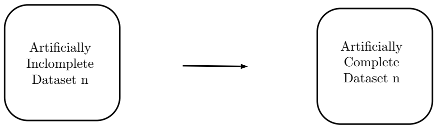
Stage 2: *We then create n testing datasets by applying the same amounts of missingness from the original dataset to CC. Thus we are left with a benchmark dataset, CC, and n artificial datasets with missing data which follow the same structure as the original dataset, call these artMiss.i where i is a number from 1 to n.*

Figure 2. Stage 2



Stage 3: *The next step will be to impute all artMissi's using the imputation method of choice, it is important to apply exactly the same procedure to all datasets in order to have reliable results. This will create n artificially complete datasets, called artComp.i where i is a number from 1 to n.*

Figure 3. Stage 3

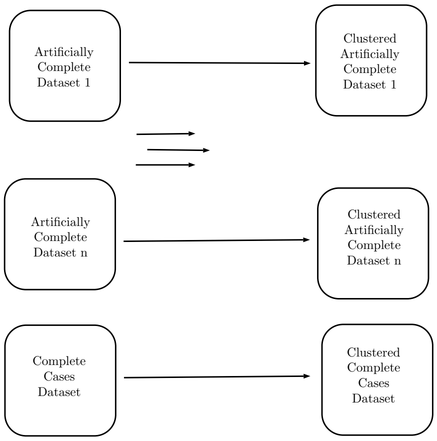


Stage 4: *In order to compare how far the imputation method has taken the original dataset, we will evaluate how far imputation has taken the dataset from the truth (truth being CC, our benchmark) using clustering methods. Thus we need to cluster our benchmark CC*

and all artificially complete datasets artComp.i ($\forall i \in [1,n]$)

we will combine all artComp.i into one dataset by averaging the information. All imputed variables will create one average variable and all the results that were there originally will be the same. By doing this will be able to comfortably compare this master artificially complete dataset with the benchmark CC.

Figure 4. Stage 4


Artificially Complete Dataset 1
Clustered Artificially Complete Dataset 1
Artificially Complete Dataset n
Clustered Artificially Complete Dataset n
Complete Cases Dataset
Clustered Complete Cases Dataset

Stage 5: *Once we have averaged the imputed datasets, we will evaluate how far imputation has taken the dataset from the truth (truth being CC, our benchmark) by clustering both datasets using the same clustering algorithm. We will be able to compare the cluster results (such as cluster centres, cluster widths, dissimilarities etc..), we will then see how artComp has moved from CC.*

Figure 5. Stage 5

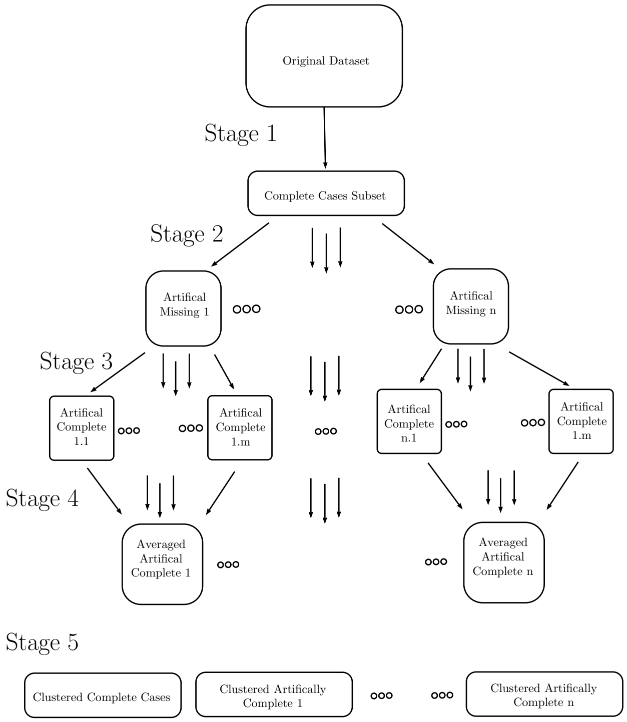
5. Discussion

6. Conclusion

7. The Framework

The process is as follows: Using a dataset with missing values, call this "OD", for original dataset, and an imputation technique, call this "Imp", you first analyse the missingness characteristics of "OD" to in order to apply them later. Then, create a new dataset by deleting any record from "OD", call this "CC", for complete cases. Using the missingness characteristics, we create more da

Figure 6. Framework flowchart



The proposed framework is as follows: In order to assess the effects of any imputation technique, the program will need a dataset with missing values, called "D", and an imputation method, called "I". Then the function "I(x)" is a function that takes a data with missing values and outputs an imputed dataset.

Specify dataset (O) and imputation method (I) Analyse dataset to obtain the missing characteristics Create subset of dataset with only complete cases (C) Apply missing characteristics to create n individual datasets out of C, called ($ArtM_1$,$ArtM_2$...$ArtM_n$) Impute all Art_n to created artificially complete datasets $ArtC_n$ Apply clustering algorithm to C and all $ArtC_n$ to create ClustC and ClustArt_n Average clustering outcomes from all ClusArt_n and compare to ClustC Analyse how far the average of all ClusArt_n have gone from ClustC Normalise the distance to give a percentage of goodness for the user.

8. Conclusion

It's better to use all the data you can but can't blindly imputation. This framework indicates whether your data

9. Discussion

Working on implementing this, ClEMI, any researcher regardless the computing ability will be able to use it.

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

[4] 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).

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

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

[7] 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`.

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

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

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

[11] 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/`.