No solution? Evaluating Multiple Imputation

Anthony S. Chapman

January 15, 2016

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

2.1 Incompleteness Issue

Huge amout of missing data, too many people not using everything they can (i.e. ignoring data with missingness)

Cite medical papers that only use a small part of the data they have due to missingness

2.2 Will it work on my data?

Even if imputation (like MICE) works on someone's data, will it work with mine? Different types and levels of missing data. Need a bench-mark but other datasets no matter how close to yours are still not yours.

Cite a couple of paper that evaluate imputation on their dataset but do not generalise on any dataset

2.3 Which imputation is best for me

There is nothing to easily compare different imputation techniques, post researches haven't got computing background... (need a way to formally say that statement, maybe number of disciplines VS computing??)

You can't compare an evaluation of one imputation on data A and a different imputation on data B, is chocolate better than bacon?

3 Possible Solutions

3.1 Incompleteness

Imputations solves incompleteness, you just have to be careful how you use it.

Can't blindly impute something as it might result in bias and unreliable results.

3.2 Testing your own data

Use your own level or missingness as a benchmark and create mini-me's as benchmark. You are the closest thing to yourself. Group theory stuff, multidimensional-mixed data distance measurements, Gower, medoids, widths and dissimilarities.

Just because it worked on someone else, doesn't mean it works for you, cite papers who test specific datasets.

3.3 Comparing Imputations

Will now be able to compare different imputations on your own dataset with "normalised " results for comparison.

4 Conclusion

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

5 Discussion

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