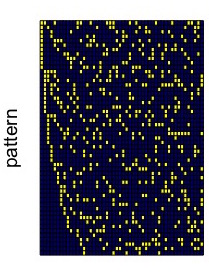
Report about 5 different imputation packages

1. Define missing value

To simplified the testing procedure and save time, I took “NeuroIm1\_Final\_AD\_vs\_NC\_training” as a dataset. The probability of missing data for each subject is 0.2, so each subject has 11 missing data.

The plot could show the current situation for the different covariate. Each column represents one subject. Yellow cells mean the missing value.

This plot reflects the missing value for subject 1:50.

From this plot, we could see the combination of missing value.

1. Test imputation packages

Since the time spent in finishing the imputation process for all 230 subjects is incredibly long for some package (like MICE), to compare the precise of different package, I tested the package depending on the first 50 columns, which include missing value.

In the meantime, for some reason, every time I try to run the Amelia, the R studio will warn that it encountered a fatal error and the session was terminated. I still trying to find the problem of this, hence, in this document, I discussed only four packages, which were MICE, MISSFOREST, HMISC and MI.

The comparison main focus on following several points -- the imputation methods, the accuracy, the time it needed and some problems may meet when using different packages.

1. The imputation methods.

In this session, I want to discuss the imputation method for different packages, since I think the imputation method may reflect the imputation accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| MICE | MissForest | Hmisc | MI |
| pmm | Random forest | pmm | ppd/pmm |

For MICE and Hmisc, the imputation method is pmm. For MI the imputation method could change from ppd to pmm.

Among all these four methods, MissForest uses random forest to impute data.

1. The accuracy

(Notes: when testing MI through imputing the first 50 columns, it always had about 2 problematic variables and then stopped the loop automatically.

However, when I changed the # of subjects to 20, this problem did not happen again. I think this may because when have too many variables, the probability of diverging in a chain is increasing, but I was not sure whether this is the real reason. I am still trying to understand the background algorithm for this method to figure out this problem. )

|  |  |  |  |
| --- | --- | --- | --- |
| MICE(MSE) | MissForest(MSE) | Hmisc(R^2) | MI(summary statistics) |
| e-00~e-02 | e-01~e-02 | 1 |  |

Based on the test result, I found that the MSE value for MICE is from e-00 to e-02. For example, among all 50 subjects, there always exist one or two subjects, whose MSE are larger than 1 but smaller than 10, although most of the MSE value are not that large. This means that this method’s accuracy is not stable.

As for MissForest, comparing with MICE, the MSE value for this method is more stable. However, there also have one or two MSE is large than 0.5. I think we could use some methods to combine different imputation results to optimize the final result.

Hmisc has the best performance, since r square is equal to 1 for all subjects.

1. The time needed

Because we are going to deal with the big data set, the imputation time has to be taken into consideration as well.

|  |  |  |  |
| --- | --- | --- | --- |
| MICE | MissForest | Hmisc | MI |
| middle | short | long |  |

MICE took quite a long time to impute all 230 subjects, which was more than one hour, not mention Hmisc.

However, when applying MissForest to impute missing values, the time needed is more acceptable, which was about 55 seconds.

1. Some problems may meet

**MICE:**

For mice, I think the main problem is the time it needed. For example, when I tried to use this method impute all missing data, the iteration process cost incredible long time, more than one hour. Considering that I just selected a subset of the whole dataset and the probability of missing was not quite high, if we apply this method to impute a missing value for big data, the time cost will be unacceptable.

In the meantime, the accuracy of this method is not stable as well. In one imputation set, the MSE for some subjects could reach 3, even 5.

**MissForest:**

Comparing to MICE, this method has better performance, when imputing the missing value for all 230 subjects. The spending time is more suitable.

However, compare with MICE, although the MSE for this method is much better, among all MSE values for all 230 subjects, sometimes, 1 or 2 MSE will larger than 50%, which is quite well. (Depending on different Xtemp\_miss sets.)

But I think we could try to figure out this problem by combining different imputation data set to seek smaller MSE.

**Hmisc:**

The main problem for Hmisc, the same as MICE, is the time it needed. Since we are going to deal with real big data, the spending time is really an important problem.

However, compare to another method, the accurate for this method is quite well. The R-square could reach 1.

**MI:**

Although I still have no idea about why this method could not impute a large number of subjects, which I am doing my utmost to figure out. I think this will be an essential problem, considering that the dataset may have hundreds of subjects.