Report2 TFoLDS

Szymon Czop

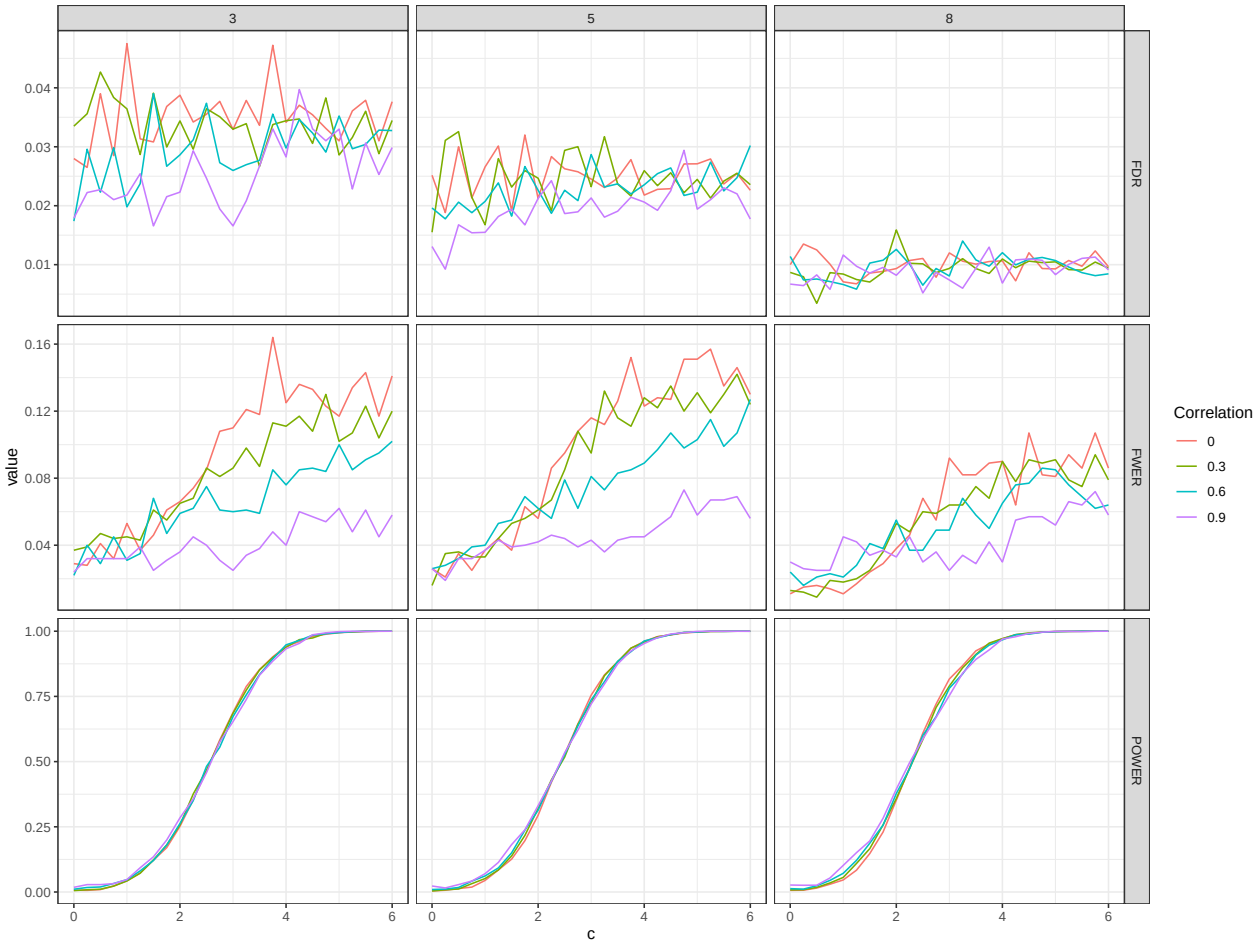
25 01 2020

In the second report, we will take on on the workshop Benjamini-Hochberg procedure. We will have a look at its behavior when it comes to false discovery proportion, FWER. and power. All these statistics will be tested for vectors of length 10. In our simulations number of non-null components will change from 3,5 to finally 8. We will check above mentioned statistic for different values of these non-zero entrance (from 0 to 6) and different correlations between random variables which our vector is made off. All this will be shown in the plot and described below it.

Note:

I combined Exercise 1 and Exercise 2 because they are the same when it comes to the experiment, the only difference is the correlation of variables in vector. All data for Exercise 1 is also included in the plot.

PLOT



Short description of the plot legend:

Value on the left side, are just values that given statistic takes.

C are values that are taken for non-null entries in the vector.

Numbers 3,5,8 at the top are the number of non-null entries for the given experiment

Colors of the lines in each chunk indicate the value of each statistic for vector with one of the correlations described in the legend.

Description of the experiment:

For each possible combination of parameters, we create a set of 1000 vectors. Then for each of the vector from this dataset, we use B-H procedure. According to its outcome, we take the expected value (average) of our scores for each statistc, calculated according to discoveries made by B-H.

POWER:

There is almost no difference when it comes to power when we try different correlations for variables in our vector. The line is almost always the same in each case. The only small difference is seen when the number of non-null (we will use n-n as a shortcut) is 8 and the value of c is between 0 and 2. For 3 and 6 n-n shapes of both curves are the same. When we have 8 n-n our power curve is growing more rapidly than the previous two and for c = 2 we have power at 50% level, when at the same time, for the rest two is about 25%.

Conlcusion : There is almost no difference what correlation we have between the variables in the vector. In the same time, if we have more n-n entries, power of the test is increasing faster

FWER:

For 3 and 5 n-n entries, FWER is very similar. The higher the correlation the lower the probability that we will have, false-positive discoveries. For the first two plots FWER has the biggest values when there is no correlation between variables and in it peek has about 16%. In the same time, when variables are strongly correlated maximum probability of type I error is belowe 8%. The smallest FWER is seen when we have 8 n-n entries. All the curves are more compact and have smaller values. Although the difference that is dependent on the correlation in the vector is still visible

Conclusion: The bigger correlation and number of n-n entires the smaller FWER

FDR:

The level of the FDR can be theoretically calculated as : $\frac{|H_0|}{n} * \alpha$ where $|H_0|$ is number of true zero values in vector , n in our case is equal to the length of the vector (for us 10) and α is significance level (0.05). Using this formula we get theoretical values for 3, 5 and 8 n-n entries come as follows: 0.035, 0.025, 0.01. So we should see the difference in plots for ich n-n entry. This is happening and it's visible that for more n-n the average of the FDR is decreasing. When we go from right to left we see that lines are making something like stairs and go lower for more n-n. Exactly as predicted in the theoretical foundations. As previously in FWER size of the correlation has a positive effect on this statistic. The bigger it is the smaller the FDR.

Conclusion: FDR is sensitive to the number of true zero values and will decrease when there are more n-n entries than another way. Again correlation matters but not as much as previous for FWER.

Made by Szymon Czop 292913