AMSI
VACATION
RESEARCH
SCHOLARSHIPS
2017-2018

Confidence Control for False Discovery Proportion

Yilun He
Supervised by Uri Keich
The University of Sydney

Vacation Research Scholarships are funded jointly by the Department of Education and Training and the Australian Mathematical Sciences Institute.

Australian Government
Department of Education and Training

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Abstract

Modern inference on large number of hypotheses usually focuses on the expectation of false discovery proportion, namely false discovery rate. However, the variability of such proportion is seldom mentioned in practice, resulting in potential instability. I implemented three confidence control methods for false discovery proportion, and applied Monte-Carlo simulation to assess their effectiveness.

1 introduction

In a statistical inference, the *significant threshold* usually determines *type-I error rate*, which is the probability that a true null hypothesis is rejected. We compare p-value obtained against thresholds (typically 10%, 5% or 1%) to determine if we reject the null hypothesis.

Usually rejecting a null hypothesis means we have one positive discovery. We do not only want to reduce the occurrence of false discovery, but also need to increase the probability that a non-null hypothesis is successfully rejected. Hence we have *type-II error rate*, the probability that an alternative hypothesis is **not** rejected. 1− type-II error rate is also called *Power*.

Type-I error rate indicates how likely you will find false result, and type-II error rate indicates how likely you will not find a true result. A good statistical test must guarantee low error rate on both measure. A balance between risk(type-I error rate) and return(Power) must be maintained.

When multiple hypotheses are tested, either p-value or significant threshold must be adjusted, as individual type-I error rate build up over large number very quickly.

In a multiple hypotheses testing scenario, we usually obtain a set of p-values from the hypotheses. Denote this set by A. N, the p-values from null hypothesis, is a subset of A. Let the size of A be m. Normally a rejection set R_t is determined based on a rejection threshold t. That is, reject all p-values that are less than t.

Family-wise error rate (FWER) is generalised version of type-I error rate. It indicates the probability that any single false discovery arises from a number of hypotheses. That is,

$$P_{\text{FWER}} = P(\frac{\#(R \cap N)}{\#R_t} > 0) \tag{1}$$

FWER is traditionally used to determine the significant threshold in such scenario. However when the number of hypothesis is very large, controlling FWER at certain level is not practical and will sometimes lead to empty rejection set. One simple example is Bonferroni correction. It divides the FWER threshold by number of hypotheses. The rejection threshold approaches 0 as the number of hypotheses goes up.

The concept of *false discovery proportion* (FDP) comes up as an alternative for FWER. It represents the composition of null hypotheses in the rejection set. Thus, FWER actually describes the probability that FDP is greater than 0.

$$\text{FDP} = \frac{\#(R \cap N)}{\#R_t} \tag{2}$$

FDP is a random variable dependent on the null and alternative distribution, number of hypotheses m, proportion of null hypotheses π_0 and rejection threshold.

Benjamini and Hochberg [1] proposed the use of false discovery rate(FDR), the expectation of FDP, and gained massive popularity in applications. Instead of trying to guarantee that FDP is zero at certain confidence level, they claim that controlling FDR at certain level is also meaningful. Using FDR to replace FWER provides a less conservative control on risk, and massively increases power.

The idea of false discovery rate massively increased the power of testing. The first FDR control algorithm proposed by BH, assumes the worst case scenario, in which all hypotheses are true nulls. This assumption leads to a very conservative rejection set when the proportion of null hypotheses, pi_0, is low. Storey[2] improved this method by estimating the expected value of null hypothesis proportion.

Those methods only controls the mean value of false discovery proportions, and they do not consider the possible variability. In fact, false discovery proportion can be highly unstable.

Many papers on false discovery research were published. Various methods are designed for different dependency assumption and confidence requirement.

However, it is very important to alert the society that FDR control might be unstable. Despite many methods that provides safer decision option are discussed. They are neither popular nor implemented.

In next section, I will introduce some FDR controlling algorithms, and demonstrate the variability of FDP when we use these algorithms by simulation.

2 False discovery rate controlling

Benjamini and Hochberg published their FDR control method together with the idea of FDR itself. The algorithm is extremely simple. He estimates the FDR when a rejection threshold is set to be $P_{(k)}$, the k-th order statistic of p-values.

$$\widehat{\mathrm{FDR}}(P_{(k)}) = \frac{mP_{(k)}}{k} \tag{3}$$

And find the largest k such that the estimated FDR is below some FDR threshold c.

This method is rather conservative amongst existing FDR control algorithms, since it does not estimate pi_0, the proportion of null hypotheses and assumes worst case value, 1. Hence the observed FDP will be much lower than the target when π_0 is low.

Storey proposed his estimator for null hypotheses proportion, and improved the algorithm such that the power will be higher even when π_0 is low. (see figure 1).

$$\widehat{\mathrm{FDR}}(P_{(k)}) = \frac{\widehat{\pi_0}mP_{(k)}}{k} \tag{4}$$

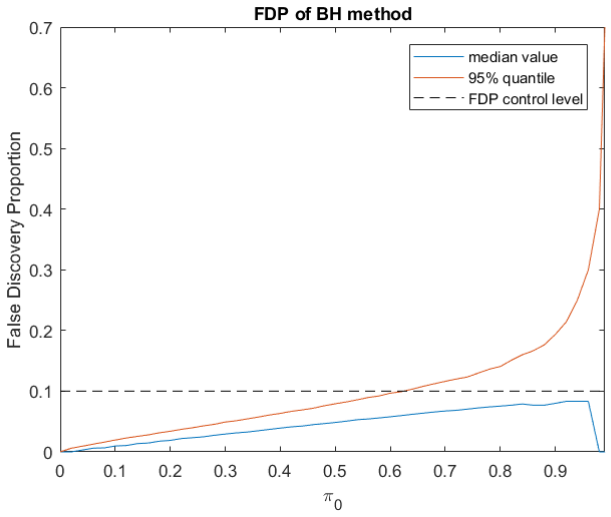

Figure 1: The empirical FDP quantiles by simulation under BH algorithm, when m = 400

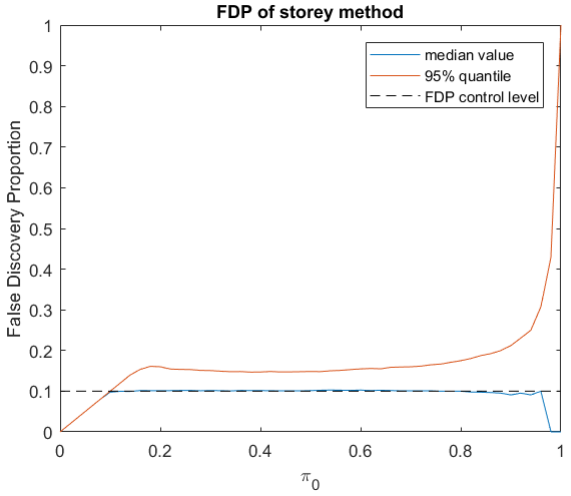

Figure 2: The empirical FDP quantiles by simulation under storey algorithm, when m = 400

This is a less conservative method. It is consistent with the expectation of FDP. However, the FDP can be much higher than this expectation, especially when m is low and pi_0 is high (See figure 2).

At each π_0 value, I repeated the experiment 10000 times, and calculated median and 95% quantile of false discovery proportion. Both BH and storey algorithm provides good control of false discovery rate, but the observed value of FDP might be much higher. The 95% quantile even goes to 1, which implies that FDR control is potentially unreliable.

3 Confidence control for false discovery proportion

I will introduce *binomial heuristic*,a simple heuristic that uses binomial distribution in this section. Strictly speaking I do not have the right to name it, since I did not invent it. I will use this name in italic to represent this method in following sections.

Controlling the value of FDP at certain confidence level has been discussed by several paper. The following approach has been mentioned by Ge and Li(2012) [3], but not tested or implemented. Consider n independent and identically distributed $Unif(0,1)$ random variable $X_1, X_2...X_n$:

$$\begin{aligned} I(X_i < t) &\sim Bernoulli(t), t \in [0,1], i = 1, ..., n \\ \sum_{i=1}^{n} I(X_i < t) &\sim Binomial(n,t) \end{aligned} \tag{5}$$

Hence, fix a rejection threshold and given π_0, the null hypothesis proportion, we can find a confidence upper bound for the number of null hypothesis landing in this rejection region using the quantile function of binomial distribution.

$$\begin{aligned} P(\text{FDP} > \frac{Q_{m\pi_0,t}(1-\alpha)}{\#R_t}) &= \alpha \\ U_{\alpha,t,\pi_0} &= \frac{Q_{m\pi_0,t}(1-\alpha)}{\#R_t} \end{aligned} \tag{6}$$

Where $Q_{n,t}$ is the quantile function of Binomial(n,t), and m is the total number of hypotheses.

However, in practice we rarely know π_0, and have to estimate it. The dependence between these two estimating steps causes potential risk of this approach, although it performs very well in practice. A confidence upper bound for π_0 can also be constructed based on the binomial distribution:

$$\begin{aligned} I(X_i > \lambda) &\sim Bernoulli(1-\lambda), t \in [0,1], i = 1, ..., n \\ \sum_{i=1}^{n} I(X_i > \lambda) &\sim Binomial(n, 1-\lambda) \end{aligned} \tag{7}$$

By assuming that all observations above a certain value λ are generated by null hypotheses, we can provide a confidence upper bound for pi_0.

$$P(m - \#R_\lambda < Q_{m\widehat{\pi_0},1-\lambda}(\alpha)) = \alpha \tag{8}$$

Where $Q_{n,t}$ is still the quantile function of Binomial, and $\#R_\lambda$ represents the rejection set when rejection threshold is λ. The selection of λ may vary, and in my experiment I used 0.5 as a default value. When all other parameters are determined, we can solve for $\widehat{\pi_0}$ and the solution will be our confidence upper bound for π_0.

If the assumption that all observed values above λ are from null hypotheses fails, the estimate will be more conservative and guarantee that the confidence level is still achieved.

Using confidence level $\frac{\alpha}{2}$ in both steps, we obtain a α-confidence upper bound for the false discovery proportion.

Ultimately, we use:

$$t_u = \max\{t|\widehat{\text{FDP}}(t) < c\} \tag{9}$$

4 Simulation results

I tested a vast range of parameter combination, and I will choose a few typical ones for demonstration in this section.

4.1 setup

The parameters used are,

$$\begin{aligned} \text{Number of experiment} \quad & n = 10000 \\ \text{Number of hypotheses} \quad & m = 10000 \\ \text{Null distribution} \quad & \text{Uniform}(0,1) \\ \text{Alternative distribution} \quad & \text{Beta}(a, 1/a) \\ \text{Binomial heuristic lambda} \quad & \lambda = 0.5 \\ \text{Confidence level for FDP} \quad & \alpha = 0.05 \\ \text{FDP control target} \quad & c = 0.05 \end{aligned} \tag{10}$$

The value of α and c determines how strict FDP is controlled. In my future research I will provide more variety of parameter combination, but for demonstration I only explain this set of parameter in detail.

I used all 5 method to control false discovery proportion at 0.05, and for confidence control methods we are expecting at most α probability that the observed FDP is higher than α. For FDR control method(BH and storey), they only want the expectation(Actually similar to median in practice) to be below $c = 0.05$.

We use 5 different false discovery control algorithms, they are:

1. Benjamini-Hochberg method[1]
2. Storey method[2]
3. Binomial heuristic
4. Genovese-Wesserman confidence envelope[4]
5. Ge-Li uniform confidence upper bound, algorithm 1[3].

First three methods are already introduced in previous sections. Genovese-Wesserman(2004) uses a uniform distribution test to determine possible null hypotheses, and Ge-Li(2012) uses Higher Criticism statistic. Please read at the original paper if you are interested in their algorithm.

I implemented these five methods in Matlab, and improved Ge-Li method by incorporating Li-Sigmund Higher Criticism approximation[5]. In close future I will translate these code into R and hopefully publish a package featuring confidence control of FDP.

In each iteration, I randomly generate $m = 10000$ p-values from two different distributions and mix them together. The proportion of null hypothesis, π_0, takes a fix value for $n = 10000$ runs. Repeat such test for 50 different π_0 range from 0 to 1.

For each run I use all 5 methods to obtain rejection thresholds for the 10000 p-values observed. Then, since we know whether a p-value is generated from alternative or null distribution, we can calculate the actual FDP in each run. Repeating this process $n = 10000$ times then we can have a empirical distribution for actual FDP under each of the 5 methods. Record the $1 - \alpha$ quantile then we can see how good these methods control FDP.

In addition, we can plot the proportion of true alternative hypotheses that are rejected by algorithms. This measure can represent the power of each algorithm in practice. Ideally we want as high power as possible while the false discovery proportion is controlled as expected. Benjamini-Hochberg and Storey methods are added for comparison.

4.2 Alternative shape Beta(0.2,5)

Let us start with a good alternative distribution. The shape of Beta(0.2,5) is as follows:

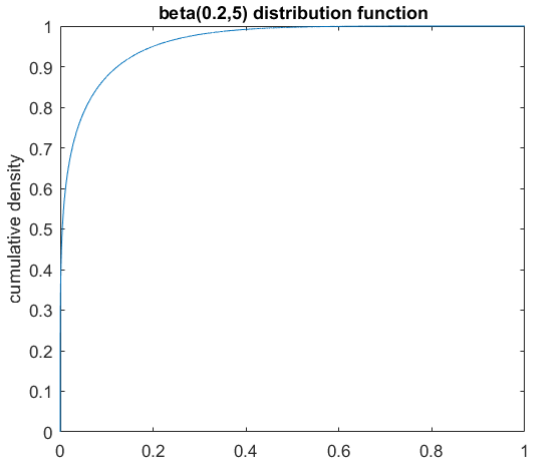

Figure 3: The cumulative distribution function of beta(0.2,5)

It is a relatively good alternative. Probability density function is not used because they are too left skewed and take extremely high density near 0. Samples from this distribution is very likely to take values near 0.

All three confidence control methods achieves the FDP control goal for all π_0 methods. The validity of BH and Storey has been verified many times in other publications, and I will skip them in my demonstration.

Since all three methods achieves the FDP control target, we can compare their power. Clearly storey has the highest power of all 5 methods for any π_0 value, as it is least conservative. But BH method has rather low

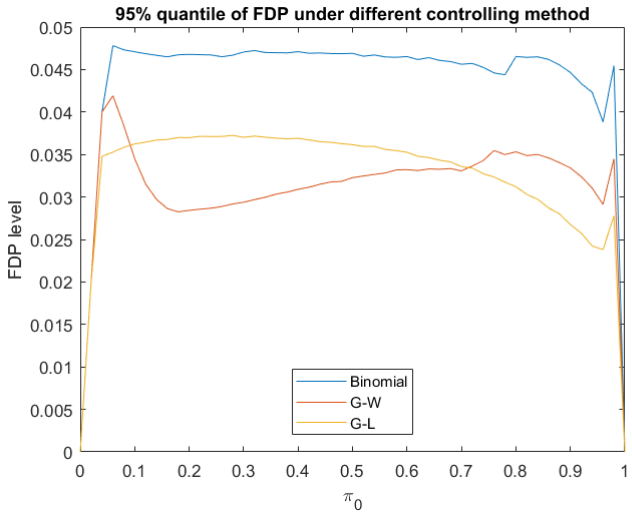

Figure 4: The empirical FDP 95% quantiles by 5 different algorithms, when m = 10000 and other parameter as declared above.

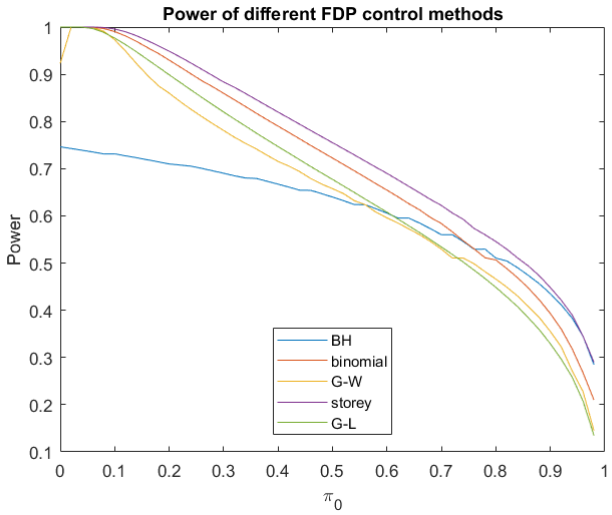

Figure 5: The median power of 5 different algorithms, when m = 10000 and other parameter as declared above.

power at low π_0 values.

Genovese&Wesserman method(G-W) and Ge&Li method has similar performance, and it is hard to compare which one is better in this scenario.

Binomial heuristic is the highest performing FDP confidence control method, and its power is very close to storey. It outperforms BH method when π_0 is less than about 0.8.

4.3 Alternative shape Beta(0.05,20)

Then an extreme alternative method is used. The observations are very likely to be very close to 0.

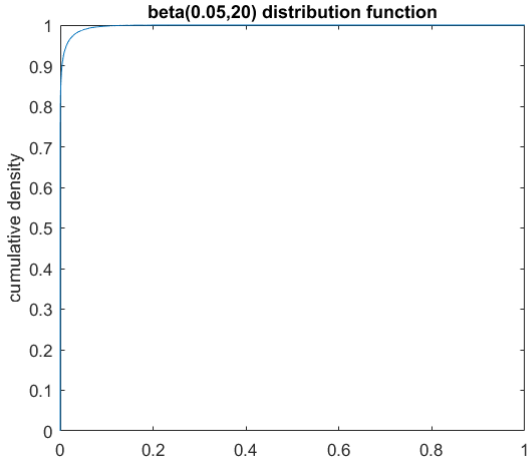

Figure 6: The cumulative distribution function of beta(0.05,20)

In this scenario, we expect higher power for all algorithms, since it is easier to distinguish alternatives from null.

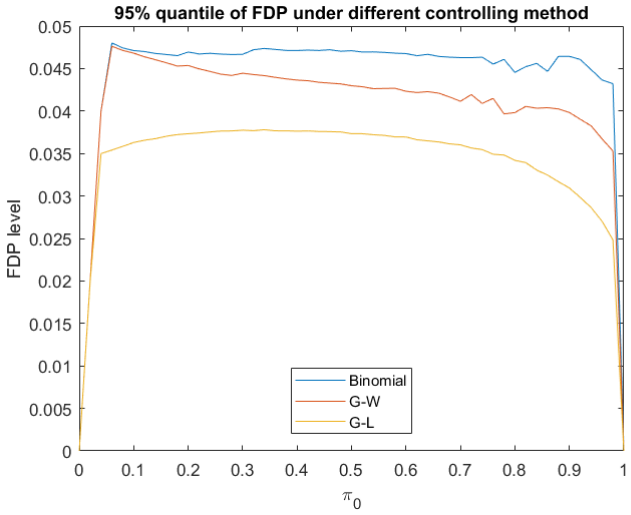

Figure 7: The empirical FDP 95% quantiles by 5 different algorithms, when m = 10000 and other parameter as declared above.

All three confidence control methods still meets the FDP control goal.

It is important to understand that, at extreme π_0 values, the FDP will be lower. When π_0 is close to 0, the total proportion of null hypothesis is below the false discovery target, and whole population will be rejected. When π_0 is very high, it is difficult to guarantee a low false discovery proportion since the population is almost filled by null hypotheses. Hence the algorithm will generate a empty rejection set, which has 0 false discovery.

Storey is also the most powerful method. BH has bad performance at low π_0 values.

Genovese&Wesserman method(G-W) is better than Ge&Li method for all π_0 in this case.

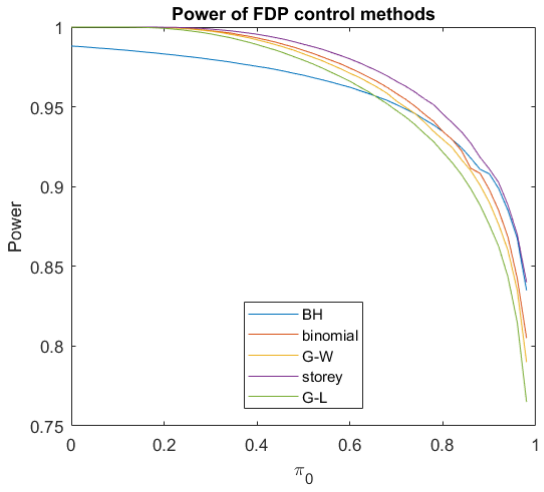

Figure 8: The median power of 5 different algorithms, when m = 10000 and other parameter as declared above.

Binomial heuristic is still the highest performing FDP confidence control method, and its power is very close to storey.

It is also notable that, all methods are more powerful than the first case.

4.4 Alternative shape Beta(0.35,3)

The last alternative distribution is very weak, and it is hard to distinguish it from null distribution.

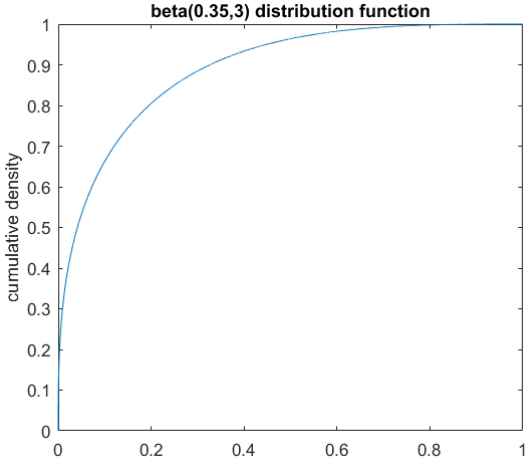

Figure 9: The cumulative distribution function of beta(0.35,3)

In this scenario, we expect lower power and more unstable outcome.

All three confidence control methods still meet the FDP control goal.

Similar statement can be made for BH, storey and binomial heuristic. They outperform other two methods by a large difference.

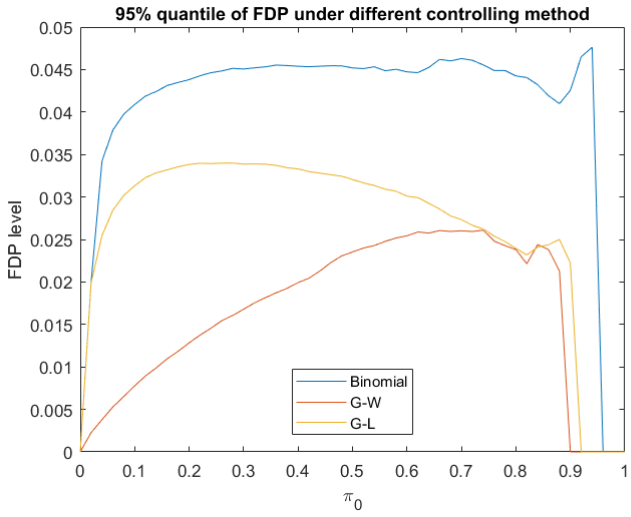

Figure 10: The empirical FDP 95% quantiles by 5 different algorithms, when m = 10000 and other parameter as declared above.

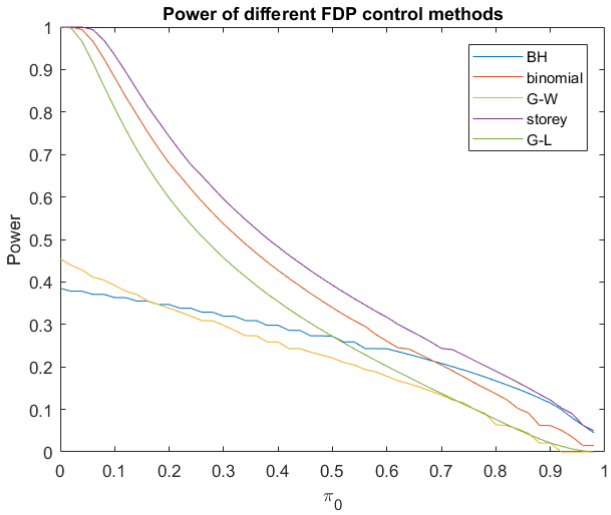

Figure 11: The median power of 5 different algorithms, when m = 10000 and other parameter as declared above.

Genovese&Wesserman method(G-W) is much worse than Ge&Li method for all π_0 in terms of power. We may conclude that, Ge-Li method is better when alternatives are weak and G-W method is only suitable for stronger alternatives.

Binomial heuristic is still the highest performing FDP confidence control method, and its power is very close to storey.

Almost all algorithm obtain a power close to 0 when π_0 proportion is too high, because the alternative is very similar to null hypothesis.

5 discussion

There are a few important findings in my investigation and simulation.

The variability of false discovery proportion decreases as number of hypotheses goes up, but the convergence speed can be slow considering a practical scale. FDP is more volatile when pi_0 is high.

Benjamini-Hochberg method, although controls false discovery rate only (Expectation control), is very conservative when pi_0 is low. Potentially this will lose power of the test.

Storey's method achieves highest power but potentially the false discovery rate can be far from the control target. This issue should be addressed in downstream analysis.

Binomial heuristic is high-performing and achieves very good power while confidence control requirements are met. I am still attempting to complete a formal proof for this heuristic, but before that exhaustive simulation can grant the heuristic much reliability.

Ge and Li's method is more stable than Genovese and Wesserman's when the alternative is weaker. The latter achieves better power when the alternative is strong.

6 Future work

R translation of my Matlab code will be completed soon. I hope it will become a popular package used in modern biology, chemistry and medical research.

More simulation scenarios shall be tested against the validity of binomial heuristic. If the method actually works, we are building a very good FDP confidence control method.

In addition, all above simulation and developments are based on the fact that all hypotheses are mutual independent. It will be necessary to discuss the cases when dependency between hypotheses exist.

7 Acknowledgement

Special thanks to Fan Wang, who was cooperating with me in the summer of 2016-2017 on this topic.

References

[1] Benjamini, Y., Hochberg, Y., 1995, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological) Vol. 57, No. 1 (1995), pp. 289-300

[2] Storey, J., Taylor, J., Siegmund, D., 2004, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, J. R. Statist. Soc. B (2004) 66, Part 1, pp. 187205

[3] Ge, Y., Li, X., 2012, Control of the False Discovery Proportion for Independently Tested Null Hypotheses, Journal of Probability and Statistics Volume 2012, Article ID 320425

[4] Genovese, C., Wasserman, L., 2006, Exceedance Control of the False Discovery Proportion, Journal of the American Statistical Association, Vol. 101, No. 476 (Dec., 2006), pp. 1408-1417

[5] Li, J., Siegmund, D., 2015, HIGHER CRITICISM: p-VALUES AND CRITICISM, The Annals of Statistics 2015, Vol. 43, No. 3, 13231350