Report on Independent Hypothesis Weighting

Xihan Qian

2024-03-01

1. Introduction

Numerous techniques have been devised for the analysis of high-throughput data to accurately quantify biological features, including genes and proteins. To ensure the reliability of discoveries, the false discovery rate (FDR) has become the dominant approach for setting thresholds. The methods for controlling FDR primarily rely on p-values, among which the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg 1995) and Storey's q-value(Storey 2002) are popular.

However, Ignatiadis, Klaus, Zaugg, and Huber suggest that FDR methods based solely on p-values exhibit suboptimal power when the individual tests vary in their statistical properties. (Ignatiadis et al. 2016). When these methods focus exclusively on p-values, they overlook potentially relevant covariates. For example, in RNA-seq differential expression analysis, one such covariate could be the normalized mean counts of genes. Intuitively, genes with higher counts are likely to have greater power in detection compared to those with lower counts, and it would be optimal to include this relationship in the analysis.

To address this limitation, a novel approach known as independent hypothesis weighting (IHW) has been introduced. IHW improves the power of multiple hypothesis testing while controlling the FDR rate. The key innovation of IHW is that it recognizes that not all tests have the same power or the same prior probability of being true. It allows for the assignment of different weights to different hypotheses based on covariates that are predictive of the test's power or its probability of being a true discovery. These covariates can be anything from the biological characteristics of the genes being tested to the technical aspects of the measurements. More explorations will be done in the following sections on this method.

2. Methodology and Simulation

There have been existing attempts to increase power by using covariates and this sections starts off by introducing some known methods.

2.1 Weighted and Group weighted BH procedure

The weighted BH method has m hypotheses $H_1, H_2, ..., H_m$ and m weights $w_1, w_2, ..., w_m \geq 0$ that satisfy $\frac{1}{m}\sum_{i=1}^{m} w_i = 1$. After the weights are obtained, the BH procedure is applied on the modified p-values: $\frac{p_i}{w_i}$. In this scenario, selecting the weights before observing the p-values is essential, relying on prior knowledge or information indicating the likelihood of some hypotheses being true over others. This requirement presents a significant challenge, one that IHW seeks to address, and it utilizes the grouped weighted BH procedure (GBH). In this method, there are G groups where each X_i takes the same value within each group. GBH first estimates the proportion of null hypotheses by $\hat{\pi}_0(g)$, then weights the hypotheses proportionally to $\frac{1-\hat{\pi}_0(g)}{\hat{\pi}_0(g)}$, and finally apply the BH procedure. However, the asymptotic theories in this method doesn't function well when the number of hypotheses $\frac{m}{G}$ is finite (Ignatiadis and Huber 2021). To resolve this issue, cross-weighting is used, where the idea is analogous to cross-fitting in regression settings, and this gives rise to the naive version of IHW.

2.2 Naive IHW and two-groups model

This version, also abbreviated as IHW-GBH since it is based off GBH, first divides the hypothesis tests into G groups based on the values of covariate $X = (X_1, X_2, \dots, X_m)$, with m_g number of hypotheses in the g-th group. It is also assumed that we have access to the p-values $P = (P_1, P_2, \dots, P_m)$, which is independent of X under the null. With this setup, $\sum_{g=1}^{G} m_g = m$. Then the weighted BH procedure is applied with each possible weight vector $w = (w_1, w_2, \dots, w_G)$, while the optimal w^* is the vector that lead to the most rejections. This method extends the BH procedure, making it pertinent to discuss the associated maximization problem. This problem stems from the two groups model (Efron 2008), which is a Bayesian framework that explains the BH procedure. Formally, assume that H_i takes values 0 or 1, and $\pi_0 = \mathbb{P}(H_i = 0)$. The distributions are as follows:

$$\begin{aligned} H_i &\sim \text{ Bernoulli } (1 - \pi_0) \\ P_i \mid H_i = 0 &\sim U[0, 1] \\ P_i \mid H_i = 1 &\sim F_1 \end{aligned}$$

The marginal distribution for p-value P_i is then

$$P_i \sim F(t) = \pi_0 t + (1 - \pi_0) F_1(t)$$

With this, the Bayesian FDR becomes:

$$\text{Fdr}(t) = \mathbb{P}\left[H_i = 0 \mid P_i \leq t\right] = \frac{\pi_0 t}{F(t)}$$

A natural empirical estimator for the CDF would be the ECDF, and it can be written in terms of $R(t)$, which denotes the total number of rejections:

$$R(t) = m\widehat{F}(t) = \sum_{i=1}^{m} \mathbf{1}_{\{P_i \leq t\}}$$

Hence if $\widehat{\pi_0}$ is an estimator of π_0,

$$\widehat{\text{Fdr}}(t) = \frac{\widehat{\pi_0} t}{\widehat{F}(t)} = \frac{\widehat{\pi_0} m t}{R(t)}$$

If a conservative estimate is made: $\widehat{\pi_0} = 1$, then

$$\widehat{\text{Fdr}}(t) = \frac{mt}{R(t)} \tag{1}$$

With this, the optimization problem is:

$$\text{maximize } R(t), \text{ s.t. } \widehat{\text{Fdr}}(t) \leq \alpha, t \in [0, 1]$$

The corresponding estimator in IHW-GBH is of the form

$$\widehat{\mathrm{Fdr}}(t, \mathbf{w}) = \frac{mt}{R(t, \mathbf{w})} = \frac{\sum_{g=1}^{G} m_g w_g t}{R(t, \mathbf{w})}$$

where $R(t, \mathbf{w}) = \sum_{i=1}^{m} \mathbf{1}_{\{P_i \leq w_g t\}}$ is the number of rejections in bin g. Now the optimization problem is:

$$\text{maximize}\, R(t, \mathbf{w}), \text{ s.t. } \widehat{\mathrm{Fdr}}(t, \mathbf{w}) \leq \alpha$$

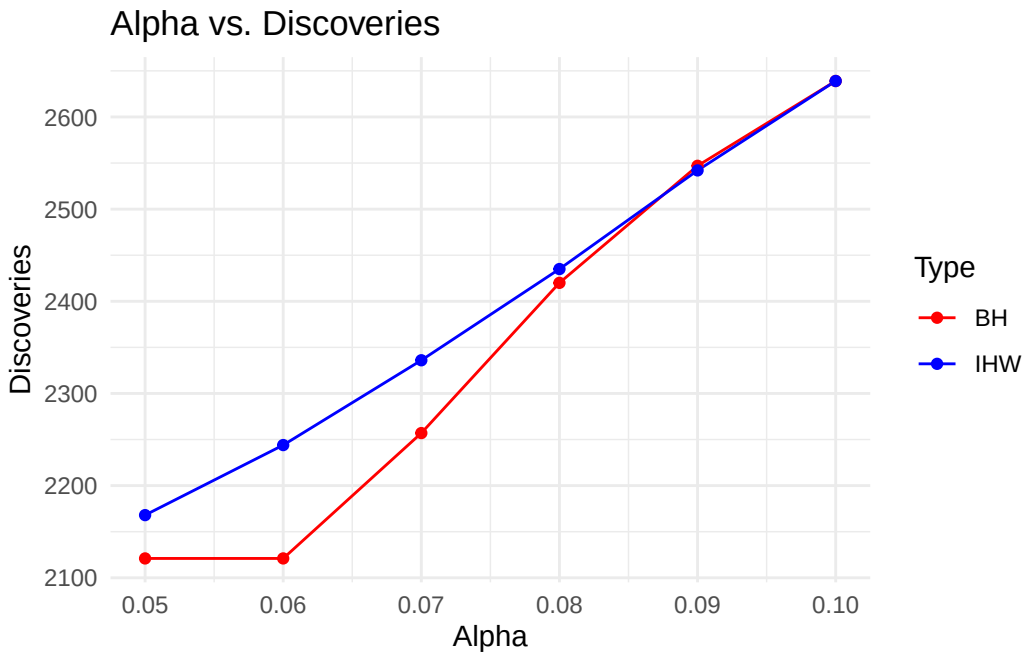
However, with this approach, there are also some disadvantages including potential loss of Type I error, complications in solving the maximization problem, and its inability to scale when large number of tests are present. That is why modifications are made, leading to the IHW method.

2.3 IHW

In the first modification, the ECDF $\hat{F}_g$ of the p-values in group g is replaced by the least concave majorant version called the Grenander estimator $\tilde{F}_g$. With this the maximization problem can be efficiently solved. The second modification involves randomly splitting the hypotheses into K folds, where K is usually taken to be 5. Then the maximization problem with the previous modification is applied to the remaining folds, leading to a weight $\widetilde{w} = (\widetilde{w}_1, \ldots, \widetilde{w}_G)$. The independence criterion between the hypotheses would guarantee the p-value P_i to be independent of the assigned weight w_i when the null hypothesis is true. The third modification ensures that the weights learned with $K - 1$ folds can be generalized to the held-out fold by adding a regularization parameter λ. The specific constraints are customized to whether the covariates are ordered or not.

2.4 Simulation

Wanting to see how this packages work exactly, I used one of the datasets suggested by the paper using RNA-seq data with read counts for genes as a covariate (Bottomly et al. 2011). It is worth noting that in the original dataset provided in the paper by Bottomly, the column of read counts is only binary, providing information of whether the counts are considered low or not based on a threshold after a log transformation. Without knowing the specific transformation applied, it is hard to achieve the original column, hence I self-generated a column of count reads using the negative binomial distribution. After trying some combinations, I chose the dispersion parameter to be 0.2 while the mean read counts is 1,000. The plot generated is as follows. There is a discrepancy between this plot and that generated in the paper possibly because the difference in the read counts, however, it can still be observed that the number of discoveries of IHW is no fewer than that of BH in most cases.



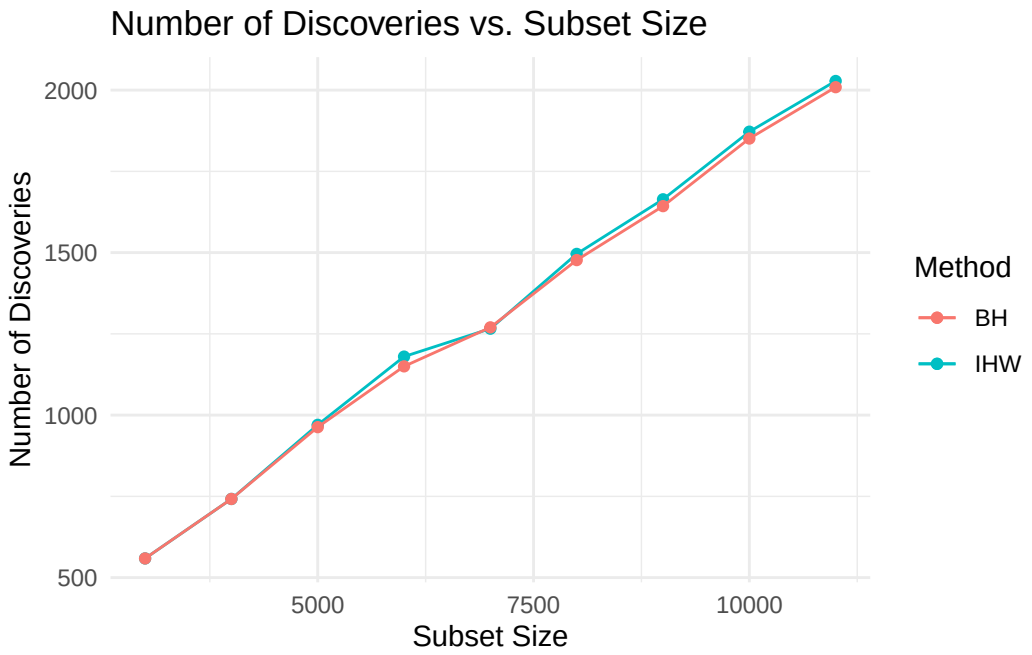
2.5 Discussions

One challenge that this method faces is that it largely depends on selecting an appropriate covariate that influences the power of each hypothesis test but is independent of the p-values under the null hypothesis. Identifying such a covariate is challenging and crucial; an unsuitable choice can diminish IHW's benefits or even degrade performance compared to traditional correction methods. This is a reason of why the plot above doesn't show great improvements. Another possible limitation is when there are datasets with high heterogeneity, where the covariate's relationship to test power varies significantly. In such cases, IHW's ability to accurately weight hypotheses could be compromised, potentially leading to less effective multiple testing correction.

3. Extensions

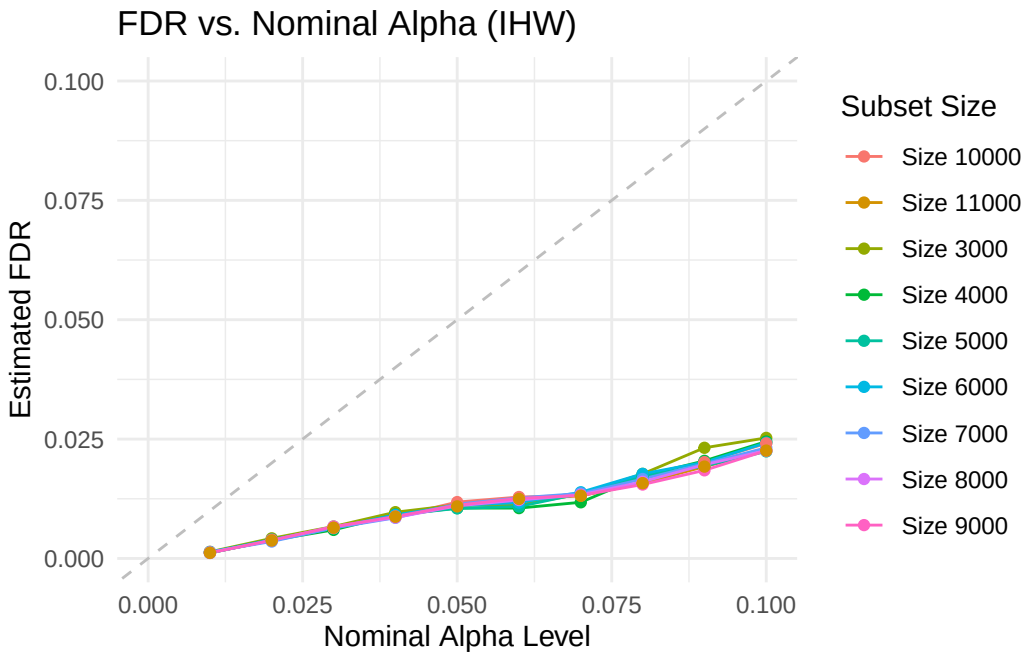
3.1 IHW performance through dataset

In statistical analysis, especially with multiple hypothesis testing, model performance is influenced by the number of tests. The variations observed across different numbers of tests can shed light on the efficiency and reliability of the model. In this section, these variations are looked into, while seeking to understand their underlying causes and proposing potential improvements. The number of tests in the analysis directly correlates with the number of genes under examination, as each gene represents a unique hypothesis to be tested. Given this relationship, analyzing subsets of the gene dataset becomes a viable strategy to examine the impact of test quantity on model performance. The RNA-seq dataset is used again here, with different sizes of subsets taken to firstly explore how the number of discoveries fluctuates. The plot is shown as follows:



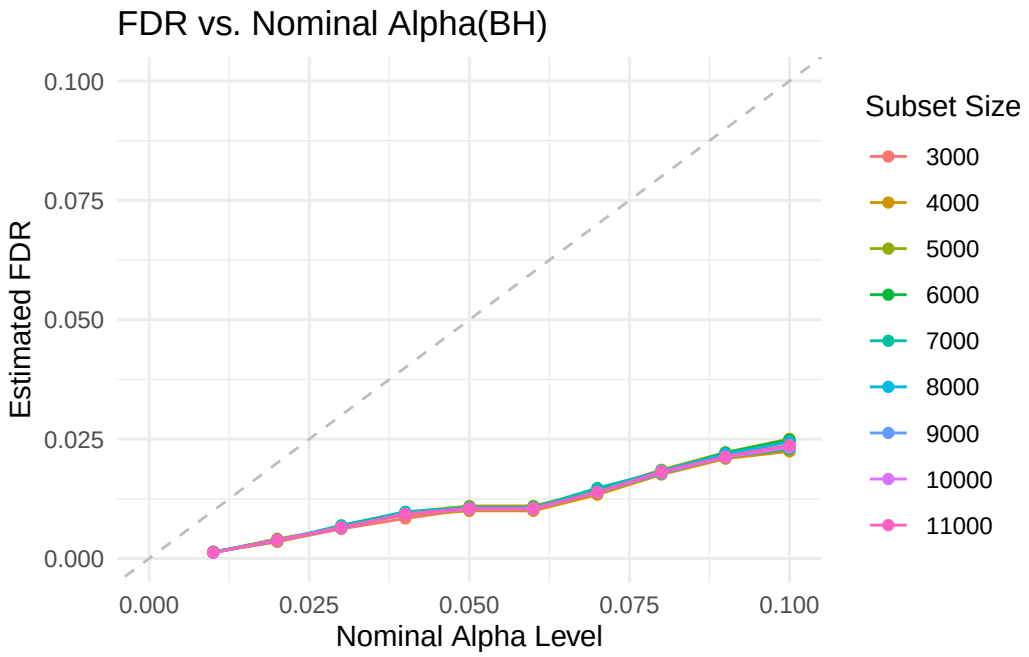
The nominal α level is set to be 0.01 here. I have tried various values ranging from 0.01 to 0.10 while the plot doesn't differ much. First off, it can be observed that as subset size grows, there are more discoveries for both IHW and BH. This can be due to several reasons. As the size grows, the statistical power is increased, leading to more rejections. It is also possible that larger subsets may enable the detection of smaller effect sizes that would be indiscernible in smaller subsets due to insufficient power. The ability to identify these smaller effects contributes to the overall increase in the number of discoveries as subset size grows. However, the number of discoveries for BH and IHW seem similar in general, with that of IHW to be higher at some points. This doesn't seem desired, but I wanted to explore and interpret through more plots.

I thought it would also be intereting to generate a plot of nominal α and FDR for comparisons. I applied it usin the IHW method first, and obtained the following:



Initially, I found it quite surprising that the curves were well below the $y = x$ line, suggesting that IHW is more conservative than expected. This observation was unexpected, especially since the plots in the paper indicated that for IHW, the curves should be closer to the diagonal line. The main reason why I think this is happening is because of the fact that I generated the covariate column myself through negative binomial distributions. In the context of IHW, the covariate should be informative about the power of each test, while randomly generating covariates might violate this condition. As for the effect of the subset sizes, it seems that the curves don't differ much as the size grows. One likely explanation is that the true positive rate could be homogeneous across the subsets. In this case the FDR control mechanisms will behave similarly across different sizes. This means that the proportion of true positives to false positives remains constant, and thus the estimated FDR will be similar. Another reason could be that all subsets are large enough to have adequate power. Then the increase in subset size may not lead to a proportionate increase in the number of rejections because the power of the test is already high in the smaller subsets.

It then occurred to me then that it could be beneficial to also have a similar plot but with the BH procedure. Hence I also produced the following plot. Since there are multiple sizes I tried, I placed the two methods in two separate plots for comparisons.

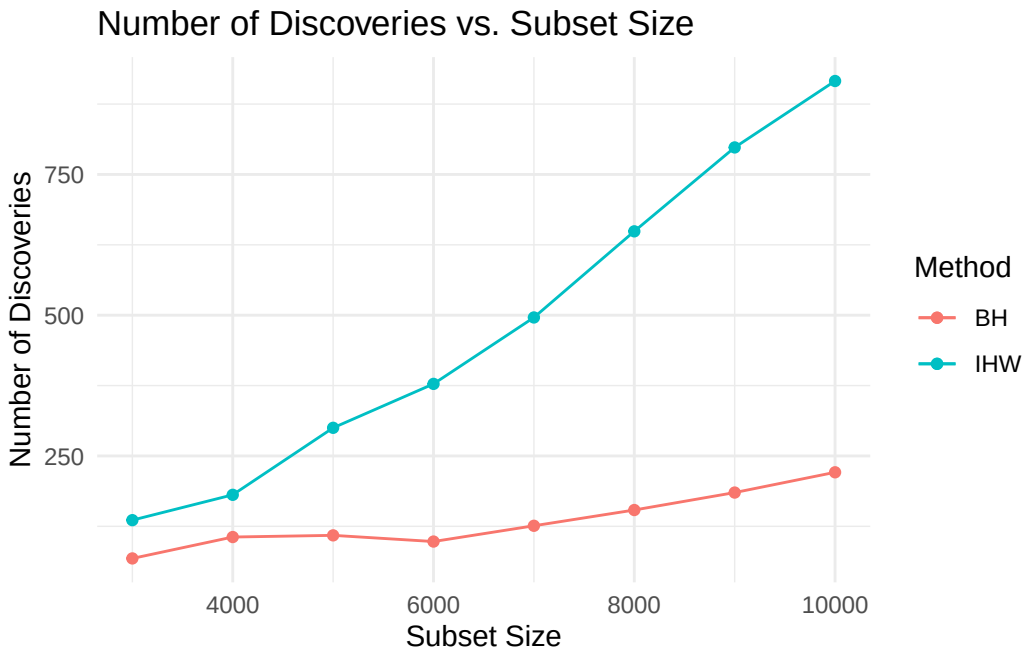


This plot with nominal α and estimated FDR using the BH procedure looks similar to that with IHW, which also matches what the paper suggested with one of the datasets in figure 2(f) (Ignatiadis et al. 2016).

Although the results are not desired since the number of discoveries for IHW doesn't exceed BH by much, indicating that there's not an improvement by using IHW, I wanted to keep these plots and analyses because they are a great indication that this happens when the covariate is not informative. This is further explored through simulations of my own, while intentionally enforcing the covariates to be informative.

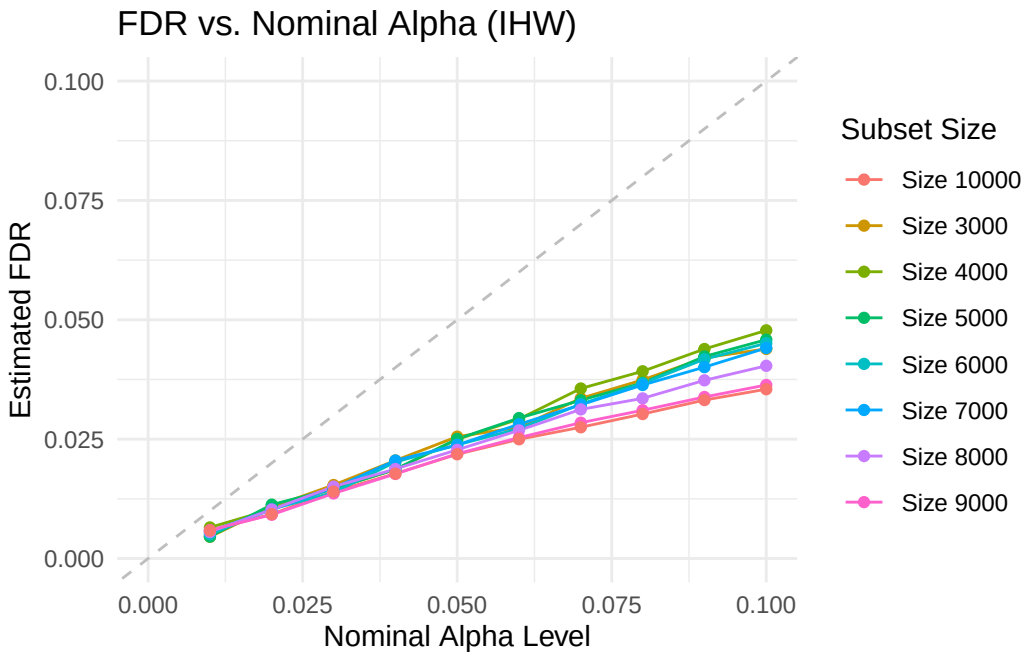
3.2 IHW performance through simulations

In this section I simulated p-values and covariates myself, specifically requiring both of them to relate to the effect size, which was generated through a normal distribution. The following procedures are similar to the previous subsection, however, the results are drastically different. The number of discoveries is shown as:

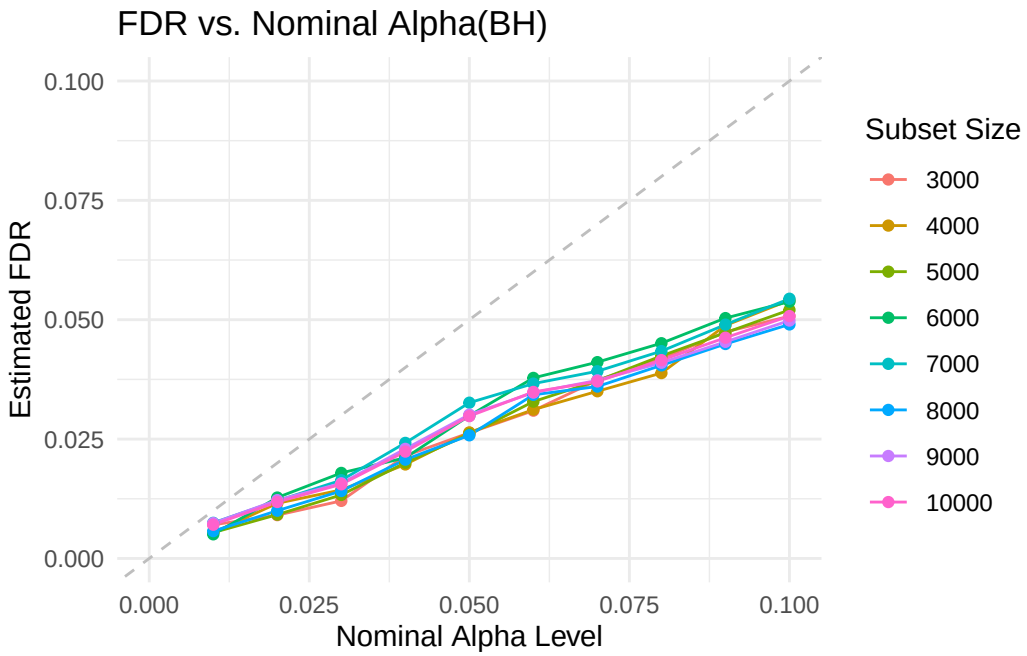


The nominal α is chosen to be 0.05 here. As it can be seen, the number of discoveries for IHW grows more than BH as the subset size increases. The discrepancy between the discoveries is now due to the fact that the covariate is informative.

Again, plots of nominal α and estimated FDR are plotted for IHW:



This is still similar to that by using the BH procedure:



However, comparing with before when the RNA-seq dataset was used and the covariate was not informative, the curves are more closely aligned with the diagonal line. Hence with better chosen covariates, these methods can be found less conservative.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Bottomly, Daniel, Nicole AR Walter, Jessica Ezzell Hunter, Priscila Darakjian, Sunita Kawane, Kari J Buck, Robert P Searles, Michael Mooney, Shannon K McWeeney, and Robert Hitzemann. 2011. "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays." *PloS One* 6 (3): e17820.

Efron, Bradley. 2008. "Microarrays, Empirical Bayes and the Two-Groups Model."

Ignatiadis, Nikolaos, and Wolfgang Huber. 2021. "Covariate Powered Cross-Weighted Multiple Testing." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83 (4): 720–51.

Ignatiadis, Nikolaos, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. 2016. "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing." *Nature Methods* 13 (7): 577–80.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64 (3): 479–98.