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. In these cases, we reject hypothesis i if the p-value is no more than some threshold $\hat{t}$.

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. **?**. 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.

If an additional covariate $X_1, X_2, \dots, X_m$ is included for each of the m tests, a popular method involves first filtering out some hypotheses for which $X_i < x$, based on a predetermined x, and then applying the BH procedure. However, Bourgon, Gentleman, and Huber (Bourgon, Gentleman, and Huber 2010) point out that this approach could result in some loss of Type I error control and requires the covariate to be independent of the p-values. This paper proposed a new method called independent hypothesis weighting (IHW), which generates weights based on data and is built upon Grouped Benjamini-Hochberg (GBH). More explorations will be done in the following sections on this method.

2. Theories

2.1 Weighted and Group weighted BH procedure

A generalization of the independent filtering method is the weighted BH method with m hypotheses $H_1, H_2, \dots, H_m$ and m weights $w_1, w_2, \dots, w_m \geq 0$ 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: $\frac{p_i}{w_i}$. Assuming that we only want $\tilde{m}$ hypotheses to be retained among all m, we would test the remaining p-values based on $\alpha\frac{i}{\tilde{m}}, i = 1, \dots, \widetilde{m}$. This is equivalent to assigning $\frac{m}{\widetilde{m}}$ as weights to the retained p-values and 0 for others and then applying the BH procedure. However, in this case the weights have to be determined prior to seeing the p-values, while this could be challenging. This is exactly what IHW is trying to investigate, and it is closely related to 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 groups based on the values of covariate $X = (X_1, X_2, \dots, X_m)$ and assume we also have access to the p-values $P = (P_1, P_2, \dots, P_m)$. Similar to GBH, the hypotheses are divided into G groups, with m_g number of hypotheses in the g-th group. 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. The maximization problem related to this method is a variation derived from the two groups model (**efron2008microarrays?**), 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.

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.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent Filtering Increases Detection Power for High-Throughput Experiments." *Proceedings of the National Academy of Sciences* 107 (21): 9546–51.

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.

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.