Report on Independent Hypothesis Weighting

Xihan Qian

2024-03-01

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. (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. 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.

A generalization of the independent filtering method is the weighted BH method, where m weights $w_1, w_2, \dots, w_m \geq 0$ are introduced and $\frac{1}{m}\sum_{i=1}^{m} w_i = 1$. Then the BH procedure is applied on the modified p-values: $\frac{p_i}{w_i}$. Assuming that we only want $\widetilde{m}$ hypotheses to be retained among all m, we would test the remaining p-values based on $\alpha\frac{i}{\widetilde{m}}, i = 1, \dots, \widetilde{m}$. 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. The weights are hard to be obtained in practice, hence the newly proposed method, independent hypothesis weighting (IHW) generates weights based on data. The naive version first divides the hypothesis tests into groups based on the values of covariate $X = (X_1, X_2, \dots, X_m)$. Then the weighted BH procedure is applied with each possible weight vector $w = (w_1, w_2, \dots, w_m)$, while the optimal w^* is the vector that lead to the most rejections. In this case, the decision rule becomes we reject hypothesis i if p-value is no more than $\hat{t} \cdot \widehat{W}^{-\ell}(X_i)$, where $i \in I_\ell$. Here $I_\ell, \ell = 1, \dots, K$ is a partition of the hypotheses into K disjoint folds, while ensuring the independence between the covariate and the p-values; and $\widehat{W}^{-\ell}(X_i)$ is a weight function denoting the fact that each of the functions is learned from the remaining $K - 1$ folds (Ignatiadis and Huber 2021).

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.

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.