Abstract:

The study aims to examine and contrast several techniques of controlling the false discovery rate (FDR) in the context of multiple hypothesis testing problems. So far I implemented and analyzed the classical Benjamini-Hochberg procedure, q-value estimation with π_0 adjustment, and a target-decoy approach. Using simulated data with known ground truth, I tested the effectiveness of these methods in controlling false discoveries while maintaining statistical power. The future results may demonstrate the practical implications of each approach and provide insights into their relative strengths and limitations under different conditions. The study contributes to the understanding of FDR control methods and their application.

1. Introduction:

1.1 Background

Multiple hypothesis testing occurs when a statistical analysis involves conducting numerous simultaneous tests, each with the potential to yield a "discovery". When conducting numerous simultaneous statistical tests, the probability of obtaining false positive results increases substantially, which is known as the multiple testing problem. Traditional methods of controlling familywise error rate (FWER) are often too stringent for many practical applications, potentially missing important discoveries.

The False Discovery Rate (FDR), introduced by Benjamini and Hochberg in 1995, offers a more balanced approach to multiple testing correction. FDR is defined as the expected proportion of false positives among all rejected null hypotheses. This metric has gained widespread adoption because of the ability to maintain statistical power while providing meaningful control over false discoveries.

1.2 Problem Statement

Practical implementation of FDR methods faces several challenges:
- The true proportion of null hypotheses (π_0) is typically unknown
- The choice between various FDR control methods can significantly impact research outcomes
- Different application domains may require different levels of stringency in false discovery control

1.3 Objectives

The study aims to:
1. Implement and compare approaches to FDR control
2. Evaluate the performance of these methods using simulated data with known ground truth
3. Analyze the trade-offs between false discovery control and statistical power
4. Provide practical insights into method selection based on specific analysis requirements

2. Theoretical Framework

2.1 Statistical Foundations

The fundamental challenge in multiple hypothesis testing is in simultaneously evaluating numerous statistical hypotheses while controlling false discoveries. Consider a scenario where we test m null hypotheses, of which m_0 are truly null and $m_1 = m - m_0$ are truly alternative. The outcomes can be categorized as follows:

	Declared non-significant	Declared significant	Total
True null	TN	FP	m_0
True alternative	FN	TP	m_1
Total	m-R	R	m

Where:
- TN: True Negatives
- FP: False Positives
- FN: False Negatives
- TP: True Positives
- R: Total number of rejected null hypotheses

2.2 False Discovery Rate

The False Discovery Rate is defined as:

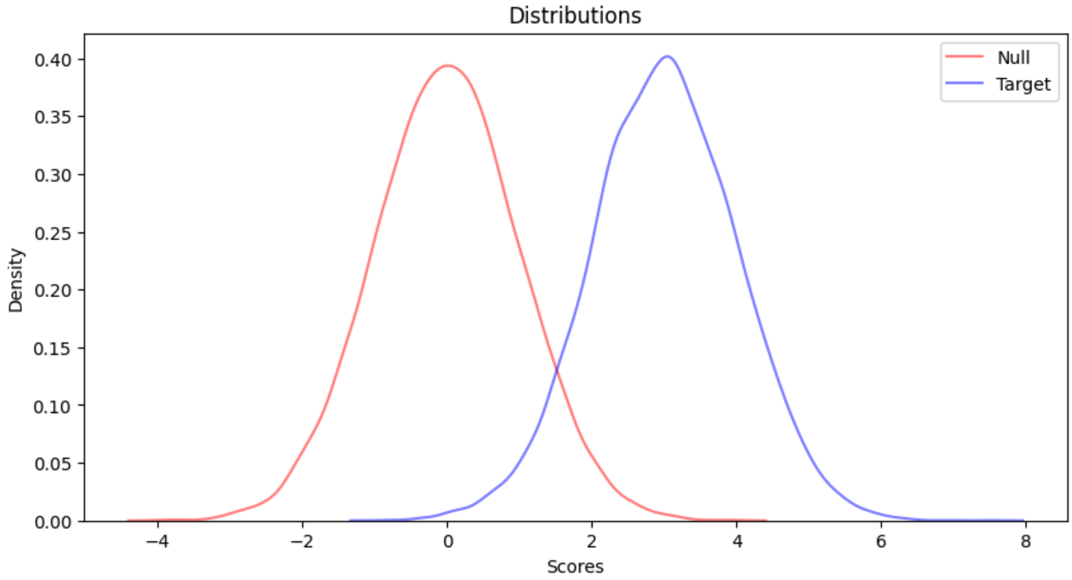
$$\mathrm{FDR} = \mathrm{E}[\frac{FP}{R} \mid \mathrm{R} > 0]\cdot \mathrm{P}(\mathrm{R} > 0) \quad [1]$$

This represents the expected proportion of false discoveries among all discoveries made. Unlike the familywise error rate (FWER), which controls the probability of making even one false discovery, FDR allows for a more balanced approach to error control.

3 Methods for FDR Control

3.1 Setting

I generated two overlapping normal distributions, consisting of 10k observations each:
- A null distribution centered at 0 (representing true null hypotheses)
- A target distribution centered at 3 (representing true alternative hypotheses)



The overlapping nature of the distributions is intentional, as it reflects a realistic testing environment with the challenge of distinguishing between true signals and random noise, a core problem that these methods are designed to address.

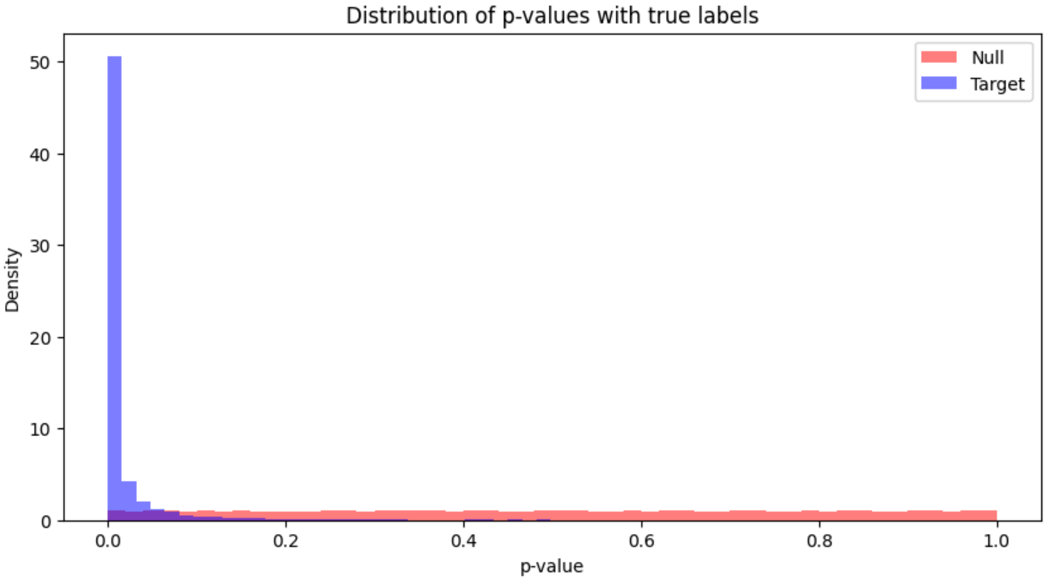
3.2 Benjamini-Hochberg Procedure

The classical BH procedure operates as follows:

1. Choose a significance level α (e.g., α=0.05).
2. Arrange the p-values p_1, p_2, …, p_m from m hypothesis tests in ascending order: $p_1 \leq p_2 \leq \cdots \leq p_m$

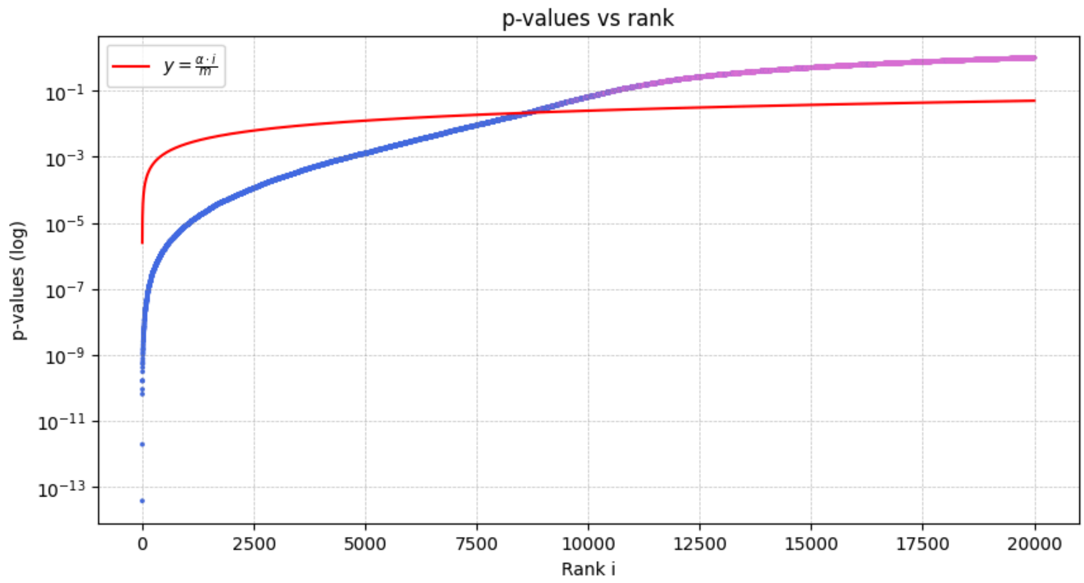
3. For each rank i (where i=1,2,…,m), calculate threshold: $\frac{i \cdot \alpha}{m}$
4. Find the largest i such that: $p_{\mathrm{i}} \leq \frac{i \cdot \alpha}{m}$
5. Reject the null hypotheses for all tests with p-values p_1, p_2, …, p_{i}

The p-values were calculated using the right-tailed probability: $P(X \geq x) = 1 - \Phi(x)$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. This calculation represents the probability of observing a test statistic as extreme or more extreme than the one observed, under the null hypothesis.



On the plot it is highlighted from which distribution the p-values originate, however in real setting this information is not available. It can be seen that under the null hypothesis p-values are uniform and for the target distribution most values are shifted to the right (mean 3) and therefore most of them will get very small p-values.

The choice of significance level α = 0.05 for FDR control means we are willing to accept that 5% of the declared discoveries may be false. 5% was chosen since it has been a widely accepted standard in statistical hypothesis and will be further explored in my study.



The Benjamini-Hochberg procedure establishes a critical threshold line $y = \frac{\alpha i}{m}$, where i is the rank of a p-value when sorted in ascending order, m is the total number of tests, and α is the desired false discovery rate (FDR).

For lower ranks, corresponding to the smallest p-values, the threshold is the most stringent, requiring these p-values to be very small to qualify as significant. As rank increases and p-values grow larger, the threshold becomes progressively more lenient, accommodating larger p-values as significant discoveries. The threshold peaks at α, reflecting the proportion of allowable false discoveries.

This dynamic thresholding prioritizes detecting strong signals (small p-values) early while gradually widening the acceptance criteria to allow weaker signals, provided they do not exceed the controlled FDR.

- The procedure found 8685 significant discoveries at 5% FDR
- The actual FDR achieved 0.0263 was below the target of 0.05, indicating conservative control.
- The statistical power achieved 0.8457 demonstrates good ability to detect true positives.

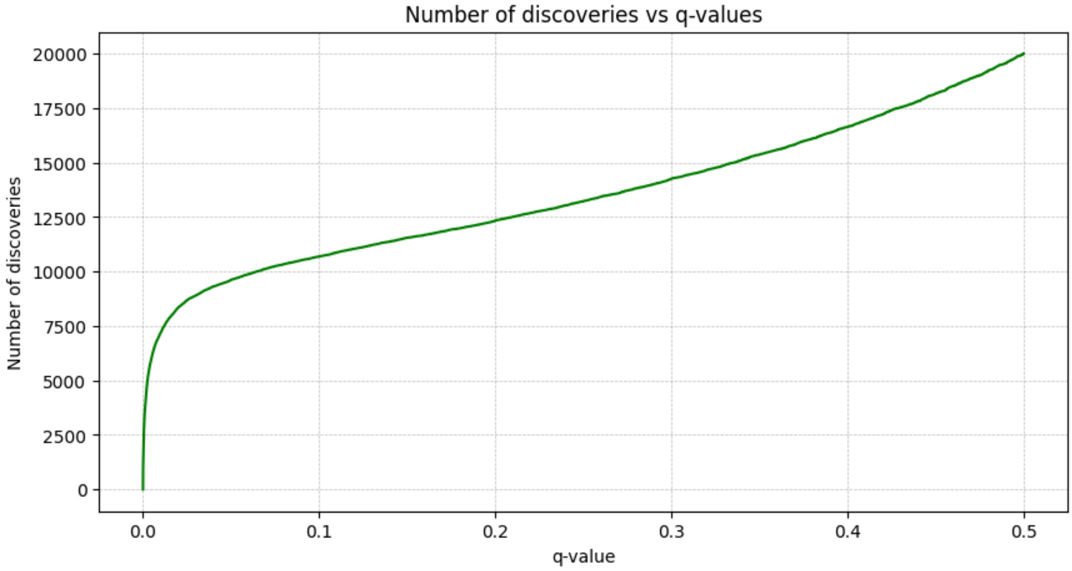
The classical BH procedure assumes that all hypotheses under consideration are null (π_0=1), meaning the proportion of true null hypotheses is 100%. In practice, this assumption often leads to overly conservative results, as can be seen in my setting, where the actual $\pi_0 = 0.5$. To address this limitation, the concept of q-values extends the BH framework by incorporating an estimate of π_0.

3.3 Q-value estimation with π_0 adjustment

The concept of q-values refines the B-H framework by providing an intuitive approach to controlling the false discovery rate. A q-value, in essence, represents the minimum FDR threshold (or significance level, α) at which a particular test would be considered significant. While the Benjamini-Hochberg (BH) procedure directly compares p-values to a threshold determined by α, q-values allow for a more flexible interpretation. They allow adjustments to α without requiring a full recalculation for the dataset.

The q-value approach is the following: [2]

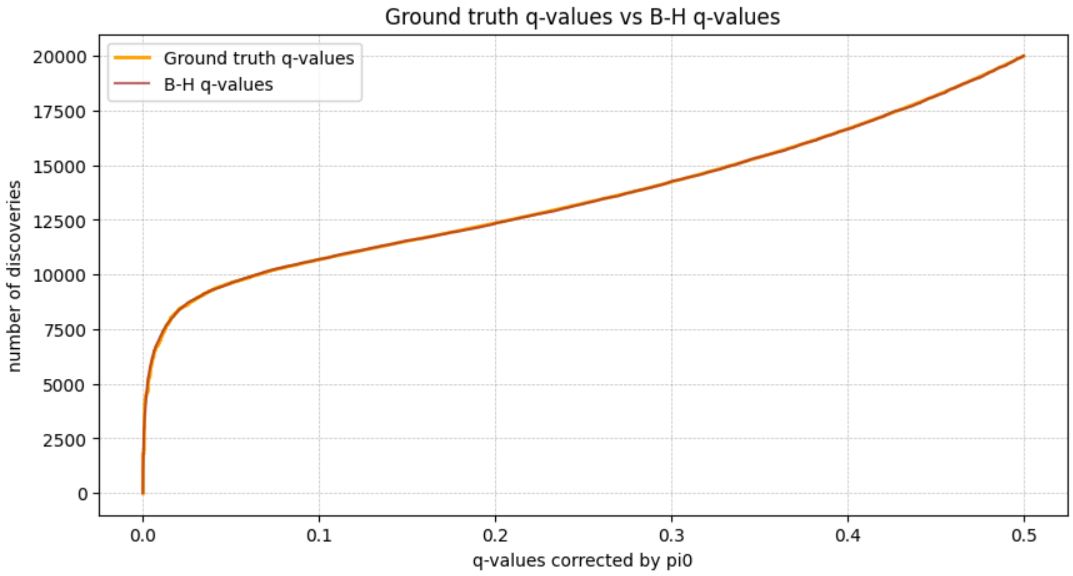
1. Arrange the p-values p_1, p_2, ..., p_m from m hypothesis tests in ascending order: $p_1 \leq p_2 \leq \cdots \leq p_m$
2. The proportion of true null hypotheses π_0, is estimated to account for the proportion of tests that are expected to yield no true discoveries.
3. For each p-value at rank i calculate $q(i) = \frac{m \cdot p(i) \cdot \pi 0}{i}$
4. Start from the largest rank and adjust each q-val as $q(i) = \min(q(i), q(i+1))$
5. Set an FDR threshold and identify all hypotheses below it as significant discoveries.



The estimation of π_0 will be further explored, but setting it equal to 0.5, which is the true proportion for the generated dataset, resulted in 9624 significant discoveries, a notable increase over the original B-H, and 0.9138 as statistical power at the same $\alpha = 0.05$.

In the region of very small q-values (q≤0.05), the curve shows a steep increase in the number of discoveries. This indicates that the most significant hypotheses (with the smallest q-values) are detected early, consistent with the prioritization of strong signals in the q-value framework.

I calculated the cumulative ground truth FDR , derived from true labels of the scores as $\frac{FP}{TP + FP}$. Essentially it is a true proportion of false discoveries up to each rank assuming the threshold is put there.



Controlling the FDR is more flexible with the q-values. By comparing each hypothesis's q-value to any desired FDR threshold, q-values allow assessing significance dynamically.

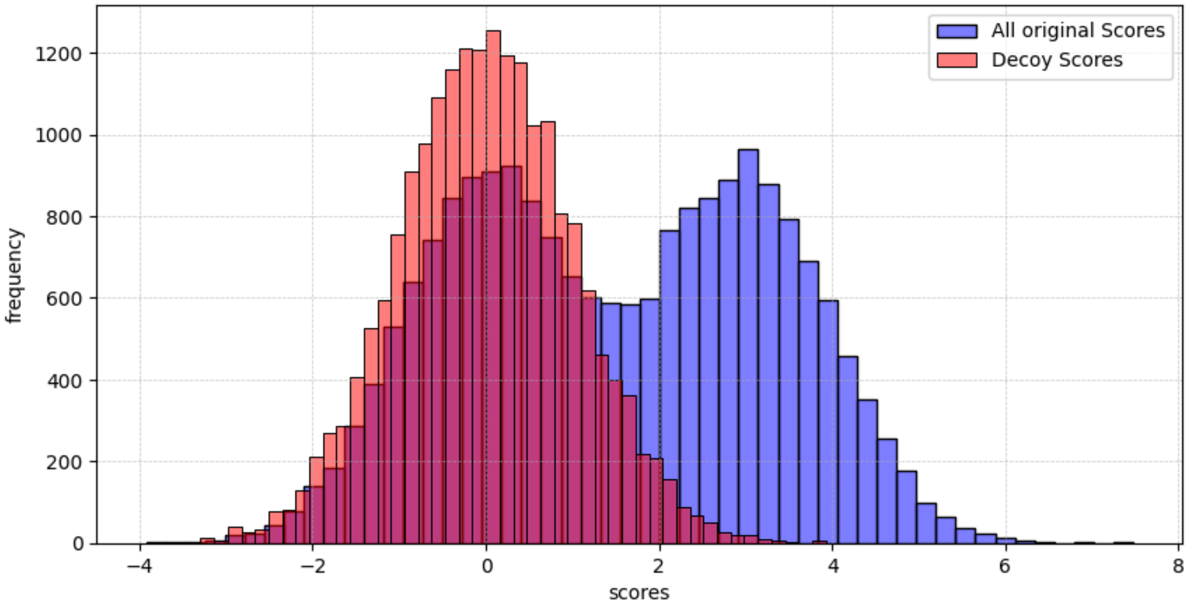
3.4 Target-decoy approach

The Target-Decoy approach is an alternative and widely used method for controlling the False Discovery Rate in fields like proteomics and genomics. The technique involves creating a "decoy" set of data that mimics the structure of the target data but is known to contain no true signals. By comparing results from the target and decoy datasets, the FDR can be estimated empirically without relying on theoretical assumptions about the distribution of p-values.

The procedure typically involves the following steps: [3]

1. Generate decoy data: Construct a decoy dataset with the same structure as the target dataset (actual experimental data, representing the hypotheses to be tested), ensuring it represents only null hypotheses.
2. Combine target and decoy datasets: Analyze both datasets together using the same statistical methods to assign scores to each hypothesis.
3. Rank by score: Arrange all hypotheses (target and decoy) in descending order of their statistical significance (e.g., scores).
4. Estimate FDR

I created a decoy group from the original null distribution.



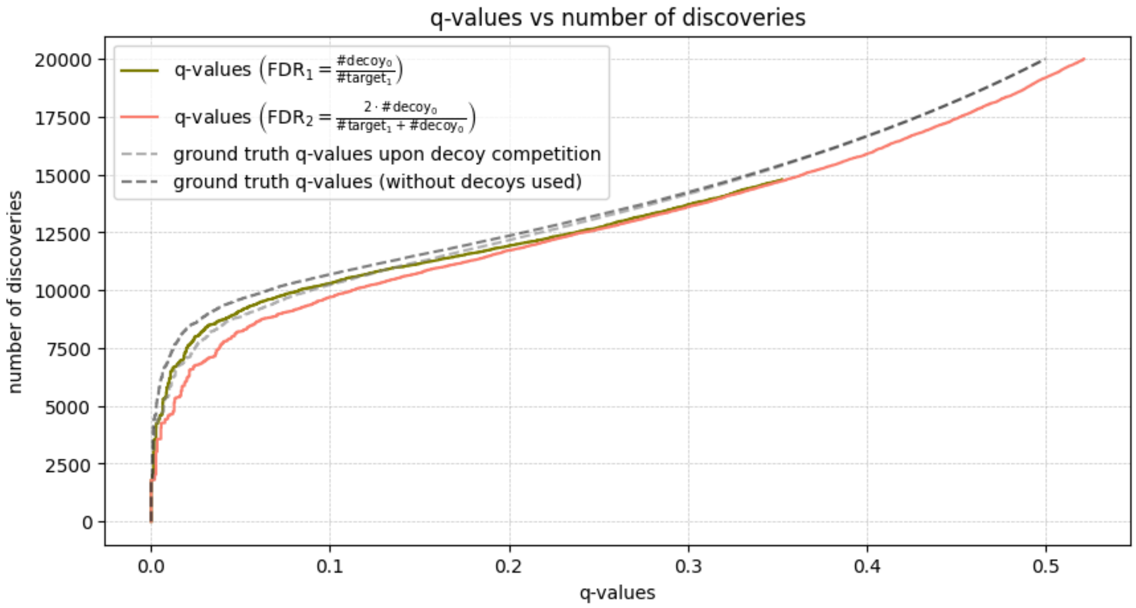
Two methods for estimating FDR in the target-decoy approach were implemented:

$$\mathrm{FDR}_1 = \frac{\#decoy(pred.label=0)}{\#target(pred.label = 1)}$$

The numerator counts decoy scores that were classified as non-significant (pred.label=0), while the denominator counts target scores classified as significant (pred.label=1).

$$\mathrm{FDR}_2 = \frac{2 \cdot \#decoy(pred.label{=}0)}{\#target(pred.label = 1) + \#decoy(pred.label{=}0)}$$

The second FDR considers both target and decoy matches in the denominator. FDR_2 is a more conservative estimate of the false discovery rate. By doubling the number of decoy identifications, it accounts for the possibility that the decoy set might not perfectly mimic the false positive rate of the target set.



The Target-Decoy approach differs from the Benjamini-Hochberg (BH) procedure and the q-value framework in that it does not rely on assumptions about the distribution of p-values. It uses the decoy dataset to empirically estimate the FDR.

However, the method's effectiveness depends on:
- The quality of the decoy generation process
- The assumption that false positives are equally likely to match target or decoy data
- The competitive scoring between target and decoy matches

To be continued

References:

[36] Benjamini, Yoav, and Yosef Hochberg. “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.

[37] Storey, John D. “A Direct Approach to False Discovery Rates.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 3, 2002, pp. 479–98.

[38] Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods, 4(3), 207-214.