Distribution-free multiple testing

Ghita Halib, Samy Vilhes

March 2024

1 Framework

1.1 Context

In the statistical framework of multiple testing, our focus lies in testing a significant number n of hypotheses, denoted as $\mathbf{H}_1, ..., \mathbf{H}_n$. Each hypothesis $\mathbf{H}_i$ is associated with a test statistic $\mathbf{X}_i$. Our objective is to reject the null hypothesis $\mathbf{H}_i$ for large values of $\mathbf{X_i}$. As the number n increases, it becomes more likely to obtain unusual values of $\mathbf{X}_i$ under $\mathbf{H}_0$, leading to erroneous rejection of our hypothesis.
The paper of interest investigates two procedures:

- The Benjamini-Hochberg method (BH), which requires knowledge of the distribution under the null hypothesis.
- The Barber and Candès method (BC), which only requires the assumption that the distribution of our observations under the null hypothesis is symmetric.

The goal of these 2 procedures is to control the False Discovery Rate (FDR)

1.2 The metrics

Let's call

- $\mathcal{F} = \{i \in \{1, ..., n\} : \mathbf{H}_i = (\mu_i > 0)\}$ indexing the false null hypothesis
- $\mathcal{R} = \mathcal{R}(X)$ indexing the hypothesis that the procedure rejects

The FDR is defined with this formula :

$$\mathrm{FDR}(\mathcal{R}) = \mathbb{E}[\mathrm{FDP}(\mathcal{R}(X))]$$

$$\mathrm{FDP}(\mathcal{R}) = \frac{\#\ \mathcal{R}\backslash\mathcal{F}}{\#\ \mathcal{R}}$$

The False Non-Discovery Rate (FNR) is defined with this formula :

$$\mathrm{FNR}(\mathcal{R}) = \mathbb{E}[\mathrm{FNP}(\mathcal{R}(X))]$$

$$\mathrm{FNP}(\mathcal{R}) = \frac{\#\ \mathcal{R}\backslash\mathcal{F}}{\#\ \mathcal{R}}$$

1.3 Threshold procedures

Both the BH and BC methods are considered as threshold procedure. A multiple testing procedure is of treshold type if :

$$\mathcal{R}(\mathbf{X}_1, ..., \mathbf{X}_n) = \{i : \mathbf{X}_i \geq \tau(\mathbf{X}_1, ..., \mathbf{X}_n)\}$$

1.4 The detection boundary

Prior research demonstrates a first-order asymptotic result.

Let $m \leq n$ false hypothesis. The μ_i associated are set to $\mu_i = (\gamma r \log n)^{\frac{1}{\gamma}} = \sqrt{2r \log n}$ (in the normal model) and the others are set to 0.

Let set a $\beta > 0$ and moreover $\beta > \frac{1}{2}$ in order to be in the sparse regime.

An interesting parameterization to consider is $\frac{m}{n} \sim n^{-\beta}$

The detection boundary was shown to be located at $r = \rho(\beta)$ with :

$$\rho(\beta) = \begin{cases} \beta - \frac{1}{2}, & \beta \in]\frac{1}{2}, \frac{3}{4}] \\ (1 - \sqrt{1-\beta})^2, & \beta \in]\frac{3}{4}, 1] \end{cases}$$

When $r < \rho(\beta)$, all tests exhibit a risk of at least 1 in the large sample limit, akin to random guessing. And when $r > \rho(\beta)$ the likelihood ratio test has risk 0 in the large sample limit.

In our context, assuming that $\beta \in [0, 1[$ and $0 \leq r < \beta$, the risk of any threshold procedure has limit inferior at least 1 as $n \to \infty$.

But if instead $r > \beta$, then the risk of the BH method tends to 0 as $n \to \infty$

The "detection boundary" is defined by the equation $r = \rho(\beta)$ while the equation $r = \beta$ defines the "selection boundary".

For $\beta \in]0, 1[$, $\beta > \rho(\beta)$, this confirm us that "selection" is harder than "detection"

2 Optimal Strategies in Statistical Multiple Testing

2.1 Model Settings

In the context of the article, we assume that the test statistics $(X_i)_{i \in [n]}$ are independent where $X_i \sim \Psi_i$[1], $\forall i \in [n]$. The function Ψ_i represents the shifted survival function $\Psi(\cdot - \mu_i)$, where the location parameter μ_i is determined based on the following hypothesis:

- Under the null hypothesis $\mathbf{H}_i$, the location parameter μ_i is set to zero ($\mu_i = 0$).

[1] This notation means that $\Psi_i(t) = P(X_i \geq t)$.

- Under the alternative hypothesis, the location parameter μ_i is greater than zero ($\mu_i > 0$).

We remark that the Ψ function corresponds to the survival function of the X_i's under the null hypothesis.

We limit our analysis to the case where the function Ψ is Asymptotically Generalized Gaussian (AGG) on the right with exponent $\gamma > 0$. This condition is characterized by the following equality:

$$\lim_{x \to \infty} x^{-\gamma} \log \Psi(x) = -\frac{1}{\gamma}. \tag{1}$$

This class of function includes the normal distribution ($\gamma = 2$) and the double exponential distribution ($\gamma = 1$). In our case we assume that γ [2]≥ 1 ensuring that the null distribution has a sub-exponential right tail, maintaining robustness against extreme values.

We also introduces a prior distribution on the location parameters $(\mu_i)_{i \in [n]}$. Specifically, from the total of n parameters, m randomly chosen μ_i's are assigned a non-zero value ($\mu > 0$) to indicate signal strength, while the remaining are set to zero. Following the AGG model assumption, the parameters m and μ are set as follows:

$$m = \left\lfloor n^{1-\beta} \right\rfloor, \text{ with } 0 < \beta < 1 \text{ (fixed)}, \tag{2}$$

and

$$\mu = \mu_\gamma(r) = (\gamma r \log n)^{1/\gamma}, \text{ with } r > 0 \text{ (fixed)}. \tag{3}$$

These equations highlight the critical functions of the parameters β and r in our multiple testing scenario: β primarily regulates the sparcity of true positive findings within the tested set, directly influencing the expected number of false hypotheses, where a higher β leads to a lower count of m(e.g. the number of genuinely non-null hypotheses), thus indicating fewer false positives and a more stringent selection of truly significant hypotheses. Concurrently, r adjusts the detectable signal strength, modulating how distinctly a real effect stands out against the background noise assumed by the null hypothesis. This dynamic between β and r fine-tunes the sensitivity and specificity of statistical tests, ensuring effective detection of true positives while reducing the risk of false discoveries.

2.2 Theorem 1

Building on the framework outlined previously, the following theorem, as presented in the article, can be applied:

[2] γ serves as an exponent characterizing the tail behavior of the distribution of Ψ. A larger γ indicates a faster decay of the tail, implying that extreme values are less likely. Conversely, a smaller γ suggests a heavier tail, indicating that extreme values are more probable.

Theorem 1. *Given a location model where the base distribution adheres to the Asymptotically Generalized Gaussian (AGG) framework with exponent $\gamma \geq 1$, and considering the prior distribution on the location parameters as described, alongside the parameterization equations* (2)-(3)*, it is posited that if $r < \beta$, then the oracle risk*[3] *defined as follow :*

$$FDP(R_t(X)) + FNP(R_t(X)), \quad R_t(X) := \{i : X_i \geq t\}. \tag{4}$$

approaches or exceeds 1 as $n \to \infty$.

This theorem provides that when the intensity of true effects (e.g., r), which reflects the detectability of genuine findings, is lower than the sparsity level (e.g., β), indicative of the proportion of true non-null hypotheses, the efficacy of statistical testing strategies significantly diminishes. Essentially, as n increases towards infinity, the testing procedure's ability to discern true effects from background noise deteriorates, rendering it as effective as random conjecture. Indeed, a lower r relative to β underscores a scenario where the instances of significant deviations are too subtle and infrequent, making them increasingly difficult to identify amid the noise as more hypotheses are tested.

It is noteworthy that as the lower limit of the oracle risk, serving as a benchmark for evaluating the performance of a procedure under ideal conditions with perfect knowledge of the distributional parameters, approaches or surpasses 1, it signals that, even under the most favorable scenarios, any multiple testing procedure struggles to outperform the effectiveness of random selection in situations characterized by sparse and weak effects (e.g., when $r \leq \beta$).

Remark: The proof hinges on establishing that the number of Type I and Type II errors in our procedure adhere to specific binomial distributions.

The whole proof of theorem 1 is provided in the last section of the article.

2.3 The BH Method

The Benjamini-Hochberg (BH) procedure is a method used for controlling the FDR in multiple hypothesis testing scenarios. Given a set of n independent hypothesis tests with corresponding p-values $p_1, \ldots, p_n$ and a desired FDR level q, the procedure is as follows:

1. Order the p-values in increasing order: $p_{(1)} \leq \cdots \leq p_{(n)}$.
2. Find $k = \max\left\{i \in [n] : p_{(i)} \leq \frac{i}{n}q\right\}$.
3. Reject all hypotheses corresponding to $p_{(1)}$ to $p_{(k)}$

[3]The article introduces the concept of the oracle procedure as an idealized theoretical construct that assumes access to perfect information—namely, the set $\mathcal{F}$, which contains all false null hypotheses (e.g., $|\mathcal{F}| = m$)

Alternatively, when applying the BH procedure in terms of test statistics X_i as in the article, assuming they are ordered in decreasing order as $X_{(1)} \geq \cdots \geq X_{(n)}$, the threshold τ_{BH} is defined using:

$$\tau_{\mathrm{BH}} = X_{(\iota_{\mathrm{BH}})}, \quad \iota_{\mathrm{BH}} := \max\left\{ i \in [n] : X_{(i)} \geq \Psi^{4-1}\left(\frac{iq}{n}\right) \right\}. \tag{5}$$

To demonstrate the equivalence of ι_{BH} and k, consider:

$$X_{(i)} \geq \Psi^{-1}\left(\frac{iq}{n}\right) \Leftrightarrow \Psi(X_{(i)}) \leq \frac{iq}{n}^{5} \Leftrightarrow p_{(i)} \leq \frac{iq}{n}^{6}$$

This equivalence establishes that the criterion for rejecting hypotheses based on ordered p-values and the criterion based on the comparison of test statistics to a threshold determined by Ψ^{-1} are essentially expressing the same condition for controlling the FDR at the level q.

2.4 Theorem 2

For the rest of our study, we maintain the assumption that:

$$q = q(n) > 0 \text{ such that } n^a q(n) \to \infty \text{ for all fixed } a > 0 \tag{6}$$

Hence we can apply the following theorem :

Theorem 2. *In the setting of Theorem 1, if instead $r > \beta$, then the BH procedure with q satisfying* (6) *has FNR tending to 0 as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0 since the procedure*

This theorem underscores the effectiveness of the BH procedure in multiple hypothesis testing within the AGG model. Specifically, it demonstrates that when the strength of the signal (r) exceeds the signal sparsity (β), and the control parameter $q(n)$ satisfies (6), the FNR of the BH procedure approaches zero asymptotically as the number of hypotheses (n) grows infinitely large. This convergence of FNR towards zero, with the control of the FDR, highlights the BH procedure's ability to balance the trade-off between minimizing the risk of falsely rejecting true null hypotheses (Type I error) and controlling the risk of incorrectly accepting null hypotheses (Type II error). This implicates the convergence towards zero of the oracle risk, further solidifying the BH procedure's capability to approach optimal performance in identifying true positives while

[4] Note that Ψ_i must be known under the null hypothesis (hence denoted as Ψ) to apply the BH procedure.

[5] We recall that Ψ is a non-increasing function.

[6] We remark on the relationship between ordered p-values and ordered test statistics: the p(i) p-values corresponds to the X(i) statistics defined above

effectively controlling errors.

Proof Sketch for Main Results: The proof revolves around the crucial remark that for any multiple testing procedure, FNR $\to 0$ if and only if FNP $\to 0$ in probability. Let's delve into the proof:

:⇒
Considering the FNR as the expected value of FNP, Markov's inequality is applied for any $\epsilon > 0$:

$$\mathbb{P}(\text{FNP} > \epsilon) \leq \frac{\mathbb{E}[\text{FNP}]}{\epsilon} \tag{7}$$

As FNR (and thus $\mathbb{E}[\text{FNP}]$) approaches zero, it implies that for all $\epsilon > 0$, $\mathbb{P}(\text{FNP} > \epsilon) \to 0$. This demonstrates that FNP converges to zero in probability.

⇐:
Given that FNP is converging to zero in probability and is bounded within $0 \leq \text{FNP} \leq 1$, the Dominated Convergence Theorem comes into play:

$$\lim_{n\to\infty} \mathbb{E}[\text{FNP}] = \mathbb{E}[\lim_{n\to\infty} \text{FNP}] = 0 \tag{8}$$

which allows the interchange of limit and expectation, thereby ensuring that FNR trends towards zero.

Thus, to prove Theorem 2, it is sufficient to demonstrate that the $\text{FNP}(\mathcal{R}_{\tau_{BH}})$ converges in probability to zero as n tends to ∞.

The whole proof of theorem 2 is provided in the last section of the article.

2.5 The BC Method

The Barber-Candès (BC) procedure, like the BH method, controls the FDR. However, unlike the BH procedure, the BC method does not require knowledge of the survival function of X_i under the null hypothesis. Operating under the assumption of symmetry to control the FDR at a level q, the BC procedure sets a data-dependent threshold, τ_{BC}, defined as:

$$\tau_{\text{BC}} = \inf\left\{t \in |\mathbf{X}| : \widehat{\text{FDP}}(t) \leq q\right\} \tag{9}$$

where $|\mathbf{X}|$ denotes the set of the sample's absolute values and the estimated False Discovery Proportion given by:

$$\widehat{\text{FDP}}(t) = \frac{1 + \#\{i : X_i \leq -t\}}{1 \vee \#\{i : X_i \geq t\}} \tag{10}$$

measures observation asymmetry for $|X_i| \geq t$. This approach, which does not rely on the null distribution's specifics, effectively controls the FDR at level q.

Proof Sketch for Main Results: Let proove the BC method:
We have

$$FDP(\mathcal{R}_{\tau_{BC}}) = \frac{\#\{i \in [n] : \mu_i = 0 \text{ and } X_i \geq \tau_{BC}\}}{\#\{i \in [n] : X_i \geq \tau_{BC}\} \vee 1} \tag{11}$$

where τ_{BC} is defined in (9).
Thus

$$\begin{aligned} FDP(\mathcal{R}_{\tau_{BC}}) &= \frac{1 + \#\{i \in [n] : \mu_i = 0 \text{ and } X_i \leq -\tau_{BC}\}}{\#\{i \in [n] : X_i \geq \tau_{BC}\} \vee 1} \cdot \frac{\#\{i \in [n] : \mu_i = 0 \text{ and } X_i \geq \tau_{BC}\}}{1 + \#\{i \in [n] : \mu_i = 0 \text{ and } X_i \leq -\tau_{BC}\}} \\ &\leq \frac{1 + \#\{i \in [n] : X_i \leq -\tau_{BC}\}}{\#\{i \in [n] : X_i \geq \tau_{BC}\} \vee 1} \cdot \frac{\#\{i \in [n] : \mu_i = 0 \text{ and } X_i \geq \tau_{BC}\}}{1 + \#\{i \in [n] : \mu_i = 0 \text{ and } X_i \leq -\tau_{BC}\}} \\ &= \widehat{\mathrm{FDP}}(\tau_{BC}) \cdot \frac{\#\{i \in [n] : \mu_i = 0 \text{ and } X_i \geq \tau_{BC}\}}{1 + \#\{i \in [n] : \mu_i = 0 \text{ and } X_i \leq -\tau_{BC}\}} \\ &\leq q \cdot \frac{\#\{i \in [n] : \mu_i = 0 \text{ and } X_i \geq \tau_{BC}\}}{1 + \#\{i \in [n] : \mu_i = 0 \text{ and } X_i \leq -\tau_{BC}\}} \end{aligned} \tag{12}$$

Given that τ_{BC} represents the first instance when the $\widehat{\mathrm{FDP}}(t)$ falls below q, we can interpret τ_{BC} as a stopping time for the supermartingale $V_+(T)/(1+V_-(T))$, where $V_\pm(t) = \#\{i \in [n] : \mu_i = 0 \text{ and } |X_i| \geq t \text{ and } \mathrm{sign}(X_i) = \pm 1\}$. Thus applying the optional stopping theorem on this supermarty-ingal, we find:

$$\mathbb{E}\left[\frac{V_+(T)}{1 + V_-(T)}\right] \leq \mathbb{E}\left[\frac{V_+(0)}{1 + p_0 - V_+(0)}\right] \leq 1 \tag{13}$$

where the last steps leverages a property of the binomial distribution $V_+(0)$ is modeled as a Binomial$(p_0, 1/2)$ random variable. This result, combined with the (12) inequality, confirms the control of FDR on the BC procedure .

2.6 Theorem 3

Likewise for the BH-method there is a theorem that ensure the convergence to zero of the oracle risk when the number of hypothesis increase in our framework:

Theorem 3. *In the setting of Theorem 1, and assuming that the null distribution Ψ is symmetric about 0 , if instead $r > \beta$, then the BC procedure with q satisfying* (6) *has FNR tending to 0 as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0 since the procedure has FDR $\leq q$.*

Following Theorem 2, Theorem 3 explores the BC method's efficacy, this time under the expanded assumption that the null distribution is symmetric about zero. Leading to conclusions for the BC procedure that mirror the conclusions drawn for the BH procedure in Theorem 2.

Idea of proof: As established in the proof of Theorem 2, the proof of Theorem 3 relies on demonstrating the convergence in probability of $\mathrm{FNP}(\mathcal{R}_{\tau_{BC}})$

(where τ_{BC} is defined in (9)) to 0 as $n \to \infty$.

We establish $\tau = |X|_{(\iota_{\mathrm{BC}}+1)}$ if $\iota_{\mathrm{BC}} < n$, and $\tau = 0$ if $\iota_{\mathrm{BC}} = n$, with $\tau_{\mathrm{BC}} = X_{(\iota_{\mathrm{BC}})}$.

It follows from the definition of τ and τ_{BC} that $\tau \leq \tau_{\mathrm{BC}}$. Therefore, by definition of $\mathrm{FNP}(\mathcal{R}_\tau)$ and $\mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}})$, we have $\mathrm{FNP}(\mathcal{R}_\tau) \leq \mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}})$. Additionally, $\mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}})$ is an upper bound on $\mathrm{FNP}(\mathcal{R}_\tau)$ by at most $\frac{1}{m}$ due to the definition of the FNP and the fact that τ_{BC} is derived from the ordered statistics. This is because adding τ_{BC} to the set of rejections at most increases the proportion of false negatives by $\frac{1}{m}$. Hence, we obtain the following inequality:

$$\mathrm{FNP}(\mathcal{R}_\tau) \leq \mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}}) \leq \mathrm{FNP}(\mathcal{R}_\tau) + \frac{1}{m}.$$

Thus, to show $\mathrm{FNP}(\mathcal{R}_{\tau_{\mathrm{BC}}})$ in probability, we have to show $\mathrm{FNP}(\mathcal{R}_\tau) \to 0$ in probability.
The whole proof of theorem 3 is provided in the last section of the article.

3 Numerical Experiments

3.1 Settings

For the simulation, they tested $n = 10^5$ hypotheses, setting the FDR control level at $q = 0.05$. Since we utilize the BH procedure, knowledge of the distribution under the null hypothesis is required. Alternatively, for the BC procedure, the distribution under the null hypothesis must be symmetric around 0. So, the distributions used for simulation are the normal distribution and the double exponential distribution, both with a scale of 1. The location parameter is selected accordingly to (3)

The constant β takes few values to show different sparsity levels and the constant r takes values in $[0, 1]$. We draw m observations from the non-null hypothesis accordingly to (2). So we have $n - m$ observations under the null-hypothesis.

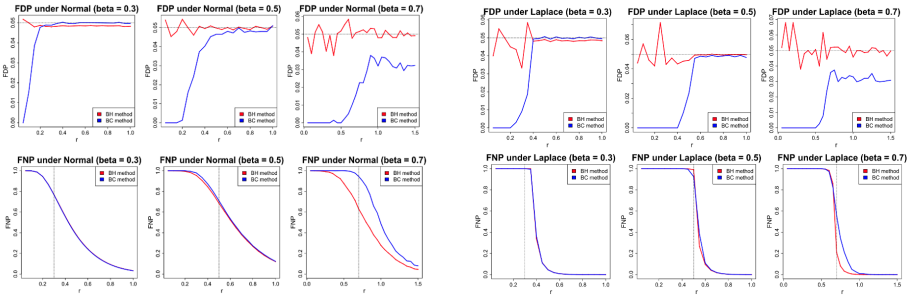
3.2 Results

3.2.1 Fixed sample size

Concerning the FDR, which we aim to control at the level $q = 0.05$, the BC method proves to be more conservative compared to the BH method. In essence, the BC method exhibits a lower FDR than the BH method, leading to fewer rejections.

For $\beta \in \{0.3; 0.5\}$, both methods demonstrate similar performance for moderately small values of r. However, when $\beta = 0.7$, the BC method exhibits reduced effectiveness. This can be attributed to a sparsity issue: in this scenario,

with $m = 31$, the number of false hypotheses is insufficient to fully manifest the power of the BC method.



(a) Normal model (b) Laplace model

Figure 1: Simulation results showing the FDP and FNP in three different sparsity settings from the paper

Due to time and computational constraints, we currently present results solely under the Normal model and conduct only 10 simulations for each setting, compared to the 500 simulations in the paper.

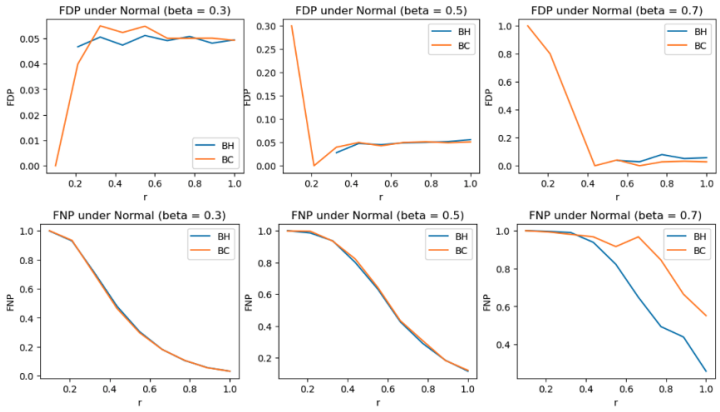
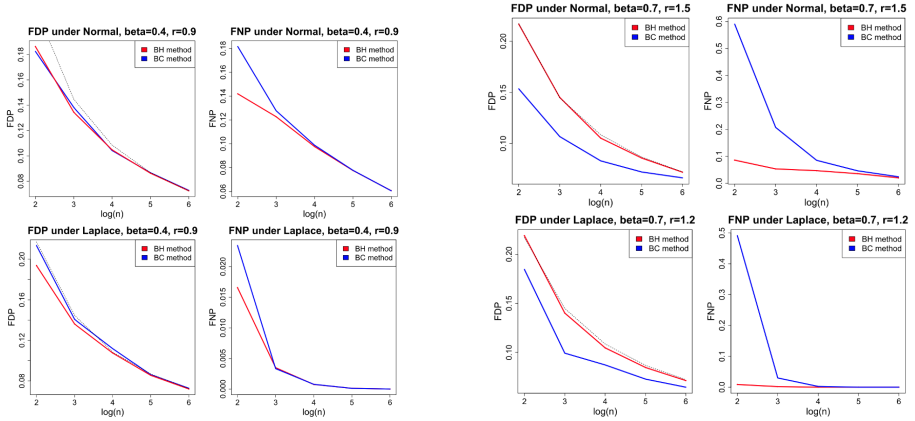


Figure 2: Simulation results showing the FDP and FNP in three different sparsity settings by oursleves

Regarding the FNP, our results closely resemble those reported in the paper. However, we encountered a slight discrepancy in FDP for small values of r (the signal), likely due to issues in the code implementation for the BH method. Nevertheless, it is reassuring to observe that as r increases, the FDP stabilizes around our specified control level of the FDR q.

3.2.2 Varying sample size

Previously, the sample size remained constant; however, it will now serve as a parameter that we vary to observe its effects on the False Discovery Proportion (FDP) and False Discovery Rate (FDR). As we increase the sample size, we will concurrently decrease the control level of the FDR to $q = \frac{1}{\log n}$. Additionally, we select pairs of (β, r) with $\beta < r$.



(a) First setting (b) Second setting

Figure 3: Simulation results showing the FDP and FNP varying the sample size for different settings from the paper

In the initial scenario, both the FDP and FNP decrease rapidly with increasing n. Regarding the FNP, the BH method outperforms BC method for $n \leq 10^3$, after which they exhibit comparable performance.
In the second scenario, the BC method demonstrates a lower FDP while displaying a higher FNP for both distributions, indicating its greater conservatism for small values of n. However, as n increases, the results become increasingly similar for both methods.

3.3 Conclusion

In conclusion, our findings emphasize the importance of considering the parameters β (sparsity), r (signal), and n (sample size) as factors that influence the performance of the BH and BC methods for FDR and FNR control. BH, reliant on knowledge of the null hypothesis distribution, demonstrates greater efficacy for smaller sample sizes, while BC, relying only on symmetry, tends to be more conservative initially. However, as sample size increases, the distinctions in performance between the two methods diminish.

4 Bibliography

Ery Arias-Castro and Shiyun Chen, ***Distribution-free multiple testing***, https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-1/Distribution-free-multiple-testing/10.1214/17-EJS1277.full

Benjamini, Y. and Y. Hochberg, ***Controlling the false discovery rate: A practical and powerful approach to multiple testing***, http://www.ams.org/mathscinet-getitem?mr=1325392

Foygel-Barber, R. and E. J. Candès , ***Controlling the false discovery rate via knockoffs***, http://www.ams.org/mathscinet-getitem?mr=3375876