False Discovery Control in Multiple Testing: A Selective Overview of Theories and Methodologies

Jianliang He[*]

Abstract

Contemporary data analysis often involves large-scale hypothesis testing, a fundamental problem in statistical inference and model selection. With the paradigm shift in data collection patterns, modern datasets often possess distinct characteristics, including large magnitude, sparse signal, rich auxiliary information, and nuanced dependence structures. This poses significant challenges and necessitates innovations in theories and methodologies. In this paper, we focus on multiple testing with false discovery rate (FDR) control and offer a selective overview of conventional methods and the recent advancements in this domain, including the choice of summary statistics, estimation, and detection. We start with the celebrated Benjamini-Hochberg (BH) procedure and discuss the issue of sparsity from both frequentist and Bayesian perspectives. Following the classical framework, we move to the problem of accommodating side information, and discuss the issue of dependence. A brief introduction to related topics is also provided.

Contents

[*] Department of Statistics and Data Science, Fudan Univeristy. Email: `hejl20@fudan.edu.cn`.

1 Introduction

The issue of multiple comparisons arises when conducting numerous hypothesis testings (Hsu, 1996). In such cases, where there is a collection of hypotheses to be tested, the challenge is to distinguish between null and non-null hypotheses while controlling the error rates. This issue is fundamental in statistical inference and has prompted the development of various procedures. Failure to correct for multiplicity can lead to serious concerns regarding reproducibility, publication bias, and p-hacking in scientific research (Ioannidis, 2005; Head et al., 2015). Specifically, multiplicity is inherently connected to the reproducibility of scientific findings. Goodman et al. (2016) claimed that multiplicity, combined with incomplete reporting, might be the largest contributor to the non-reproducibility or falsity of published claims. Multiplicity adjustment can greatly enhance the reproducibility of results from psychology experiments (Zeevi et al., 2020). This article reviews the recent advancements in this field, particularly on multiple testing with false discovery rate (FDR) control.

The generic problems of multiple testing that arise from feature selection and anomaly detection have long been acknowledged. For instance, consider the prostate dataset (§2, Efron, 2012), which contains a genetic expression for $n = 6033$ genes measured on $m_1 = 52$ patients with prostate cancer and $m_0 = 50$ normal control subjects. Let $X_{ij}^{(k)}$ indicate the expression level on gene i for patient j, where superscripts k indicate whether the observation was collected from patient populations. To identify whether genes have a causal link to the development of prostate cancer, we can formalize the problem by testing the n null hypothesis $H_{0,i} : \mathbb{E}[X_i^{(0)}] = \mathbb{E}[X_i^{(1)}]$ for all $i \in [n]$ and then compute the p-value of the corresponding t-test, illustrating a typical multiple testing problem. Another crucial application area of multiple testing is the detection of anomalous events in financial markets, which includes monitoring for credit card fraud, cyber intrusions, financial market anomalies, and covert communication. In financial economics, there is a growing focus on detecting extreme events in time series data, often utilizing sequential change-point analysis (Lumsdaine and Papell, 1997; Andreou and Ghysels, 2006; Fryzlewicz, 2014). One significant challenge is to identify anomalies in financial markets quickly while controlling the number of false alarms. These large-scale inference problems necessitate processing massive amounts of real-time estimates or testing thousands or even millions of hypotheses with high frequencies, highlighting the importance of multiplicity adjustment.

In many areas of modern applied statistics, from genetics and neuroimaging (Pe'er et al., 2008) to online advertising and finance (Harvey and Liu, 2015), massive datasets with thousands or even millions of variables are consistently collected by institutions and online platforms. This expansive data collection and complexity calls for new techniques for making large-scale statistical inferences, which involve simultaneously performing inferences on many study units. Following this paradigm shift in data collection, several phenomena arise particularly frequently, including *sparsity*, *auxiliary* *sequences*, and *dependence*. While these structures can be informative, they also pose challenges to the design of testing methods. In response, the statisticians aim to answer the following questions: (i) While using p-values is standard in hypothesis testing, is there a more powerful or robust statistics that can better retain the structural information or accommodate complex scenarios? (ii) With the removal of strong assumptions (e.g., i.i.d), in conventional methods, what is the minimal condition to ensure a statistical guarantee, and what is the best guarantee we can get for general cases? (iii) How can we design a more powerful testing procedure by extracting information from internal (e.g., sparsity and dependence) or external (e.g., auxiliary sequence) structure? (iv) Recent advancements have shown that powerful testing procedures rely on a series of complex ranking statistics that must be estimated from data. How can we improve estimation methods or even further accommodate the inaccuracy of estimation? This review offers a selective overview of answers to the questions posed above from both the theoretical and methodological perspectives.

Notation. Throughout the paper, we denote $\mathcal{M}$ or $[m] = \{1, \dots, m\}$. Let $\mathbb{1}(\cdot)$ denote the indicator function that returns 1 if the condition is true and 0 otherwise, and $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. Consider two non-negative sequences $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$, if $\limsup a_n/b_n = 0$, then write $a_n = o(b_n)$. Let $\sigma\{S_t\}$ with $S_t = \{\ell_1, \dots, \ell_t\}$ be the sigma algebra generated by S_t.

2 Problem Formulation

Consider m null hypotheses $H_{0,1}, \dots, H_{0,m}$ and summary statistics $X_1, \dots, X_m$ with a known null distribution, e.g. p-value $P_i \sim \text{Unif}[0,1]$ and z-value $Z_i \sim \mathcal{N}(0,1)$ under the null $H_{0,i}$. A multiple testing procedure involves making simultaneous inferences on m hypotheses:

$$H_{0,i}: \text{ case } i \text{ is null} \quad \text{versus} \quad H_{1,i}: \text{ case } i \text{ is non-null}, \quad i = 1, \dots, m.$$

A testing procedure examines these summary statistics and decides which null hypotheses to reject. Let $\mathcal{H}_0 = \{i : H_{0,i} \text{ is true}\}$ be the set of true null hypotheses, $\mathcal{H}_1 = \mathcal{M}/\mathcal{H}_0$ with $\mathcal{M} = \{1, \dots, m\}$ as the set of non-null hypotheses, and $\mathcal{R} = \{i : H_{0,i} \text{ is rejected}\}$ is the rejection set. For clear presentation, let $\theta_i = \mathbb{1}(i \in \mathcal{H}_0)$ be an indicator function that gives the true state of the i-th testing problem. A selection error, or false positive, occurs if the practitioner asserts that $H_{0,i}$, is false when it is not. In multiple testing problems, such false positive decisions are inevitable if we wish to discover interesting effects with reasonable power. Instead of aiming to avoid any false positives, a practical goal is to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995) small, which is the expectation of false discovery proportion (FDP) among all selections:

$$\text{FDR}(\mathcal{R}) = \mathbb{E}\left[\text{FDP}(\mathcal{R})\right] \text{ where } \text{FDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}, \tag{2.1}$$

and a widely adopted variant marginal FDR (mFDR) (Storey, 2002) is defined as

$$\text{mFDR}(\mathcal{R}) = \frac{\mathbb{E}\left[|\mathcal{R} \cap \mathcal{H}_0|\right]}{\mathbb{E}\left[|\mathcal{R}| \vee 1\right]}. \tag{2.2}$$

More multiplicity-related error rates for simultaneous and selective inference is discussed in §A.1. Similarly, the true discovery rate (TDR) is defined as the expectation of true discovery proportion (TDP), which measures the power of the testing procedure:

$$\text{TDR}(\mathcal{R}) = \mathbb{E}\left[\text{TDP}(\mathcal{R})\right] \text{ where } \text{TDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_1|}{|\mathcal{H}_1| \vee 1}. \tag{2.3}$$

The goal is to find the optimal rejection set $\mathcal{R}$ that maximizes the $\mathrm{TDP}(\mathcal{R})$ subject to $\mathrm{FDR}(\mathcal{R}) \leq \alpha$, where $\alpha \in (0,1)$ is the target FDR level. The following proposition demonstrates that FDR and mFDR are asymptotically equivalent under large-scale scenarios.

Proposition 2.1. Let $\mathcal{R}$ be the rejection set concerning m null hypotheses following a specific decision rule, then $\mathrm{FDR}(\mathcal{R}) = \mathrm{mFDR}(\mathcal{R}) + o(1)$ if the following two conditions hold: 1) there exists an absolute constant $\eta > 0$ such that $m^{-1}\mathbb{E}[|\mathcal{R}|] \geq \eta$, 2) $\mathrm{Var}(|\mathcal{R}|) = o(m^2)$.

Proof of Proposition 2.1. Based on the definition

$$\begin{aligned}\mathrm{FDR}(\mathcal{R}) - \mathrm{mFDR}(\mathcal{R}) &= \mathbb{E}\left[\frac{|\mathcal{R}\cap\mathcal{H}_0|}{|\mathcal{R}|\vee 1}\right] - \frac{\mathbb{E}\left[|\mathcal{R}\cap\mathcal{H}_0|\right]}{\mathbb{E}\left[|\mathcal{R}|\vee 1\right]} \\ &= \mathbb{E}\left[\left(\frac{|\mathcal{R}\cap\mathcal{H}_0|}{|\mathcal{R}|} - \frac{|\mathcal{R}\cap\mathcal{H}_0|}{\mathbb{E}|\mathcal{R}|}\right)\cdot \mathbb{1}(|\mathcal{R}|>0)\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}|\mathcal{R}| - |\mathcal{R}|}{\mathbb{E}|\mathcal{R}|}\cdot\frac{|\mathcal{R}\cap\mathcal{H}_0|}{|\mathcal{R}|}\cdot \mathbb{1}(|\mathcal{R}|>0)\right] \leq \left|\frac{|\bar{\mathcal{R}}| - \mathbb{E}|\bar{\mathcal{R}}|}{\mathbb{E}|\bar{\mathcal{R}}|}\right|,\end{aligned}$$

where the second inequality results from the fact that $|\mathcal{R}\cap\mathcal{H}_0| = 0$ if $|\mathcal{R}| = 0$, and the last equation follows $|\mathcal{R}\cap\mathcal{H}_0| \leq |\mathcal{R}|$ and we write $|\bar{\mathcal{R}}| = |\mathcal{R}|/m$. Note that conditions indicate that 1) $\mathbb{E}[|\bar{\mathcal{R}}|] \geq \eta$ with absolute constant $\eta > 0$, and 2) $\mathrm{Var}(|\bar{\mathcal{R}}|) = o(1)$. Following this, we have

$$\left|\frac{|\bar{\mathcal{R}}| - \mathbb{E}|\bar{\mathcal{R}}|}{\mathbb{E}|\bar{\mathcal{R}}|}\right| \leq \frac{\left(\mathbb{E}\left||\bar{\mathcal{R}}| - \mathbb{E}|\bar{\mathcal{R}}|\right|^2\right)^{1/2}}{\mathbb{E}|\bar{\mathcal{R}}|} = \frac{\mathrm{Var}(|\bar{\mathcal{R}}|)^{1/2}}{\mathbb{E}|\bar{\mathcal{R}}|} = o(1),$$

where the inequality results from the Cauchy-Swartz inequality and then we finish the proof. □

We remark that establishing the asymptotic equivalence between False Discovery Rate (FDR) and modified False Discovery Rate (mFDR) introduces a methodological advancement in large-scale multiple testing. Under this framework, statisticians can devise methods that control the mFDR, thereby achieving an *asymptotic* control over the FDR, i.e., $\mathrm{FDR}(\mathcal{R}) \leq \alpha + o(1)$, where the limit is taken over the number of hypotheses m. Additionally, mFDR proves to be valuable in splitting-based testing procedures (Gang et al., 2023a).If we decide hypotheses into $\mathcal{M} = \mathcal{M}^1 \cup \mathcal{M}^2$ and the mFDR levels of rejections $\mathcal{R}^1$ and $\mathcal{R}^2$ are controlled at α correspondingly, we have:

$$\begin{aligned}\mathrm{mFDR}(\mathcal{R}^1\cup\mathcal{R}^2) &= \frac{\mathbb{E}\left[|\mathcal{R}^1\cap\mathcal{H}_0^1|\right] + \mathbb{E}\left[|\mathcal{R}^2\cap\mathcal{H}_0^2|\right]}{\mathbb{E}[|\mathcal{R}^1|\vee 1] + \mathbb{E}[|\mathcal{R}^2|\vee 1]} \\ &= \underbrace{\frac{\mathbb{E}\left[|\mathcal{R}^1\cap\mathcal{H}_0^1|\right]}{\mathbb{E}[|\mathcal{R}^1|\vee 1]}}_{\mathrm{mFDR}(\mathcal{R}_1)\leq\alpha}\cdot\frac{\mathbb{E}[|\mathcal{R}^1|\vee 1]}{\mathbb{E}[|\mathcal{R}^1|\vee 1] + \mathbb{E}[|\mathcal{R}^2|\vee 1]} \\ &\quad + \underbrace{\frac{\mathbb{E}\left[|\mathcal{R}^2\cap\mathcal{H}_0^2|\right]}{\mathbb{E}[|\mathcal{R}^2|\vee 1]}}_{\mathrm{mFDR}(\mathcal{R}_2)\leq\alpha}\cdot\frac{\mathbb{E}[|\mathcal{R}^2|\vee 1]}{\mathbb{E}[|\mathcal{R}^1|\vee 1] + \mathbb{E}[|\mathcal{R}^2|\vee 1]} \leq \alpha,\end{aligned}$$

where $\mathcal{H}_0^1 = \mathcal{M}_1 \cap \mathcal{R}^1$ and $\mathcal{H}_0^2 = \mathcal{M}_2 \cap \mathcal{R}^2$. Hence, compared with FDR, mFDR demonstrates robustness to splitting and merging, which is widely adopted in handling sophisticated dependence structure (Wasserman and Roeder, 2009; Dai et al., 2022a; Gang et al., 2023a).

2.1 General Framework for Multiple Hypotheses Testing

In this subsection, we present a general framework to solve the multiple hypotheses testing problem. Given summary statistics $(X_i)_{i\in\mathcal{M}}$ in correspondence with the null hypotheses $(H_{0,i})_{i\in\mathcal{M}}$, then the statistical decision framework can be summarized into the following three-fold procedure:

Step 1 **(Ranking)** Generate the ranking statistics $T_i = c_i(X_i)$ for each $i \in \mathcal{M}$, where $c_i : \mathbb{R} \mapsto \mathbb{R}$ denotes the transformation function which can be either pre-determined or data-dependent.

Step 2 **(Estimation)** Utilize a pre-determined estimation function $\widehat{\mathrm{FDP}} : \mathbb{R}\times\mathbb{R}^m \mapsto \mathbb{R}$ to estimate the $\mathrm{FDP}(\mathcal{R}_t)$ by $\widehat{\mathrm{FDP}}(t)$ for $(T_i)_{i\in\mathcal{M}}$, where $\mathcal{R}_t = \{T_i \le t : i \in \mathcal{M}\}$.

Step 3 **(Thresholding)** Given the designated FDR control level $\alpha \in (0,1)$, choose the maximal threshold of ranking statistics with estimated FDP controlled at level α:

$$t_\alpha = \sup\{t \in \mathcal{T} : \widehat{\mathrm{FDP}}(t) \le \alpha\},$$

within candidate threshold set $\mathcal{T}$ and output rejection set as $\mathcal{R}_{t_\alpha} = \{T_i \le t_\alpha : i \in \mathcal{M}\}$.

We remark that almost all existing testing methods can be attributed to the procedure above. Note that the cornerstone of a testing procedure lies in the construction of *ranking statistics* and *FDP estimators*, which are meticulously designed to maximize TDR while achieving FDR control.

3 Conventional Methodologies

Multiple testing stands out as a valuable approach for extracting meaningful insights from extensive datasets. It serves as a powerful tool for identifying significant features among multiple candidates. In genetic research, it facilitates the detection of regulatory relationships based on gene expression level data (Tusher et al., 2001; Nyholt, 2004; Sun and Wei, 2011); in finance, it helps in making trading strategies or financial asset allocation (Harvey and Liu, 2015; Wang and Ramdas, 2022); in astronomy, it's applied to capture astronomical features contained in abundant data (Miller et al., 2001); in the realm of data visualization, it serves to capture potentially interesting structure (Zhao et al., 2017). The widespread adoption of multiple testing can be attributed to its effectiveness in feature detection under a robust statistical framework with error rate control. In this section, we discuss the conventional methodologies to provide an overview of celebrated solutions, where summary statistics $(X_i)_{i\in\mathcal{M}}$ are assumed to be i.i.d generated from homogeneous pools. For detailed discussions on heterogeneous scenarios, refer to §4, and for dependent scenarios, refer to §5.

3.1 Benjamini-Hochberg (BH) procedure

Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) stands as one of the most celebrated multiple testing procedures in the modern era. Given p-value as summary statistics for each hypotheses $i \in \mathcal{M}$, denoted by $(P_i)_{i\in\mathcal{M}}$, where the null p-values are i.i.d uniform on $[0,1]$, i.e., $P_i \sim \mathrm{Unif}[0,1]$ for all $i \in \mathcal{H}_0$. From the classical (frequentist) view, we suppose that the set of true null hypotheses $\mathcal{H}_0$ is fixed, and the expectation of FDR is taken with respect to the randomness induced by $(P_i)_{i\in\mathcal{H}_0}$. Specifically, under the decision framework in §2.1, BH procedure follows

$$\textbf{BH:}\quad T_i = P_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{mt}{\sum_{i=1}^m \mathbb{1}(P_i \le t)}. \tag{3.1}$$

The BH procedure is intuitively straightforward in that it ranks hypotheses in the order of p-values, and rejects those with small p-values, i.e., stronger statistical evidence against null. The following theorem shows that BH procedure ensures a finite-sample FDR control at designated level $\alpha \in (0,1)$.

Theorem 3.1. The BH procedure controls FDR at $\alpha|\mathcal{H}_0|/m$ in finite sample.

Proof of Theorem 3.1 (Martingale). Denote $F(t) = |\mathcal{R}_t \cap \mathcal{H}_0|$ and $R(t) = |\mathcal{R}_t| \vee 1$. Note that

$$t_\alpha = \sup\{t \in (0,1) : mt \leq \alpha R(t)\}. \tag{3.2}$$

For any potential threshold $t \in [0,1]$, we define the filtration as $\mathcal{F}_t = \sigma\{(\mathbb{1}(P_1 \leq \tau), \ldots, \mathbb{1}(P_m \leq \tau)) : \tau \in [t,1]\}$. Following this, for all $\tau \leq t$, it holds that

$$\mathbb{E}[F(\tau) \mid \mathcal{F}_t] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{1}(P_i \leq \tau) \mid \mathcal{F}_t\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{1}(P_i \leq \tau) \mid \mathbb{1}(P_i \leq t)\right] = \sum_{i \in \mathcal{H}_0} \frac{\tau}{t} \mathbb{1}(P_i \leq t) = \frac{\tau}{t} F(t). \tag{3.3}$$

Notes that (3.3) indicates that $\mathbb{E}[F(\tau)/\tau | \mathcal{F}_t] = F(t)/t$ and thus $t \mapsto F(t)/t$ is a backward martingale. Furthermore, t_α is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t \in [0,1]}$, and the optional stopping theorem (Grimmett and Stirzaker, 2020) gives that $\mathbb{E}[F(t_\alpha)/t_\alpha] = F(1) = |\mathcal{H}_0|$. Thus,

$$\mathrm{FDR}(\mathcal{R}_{t_\alpha}) = \mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1}\right] = \mathbb{E}\left[\frac{F(t_\alpha)}{R(t_\alpha)} \cdot \mathbb{1}(R(t_\alpha) > 0)\right] \leq \frac{\alpha}{m} \cdot \mathbb{E}\left[\frac{F(t_\alpha)}{t_\alpha}\right] = \frac{|\mathcal{H}_0|}{m} \alpha,$$

where the third equation results from the threshold choice $t_\alpha = \sup\{t \in (0,1) : mt \leq \alpha R(t)\}$ and $\mathbb{1}(R(t_\alpha) > 0) \leq 0$, and then we complete the proof via the martingale arguments. □

Proof of Theorem 3.1 (Leave-one-out). Let $\alpha_j = \alpha j/m$. Note that the decision rule of the BH procedure can be written as $\tau^* = \max\left\{\tau \in \mathcal{M} : P_{(\tau)} \leq \frac{\tau\alpha}{m}\right\}$. Following this, the candidate thresholds can be considered discrete such that $\mathcal{T} = \{\alpha_j\}_{j \in \mathcal{M}}$. Thus, it holds that

$$\mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1}\right] = \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \frac{1}{j} \cdot \mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j) \cdot \mathbb{1}\left(P_i \leq \alpha_j\right)\right]. \tag{3.4}$$

Let $\mathcal{R}_{t_\alpha, i \to 0}$ be the rejection set by apply BH procedures on $(P_i)_{i \in \mathcal{M}}$ with P_i substituted by 0. Note that $\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j) \cdot \mathbb{1}\left(P_i \leq \alpha_j\right) = \mathbb{1}(|\mathcal{R}_{t_\alpha, i \to 0}| = j) \cdot \mathbb{1}\left(P_i \leq \alpha_j\right)$ for all $(i,j) \in \mathcal{H}_0 \times \mathcal{M}$. Thus,

$$\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j) \cdot \mathbb{1}\left(P_i \leq \alpha_j\right)\right] = \mathbb{P}(P_i \leq \alpha_j) \cdot \mathbb{P}(|\mathcal{R}_{t_\alpha, i \to 0}| = j) = \frac{\alpha j}{m} \cdot \mathbb{P}(|\mathcal{R}_{t_\alpha, i \to 0}| = j), \tag{3.5}$$

where the first equation results from the independence due to the substitution, and the last equation follows the uniformity of p-values under the null. Combine (3.4) and (3.5), we can get

$$\mathrm{FDR}(\mathcal{R}_{t_\alpha}) = \mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1}\right] = \frac{\alpha}{m} \cdot \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \mathbb{P}(|\mathcal{R}_{t_\alpha, i \to 0}| = j) \leq \frac{|\mathcal{H}_0|}{m} \alpha,$$

and then we complete the proof via the leave-one-out arguments. □

Here, we provide two different proofs of the BH procedure, one from the martingale perspective and the other from the leave-one-out perspective. These perspectives have inspired various follow-up studies, which we will discuss in detail later in this paper, e.g. in §3.2, §4.1, §5.1, and §5.2. We remark that the BH procedure is nearly optimal when the null hypotheses have exchangeable priori, and nearly all true, i.e., $|\mathcal{H}_0| \approx m$. However, in most scenarios, the BH procedure tends to be conservative due to the provable control level at $(|\mathcal{H}_0|/m) \cdot \alpha$ as shown in Theorem 3.1. Additionally, the lack of consideration for the discriminant prior information can result in significant power loss (Lei and Fithian, 2018). The remedies for over-conservativeness are presented in the following subsection, and the latter problem then gives rise to a distinct set of challenges—multiple testing with side information—which we provide a comprehensive discussion in §4.

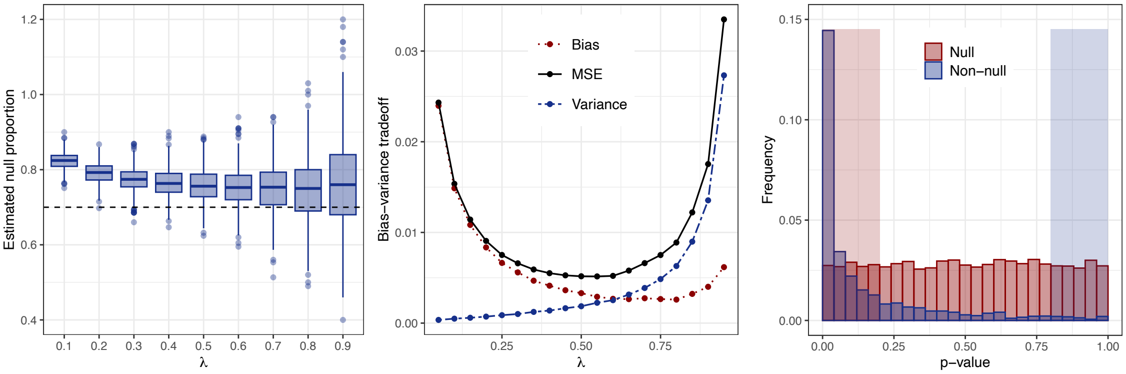

Figure 1: Illustrations of Storey's $\hat{\pi}_0$ and structural information used in BC procedure: distribution of estimation (left), bias-variance tradeoff (middle) with different $\lambda \in (0,1)$, and symmetric structure of p-value under the null, with shaded areas denoting rejection region and false rejection estimation.

3.2 π_0-Adaptive BH Procedure: Remedies for Over-Conservativeness

In this subsection, we first the present remedies for the over-conservativeness issue of BH procedure. When proportion $\pi_0 \equiv |\mathcal{H}_0|/m$ is known, the threshold can be choosed as $t_\alpha = \sup\{t \in \mathbb{R} : \widehat{\mathrm{FDP}}(t) \le \alpha/\pi_0\}$ to close the gap. However, proportion π_0 is usually unknown in practice. Nonetheless, there exists valid estimators $\hat{\pi}_0$ of π_0 such that the BH procedure with thresholding taken at $t_\alpha = \sup\{t \in \mathbb{R} : \widehat{\mathrm{FDP}}(t) \le \alpha/\pi_0\}$ continues to control the FDR under independence. Note that the most celebrated estimator of null proportion is introduced by Storey et al. (2004):

$$\hat{\pi}_0(\lambda) = \frac{1 + \sum_{i=1}^m \mathbb{1}(P_i \ge \lambda)}{m(1-\lambda)}, \quad \forall \lambda \in (0,1), \tag{3.6}$$

and more estimators of null proportion are discussed in Benjamini et al. (2006); Jin and Cai (2007), where Storey's and Benjamini's estimators are tailored for p-values and JC's estimator accommodates more general choices of summary statistics. These procedures are often called the π_0-adaptive BH. We remark that the intuition behind Storey's null proportion estimator $\hat{\pi}_0(\lambda)$ in (3.6) is that:

$$\begin{aligned}\hat{\pi}_0(\lambda) = \frac{1 + \sum_{i=1}^m \mathbb{1}(P_i \ge \lambda)}{m(1-\lambda)} &\ge \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(P_i \ge \lambda)}{|\mathcal{H}_0|\cdot(1-\lambda)} \cdot \frac{|\mathcal{H}_0|}{m} \\ &\approx \pi_0 \cdot \frac{\mathbb{P}(P_i \ge \lambda \mid i \in \mathcal{H}_0)}{1-\lambda} = \pi_0, \quad \forall \lambda \in (0,1),\end{aligned}$$

where "≈" arises from the law of large numbers. Hence, $\hat{\pi}_0(\lambda)$ is a consistently conservative estimator of π_0 for all $\lambda \in (0,1)$, and there is an inherent bias-variance trade-off in the choice of λ as in most cases when λ grows smaller, the bias of $\hat{\pi}_0(\lambda)$ grows larger, but the variance becomes smaller. To choose a proper λ, Storey (2002, 2003) propose a bootstrapping method and a spline-smoothing method, respectively. Langaas et al. (2005) investigate the choice of λ systematically and develop a class of estimators based on nonparametric maximum likelihood estimates (MLEs). With a slight modification in the rejection region, Storey et al. (2004) has also shown that the $\hat{\pi}_0(\lambda)$-adjusted BH procedure can guarantee a finite-sample FDR control at level $(1 - \lambda^{|\mathcal{H}_0|}) \cdot \alpha$ for all $\lambda \in (0,1)$ and please refer to §A.3 for a detailed discussion.

If we choose the threshold λ *adaptively* under the BH procedure by taking $\hat{\pi}_0(1-t)$ as the null proportion estimator for any rejection threshold $t \in (0,1)$, then we have

$$\widehat{\mathrm{FDP}}_{\mathrm{BH}}(t) \le \frac{\alpha}{\hat{\pi}_0(1-t)} \;\Leftrightarrow\; \frac{mt}{\sum_{i=1}^m \mathbb{1}(P_i \le t)} \le \alpha \cdot \frac{mt}{1 + \sum_{i=1}^m \mathbb{1}(P_i \ge 1-t)},$$

which is precisely the Barber-Candès procedure (Barber and Candès, 2015, 2019) by taking p-values as the summary statistics. Formally, the BC procedure can be summarized as below:

$$\textbf{BC:} \quad T_i = P_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{1 + \sum_{i=1}^m \mathbb{1}(P_i \geq 1-t)}{\sum_{i=1}^m \mathbb{1}(P_i \leq t)}, \tag{3.7}$$

where threshold follows $t_\alpha = \sup\{t \in (0, 0.5] : \widehat{\mathrm{FDP}}(t) \leq \alpha\}$. Intuitively, BC procedure utilizes the symmetry of p-values such that $\frac{1}{|\mathcal{H}_0|}\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq t)$ and $\frac{1}{|\mathcal{H}_0|}\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq 1-t)$ is close due to the law of large number. The following theorem shows that BC procedure ensures a finite-sample FDR control over the designated control level $\alpha \in (0,1)$.

Theorem 3.2. The BC procedure controls FDR at α in finite sample.

Proof of Theorem 3.2. Note that the FDR under the BC procedure is upper bounded by

$$\begin{aligned} \mathrm{FDR}(\mathcal{R}_{t_\alpha}) &\leq \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq 1 - t_\alpha)} \cdot \frac{1 + \sum_{i=1}^m \mathbb{1}(P_i \geq 1 - t_\alpha)}{\sum_{i=1}^m \mathbb{1}(P_i \leq t_\alpha)}\right] \\ &\leq \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq 1 - t_\alpha)}\right] \cdot \alpha, \end{aligned} \tag{3.8}$$

where the first inequality results from $\mathcal{H}_0 \subseteq \mathcal{M}$, and the second inequality follows

$$t_\alpha = \sup\{t \in (0, 0.5] : \widehat{\mathrm{FDP}}(t) \leq \alpha\}.$$

Define $\breve{P}_i = P_i$ if $P_i \leq 0.5$, otherwise $\breve{P}_i = 1 - P_i$. Following this, denote $\breve{\mathcal{P}}_0 = \{\breve{P}_i : i \in \mathcal{H}_0\}$ and let $\breve{P}_{(1)} \leq \cdots \leq \breve{P}_{(m_0)}$ be the order statistics over set $\breve{\mathcal{P}}_0$ with $m_0 = |\mathcal{H}_0|$. Without loss of generality, assume that the first $|\mathcal{H}_0|$ hypotheses are null, i.e., $\mathcal{H}_0 = \{1, \ldots, |\mathcal{H}_0|\}$. Consider the stopping time $J = \max\{j \in \mathcal{H}_0 : \breve{P}_{(j)} \leq t_\alpha\}$, and such J must exist since $t_\alpha \in (0, 0.5]$. Thus, it holds that

$$\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq 1 - t_\alpha)} = \frac{(1 - B_1) + \cdots + (1 - B_J)}{1 + B_1 + \cdots + B_J} = \frac{1 + J}{1 + B_1 + \cdots + B_J} - 1,$$

where let $B_i = \mathbb{1}(P_{(i)} > 0.5)$ for all $i \in \mathcal{H}_0$ and the order of $P_{(i)}$'s is inherited from $\breve{P}_{(i)}$'s ranther than the magnitude of P_i's. Note that $(B_i)_{i\in\mathcal{H}_0}$ are independent Bernoulli random variables following $B_i \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$. By the optional stopping lemma (Barber and Candès, 2015) (see Lemma C.1 for details), it holds that $\mathbb{E}\left[\frac{1+J}{1+B_1+\cdots+B_J}\right] \leq 2$ and then the theorem follows. □

We remark that while the BC procedure guarantees that FDR is no larger than α (see Theorem 3.2), and BH procedure ensures FDR at level $|\mathcal{H}_0|\alpha/m$ (see Theorem 3.1), it does not imply that BC procedure dominates BH procedure (Li and Zhang, 2023). In the situations where the non-null information is less significant, the BC procedure exhibits substantial sub-optimality, primarily due to the substitution of $\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \geq 1-t)$ with $\sum_{i\in\mathcal{M}}\mathbb{1}(P_i \geq 1-t)$. Furthermore, it is noteworthy that both BH and BC procedures achieve finite-sample FDR control. In comparison, Storey's $\hat{\pi}_0$-adaptive BH only ensures FDR control asymptotically (Storey et al., 2004). All these procedures are rooted in p-value ranking and aim to optimize power with a more concise FDP (FDR) estimation.

3.3 Local FDR: A Empirical Bayesian Intepretation

The BH (and BC) procedure is established upon a frequentist view, where $\mathcal{H}_0$ is supposed to be fixed, and expectation of FDR is taken with respect to the randomness induced by $(P_i)_{i\in\mathcal{H}_0}$. In this subsection, we re-examine the multiple testing problem from the Bayesian perspective and assume that the set of null hypotheses $\mathcal{H}_0$ is random with summary statistics $(X_i)_{i\in\mathcal{M}}$ in correspondence. Suppose that $(X_i)_{i\in\mathcal{M}}$ follows a two-group mixture model:

$$\theta_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi), \quad X_i \overset{\text{iid}}{\sim} (1-\theta_i)f_0 + \theta_i f_1, \quad \forall i \in \mathcal{M}, \tag{3.9}$$

where null distribution f_0 under $H_{0,i}$ is specified and non-null distribution f_1 under $H_{1,i}$ is unknown. In correspondence, the definition of FDR in (2.1) is modified with expectation taken over $(X_i)_{i\in\mathcal{M}}$ and $\mathcal{H}_0$ jointly, The most celebrated multiple testing procedure from Bayesian perspective is the SC procedure (Sun and Cai, 2007), which is constructed upon local false discovery rate (Lfdr) (Efron et al., 2001; Sun and McLain, 2012; Efron, 2012). Specifically, the Lfdr statistics is defined as

$$\text{Lfdr}(X_i) := \mathbb{P}(\theta_i = 0|X_i) = \frac{(1-\pi)f_0(X_i)}{(1-\pi)f_0(X_i) + \pi f_1(X_i)}, \tag{3.10}$$

and Sun and Cai (2007) has shown that the Lfdr ranking and thresholding rule is optimal in the sense that it maximizes TDR subject to a controlled FDR. Here, we first consider the *oracle* procedure, where function $\text{Lfdr}_i : \mathbb{R} \mapsto \mathbb{R}$ for all $i \in \mathcal{M}$ is known. The SC procedure follows that

$$\textbf{SC:} \quad T_i = \text{Lfdr}(X_i), \quad \widehat{\text{FDP}}(t) = \frac{\sum_{i=1}^m \text{Lfdr}(X_i)\,\mathbb{1}(\text{Lfdr}(X_i) \le t)}{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t)}. \tag{3.11}$$

We remark that the oracle SC procedure ensures a finite sample FDR control intuitively, as

$$\begin{aligned}
\text{FDR}(\mathcal{R}_{t_\alpha}) &= \mathbb{E}_{(X_i)_{i\in\mathcal{M}}}\left[\mathbb{E}_{\mathcal{H}_0|(X_i)_{i\in\mathcal{M}}}\left[\frac{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha)(1-\theta_i)}{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha)} \cdot \mathbb{1}(R(t_\alpha) > 0)\right]\right] \\
&= \mathbb{E}_{(X_i)_{i\in\mathcal{M}}}\left[\frac{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha) \cdot \mathbb{P}\left(\theta_i = 0|(X_i)_{i\in\mathcal{M}}\right)}{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha)} \cdot \mathbb{1}(R(t_\alpha) > 0)\right] \\
&= \mathbb{E}_{(X_i)_{i\in\mathcal{M}}}\Bigg[\underbrace{\frac{\sum_{i=1}^m \text{Lfdr}(X_i)\,\mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha)}{\sum_{i=1}^m \mathbb{1}(\text{Lfdr}(X_i) \le t_\alpha)}}_{\widehat{\text{FDP}}(t_\alpha) \text{ in } (3.11)} \cdot \mathbb{1}(R(t_\alpha) > 0)\Bigg] \le \alpha,
\end{aligned}$$

where the first equation arises from the Baye's theorem and the last equation results from independence and definition of Lfdr in (3.10). In many applications, the oracle Lfdr function is not known and must be estimated. We first define a weak-consistent Lfdr estimator as below.

Definition 3.3. $\widehat{\text{Lfdr}}(x)$ is a weak-consistent if $\frac{1}{m}\sum_{i=1}^m \left|\widehat{\text{Lfdr}}(X_i) - \text{Lfdr}(X_i)\right| \xrightarrow{p} 0$.

We remark that $\widehat{\text{Lfdr}}(x)$ can be estimated with a two-fold procedure nonparametrically: firstly, a consistent estimation of mixture density $f = (1-\pi)f_0 + \pi f_1$ can be obtained via standard kernel density estimation with a bandwidth chosen by cross validation (Silverman, 2018); secondly, the non-null proportion can be estimated using the frequency-based approach proposed by Jin and Cai (2007); Tony et al. (2011), sharing the form that, for fixed $\gamma \in (0, 1/2)$

$$\hat{\pi}(\gamma) = \left(1 - \frac{1}{m}\sum_{i=1}^n e^{\frac{t^2}{2}}\cos(tX_i)\right)\Bigg|_{t=\sqrt{2\gamma\log m}} = 1 - m^{\gamma-1}\sum_{i=1}^m \cos\left(\sqrt{2\gamma\log m}X_i\right), \tag{3.12}$$

which is near minimax optimal for the Gaussian mixture model, i.e., $f(x) = (1-\pi)\phi(x) + \pi \int \phi(x - \mu)\mathrm{d}H(\mu)$ where H is the mixing distribution. Compared with Storey's $\hat{\pi}_0$, the JC estimator outperforms in its consistency and Storey's $\hat{\pi}_0$ is biased in general (Langaas et al., 2005). However, the JC estimator is constrained by its dependence on the assumption of Gaussian distribution family. The following theorem shows that SC procedure ensures an asymptotic FDR control at level $\alpha \in (0, 1)$.

Theorem 3.4. Given $(X_i)_{i\in\mathcal{M}}$ i.i.d from (3.9), the SC procedure at level $\alpha \in (0, 1)$ with any weak-consistent Lfdr estimator (see Definition 3.3) that controls FDR asymptotically under assumptions: (C1) there exists *continuous* functions D_0 and D_1 such that $D_0(t) = \mathbb{P}(\mathrm{Lfdr}(X) \le t)$ and $D_1(t) = \mathbb{E}[\mathrm{Lfdr}(X)\,\mathbb{1}(\mathrm{Lfdr}(X) \le t)]$ for all $t \in [0, 1]$, and (C2) there exists constant $t_\infty \in (0, 1]$ such that $D_1(t_\infty)/D_0(t_\infty) \le \alpha$ where D_0 and D_1 are defined in (C1).

Proof of Theorem 3.4. See §B.1 for a detailed proof. □

We remark that conditions (C1)-(C2) are standard in multiple testing literature from the Empirical Bayesian perspective (Storey et al., 2004; Sun and Cai, 2007; Cao et al., 2022). Note that (C1) guarantees the continuity of the mixture model and the corresponding decision function, and (C2) ensures the existence of the critical value to asymptotically control the FDR at level α.

Power Distortion in Normalization and Standardization Suppose the practitioner obtains the original summary statistics X_i for each hypothesis from the experiments. For simplicity, assume that X_i is sampled from a normal distribution with $\mathrm{Var}(X_i) = \sigma_i^2$, and aims to test $H_{0,i} : \mu_i = 0$ versus $H_{0,i} : \mu_i \ne 0$. Following convention, it's common practice to use standardized z-values $Z_i = X_i/\sigma_i$ and normalized p-values $P_i = 2\Phi(-|Z_i|)$ for computational simplicity. Thus, in conventional FDR control approaches, the default choice of ranking statistics is the p-value. However, Sun and Cai (2007) and Fu et al. (2022) have respectively shown that normalization and standardization can lead to suboptimality. For illustrations, we provide the numerical results in Figure 2, where the oracle procedures refer to the Lfdr-based methods conditioned on p-values P_i, z-values Z_i and heteroscedastic statistics (X_i, σ_i) respectively. We remark that (i) such power distortion arises from data compression during pre-processing (standardization, normalization), resulting in a significant power loss of p-value and z-value oracle procedures compared with the heteroscedasticity-adjusted one, and (ii) the Lfdr-based procedure outperforms BH procedure since it provides a stricter FDR control and takes into account information from alternative distributions. Please refer to Fu et al. (2022) and §6 in Gang et al. (2023a) for further discussions and comparisons.

4 Multiple Hypotheses Testing with Side Information

Conventional multiple testing procedures implicitly assume that data are collected from repeated or identical experimental conditions, implying that hypotheses are exchangeable. However, in many applications, data are known to be collected from heterogeneous sources. Moreover, relevant domain knowledge, such as carefully constructed auxiliary sequences from the same dataset (Liu, 2014; Cai et al., 2019) and external covariates or prior data from secondary data sources (Fortney et al., 2015; Scott et al., 2015; Ignatiadis et al., 2016; Zhang and Chen, 2022), is often available alongside the primary dataset in many studies. Furthermore, side information can be categorized as follows: (i) continuous variables, including the minor allele frequency, the prevalence of bacterial species in genetics, and sample variance of experiments; (ii) discrete variables, including location in 1D, 2D, or 3D coordinate systems from satellite monitoring and neuroimaging, and order of gene expression data when dealing with RNA-Seq; (iii) categorical variables, including the affiliations or sub-groups.

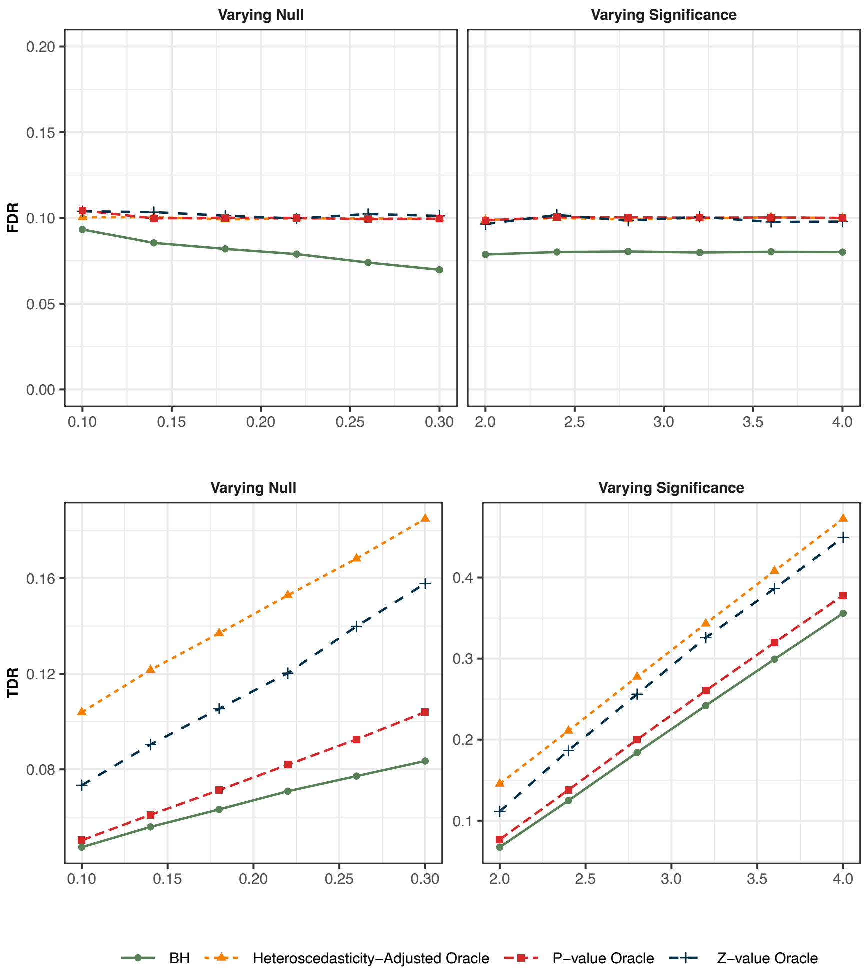

Figure 2: Numerical results of procedures based on different summary statistics: varying null cases (left) follows normal mixture model $X_i \sim (1-\pi)\cdot\mathcal{N}(0,\sigma_i^2)+\pi\cdot\mathcal{N}(2,\sigma_i^2)$ with $\sigma_i \sim \text{Unif}[0.5,4]$ and π from 0.1 to 0.3; varying significance cases (right) follows $X_i \sim 0.8\cdot\mathcal{N}(0,\sigma_i^2)+0.2\cdot\mathcal{N}(\mu,\sigma_i^2)$ with $\sigma_i \sim \text{Unif}[0.5,4]$ and μ from 2 to 4. P_i and Z_i are normalized and standardized in correspondence.

See Figure 3 for an illustration. For instance, consider testing for the association of 400,000 single-nucleotide polymorphisms (SNPs) with each of 40 related diseases. If gene-regulatory relationships are known, then we could expect SNPs near genes to be associated with related diseases.

Following this, extensive research efforts have been devoted to grouped hypotheses testing (Cai and Sun, 2009; Hu et al., 2010), ordered hypotheses testing (Lei and Fithian, 2016; Li and Barber, 2017), and covariate-adjusted hypotheses testing (Ignatiadis et al., 2016; Lei and Fithian, 2018; Zhang and Chen, 2022; Leung and Sun, 2022) to fully exploit the power of statistical inference when external or structural information is provided. Since the grouped and order hypotheses testing can be viewed as special cases of the covariate-adjusted testing, the problem with side information can generally be framed as the covariate-regulated models. For each hypothesis, the practitioner obseves a summary statistics $X_i \in \mathbb{R}$ alongside a covariate $s_i \in \mathcal{S}$ with $\mathcal{S} \subseteq \mathbb{R}^d$. Following the literature, we consider a two-component mixture model:

$$\theta_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_{s_i}) \quad X_i|s_i \overset{\text{iid}}{\sim} (1-\theta_i)f_0 + \theta_i f_{1,s_i}, \quad \forall i \in \mathcal{M}, \tag{4.1}$$

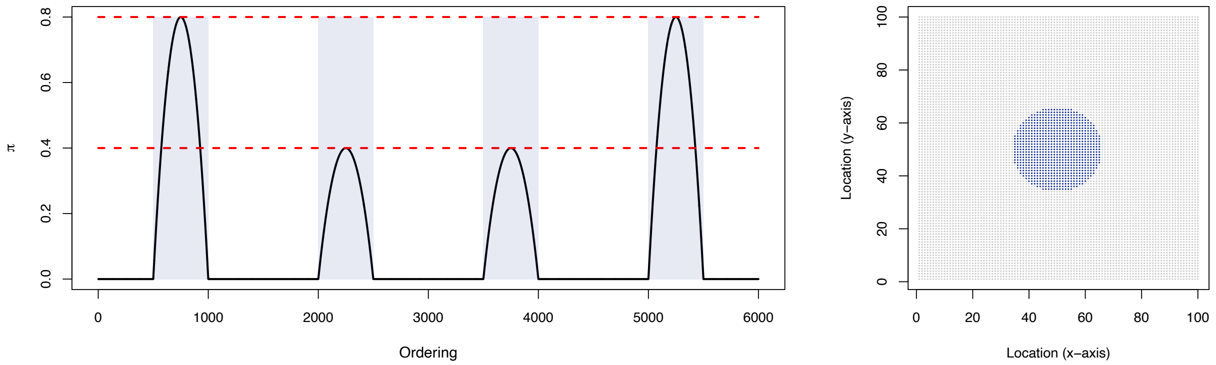

Figure 3: Illustrations of informative side information in multiple testing: ordered hypotheses (left) with shaded area as potentially significant features ($\pi_i > 0$), and two-dimensional clustering pattern (right) on a 100×100 lattice with blue dots as the ground-truth features with $\pi = 0.8$.

where the null distribution f_0 under $H_{0,i}$ is specified and the non-null distribution f_{1,s_i} under $H_{1,i}$ is unknown. Compared to the conventional two-component mixture model in (3.9), the varying null probability π_{s_i} reflects the relative importance of each hypothesis given the covariate information s_i and the varying alternative density f_{1,s_i} emphasizes the heterogeneity among the signals. In this section, we introduce two main approaches towards covariate-adjusted testings: weighted BH (see §4.1) and generalized BC procedure (see §4.2). These approaches are extensions of the standard BH and BC procedures discussed in §3.1 and §3.2, respectively.

4.1 Weighted BH (WBH) Procedure

Weighting is a widely used strategy for incorporating side information into FDR analyses (Benjamini and Hochberg, 1997; Genovese et al., 2006; Basu et al., 2018). Following the literature, let $(\omega_i)_{i\in\mathcal{M}}$ be a set of weights, where the collected p-values $(P_i)_{\in\mathcal{M}}$ is independent of weights $(\omega_i)_{i\in\mathcal{M}}$ conditioned on the set of null hypotheses $\mathcal{H}_0$. The WBH procedure can summarized below:

$$\textbf{WBH:}\quad T_i = P_i/w_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{mt}{\sum_{i=1}^{m} \mathbb{1}(P_i/w_i \le t)}, \tag{4.2}$$

where the WBH procedure is equivalently using P_i/ω_i's as the input of standard BH procedure. The following theorem posits that WBH procedure ensures FDR control under certain regulations.

Theorem 4.1. Suppose that $(P_i)_{\in\mathcal{M}}$ is independent of $(\omega_i)_{i\in\mathcal{M}}$ conditioned on $\mathcal{H}_0$ with $\sum_{i\in\mathcal{H}_0}\omega_i \le m$. The WBH procedure controls FDR at level α in finite sample.

Sketch proof of Theorem 4.1. The proof is akin to that of Theorem 3.1. Let $\alpha_j = \alpha j/m$ and $\mathcal{R}_{t_\alpha, i\to 0}$

be the rejection set by apply BH procedures on $(P_i)_{i\in\mathcal{M}}$ with P_i substituted by 0

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha}\cap\mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}|\vee 1}\right] &= \sum_{i\in\mathcal{H}_0}\sum_{j=1}^{m}\frac{1}{j}\cdot\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}|=j)\cdot\mathbb{1}\left(P_i\le\omega_i\cdot\alpha_j\right)\right]\\ &= \sum_{i\in\mathcal{H}_0}\sum_{j=1}^{m}\mathbb{P}(P_i\le\omega_i\cdot\alpha_j)\cdot\mathbb{P}(|\mathcal{R}_{t_\alpha,i\to 0}|=j)\\ &= \sum_{i\in\mathcal{H}_0}\sum_{j=1}^{m}\frac{\omega_i}{j}\cdot\frac{\alpha j}{m}\cdot\mathbb{P}(|\mathcal{R}_{t_\alpha,i\to 0}|=j)=\frac{\alpha}{m}\sum_{i\in\mathcal{H}_0}\omega_i\le\alpha, \end{aligned}\tag{4.3}$$

where the third equation results from conditional independence and the inequality follow $\sum_{i\in\mathcal{H}_0}\omega_i\le m$. Then, we complete the proof of Theorem 4.1. □

The original condition in Genovese et al. (2006) requires $\sum_{i\in\mathcal{M}}\omega_i\le m$, which is slightly stronger than the one we used. Here, we provide an intuition behind the weight condition. For exchangeable hypotheses, as demonstrated in §3.2, we can use BH procedure at level $m\alpha/|\mathcal{H}_0|$ as a remedy for the over-conservativeness, which is equivalently using P_i/ω_i with weight $\omega_i=m/|\mathcal{H}_0|$. To accommodate the side information into the testing procedure, Li and Barber (2017) proposed to use $\omega_i=\frac{1}{1-\pi(s_i)}$, which is a natural extension to $\omega_i=m/|\mathcal{H}_0|$ under the model in (4.1); and Cai et al. (2022) suggested using $\omega_i=\frac{\pi(s_i)}{1-\pi(s_i)}$ to separate the clustered nonnull p-values more effectively, motivated by the optimality theory in (§4.1, Cai et al., 2019), which we will discuss in detail later. We remark that both weights are valid under regulations, and the main challenge lies in estimating function π_s from data. For instance, Cai et al. (2022) proposed a model-free estimation via screening:

$$\hat{\pi}_\lambda(s)=1-\frac{\sum_{i=1}^m\mathcal{K}_h(s,s_i)\,\mathbb{1}(P_i\ge\lambda)}{(1-\lambda)\cdot\sum_{i=1}^m\mathcal{K}_h(s,s_i)},\quad\forall\lambda\in(0,1),\tag{4.4}$$

where $\mathcal{K}_h:\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}$ is a positive, bounded, and symmetric kernel function with bandwidth h chosen by cross-validation. The estimation in (4.4) can be viewed as a kernelized Storey's $\hat{\pi}_0$ in (3.6), screening over the collected side information $\{s_i\}_{i\in\mathcal{M}}$. We remark that WBH procedure with weights constructed upon covariate-adjusted non-null proportions shows significant power improvement with mild conditions. However, it still has limitations: (i) it neglects the varying alternative density f_{1,s_i} that emphasizes the heterogeneity among the signals; (ii) given the low convergence rate of kernel methods, the estimation of non-null is not that accurate in practical settings, leading to power loss. Fortunately, statisticians have developed the optimal theories and a series of powerful "distribution-free" tools to address these issues, which we will discuss in detail in the next section.

4.2 Generalized BC (GBC) Procedure

Following the covariate-adjusted two-component mixture model in (4.1), Cai et al. (2019) has shown that the optimal procedure for the covariate-adjusted testings is the ranking and thresholding procedure upon the conditional local false discovery rate (Clfdr), defined as

$$\text{Clfdr}_{s_i}(X_i)=\mathbb{P}(\theta_i=0|X_i,s_i)=\frac{(1-\pi_{s_i})f_0(X_i)}{(1-\pi_{s_i})f_0(X_i)+\pi_{s_i}f_{1,s_i}(X_i)}.\tag{4.5}$$

Some astute readers may observe that practitioners could apply SC procedure over estimated Clfdr's to control FDR with covariate adjustment. However, as discussed in §4.1, π_s is hard to estimate, especially in the presence of multi-dimensional side information, since the convergence rate increases exponentially with dimension $d(\mathcal{S})$, let alone estimating the entire Clfdr. To address these issues, Lei and Fithian (2018) has first proposed a two-fold solution. First, they suggested using a model-based method to estimate $\widehat{\text{Clfdr}}_s$, leading to a faster convergence rate. Second, they proposed controlling FDR using a "distribution-free" filter—generalized BC procedure. This approach provides theoretical guarantees regardless of the accuracy of $\widehat{\text{Clfdr}}_s$ estimator. Here, we adopt the GBC procedure proposed by Leung and Sun (2022), which is a natural generalization of the method used in Lei and Fithian (2018); Zhang and Chen (2022), sharing a similar high-level idea. Let $\mathcal{A} : \mathbb{R} \times \mathcal{S} \mapsto \mathbb{R}$ be an assessor function that approximates Clfdr, i.e., $\widehat{\text{Clfdr}}_s(X) = \mathcal{A}(X, s)$, and $c_s(t) = \mathbb{P}(\mathcal{A}(X, s) \leq t | H_0, s)$ be the conditional CDF under the null. Suppose that c_s is continuous and strictly increasing. Following this, the GBC procedure can be summarized as

$$\textbf{GBC:} \quad T_i = \mathcal{A}(X_i, s_i), \quad \widehat{\text{FDP}}(t) = \frac{1 + \sum_{i=1}^m \mathbb{1}(c_{s_i}(\mathcal{A}(X_i, s_i)) \geq 1 - c_{s_i}(t))}{\sum_{i=1}^m \mathbb{1}(\mathcal{A}(X_i, s_i) \leq t)}, \tag{4.6}$$

and the candidate rejection is chosen as $\mathcal{T} = (0, t_{\max}]$ with $t_{\max} = \max\{t : c_{s_i}(t) \leq 0.5 \text{ for all } i\}$. The intuition behind the GBC estimator in (4.6) is straightforward. It constructs covariated-adjusted p-values $c_{s_i}(\mathcal{A}(X_i, s_i))$ using the known f_0. Specifically, the FDR control is realized via

$$\begin{aligned}\text{FDP}(\mathcal{R}_{t_\alpha}) &< \frac{1 + \sum_{i \in \mathcal{H}_0} \mathbb{1}(\mathcal{A}(X_i, s_i) \leq t)}{\sum_{i=1}^m \mathbb{1}(\mathcal{A}(X_i, s_i) \leq t) \vee 1} \\ &= \frac{1 + \sum_{i \in \mathcal{H}_0} \mathbb{1}(c_{s_i}(\mathcal{A}(X_i, s_i)) \leq c_{s_i}(t))}{\sum_{i=1}^m \mathbb{1}(\mathcal{A}(X_i, s_i) \leq t) \vee 1} \\ &\approx \frac{1 + \sum_{i \in \mathcal{H}_0} \mathbb{1}(c_{s_i}(\mathcal{A}(X_i, s_i)) \geq 1 - c_{s_i}(t))}{\sum_{i=1}^m \mathbb{1}(\mathcal{A}(X_i, s_i) \leq t) \vee 1},\end{aligned}$$

where the " $\approx$ " follows from the fact that $c_{s_i}(\mathcal{A}(X_i, s_i))$ is a uniformly distributed random variable over $[0, 1]$ under $H_{0,i}$. The following theorem shows that given pre-determined assessor function $\mathcal{A}$, the GBC procedure ensures a finite-sample FDR guarantee at designated level $\alpha \in (0, 1)$.

Theorem 4.2. Let $\mathcal{A} : \mathbb{R} \times \mathcal{S} \to \mathbb{R}$ be a pre-determined assessor function and $c_s : \mathbb{R} \mapsto \mathbb{R}$ be the null CDF in correspondence. The GBC procedure controls FDR at α in finite sample.

Proof of Theorem 4.2. The proof is akin to that of Theorem 3.2. Please refer to §B.2 for details. □

We remark that the assessor-based GBC procedure is favorable as the FDR guarantee does not depend on the quality of approximation for Clfdr, which is desirable as Clfdr is extremely hard to estimate in practice and only affects the power of the testing procedure. Besides, compared with directly apply BH or BC procedure on covariate-adjusted p-values $c_{s_i}(\mathcal{A}(X_i, s_i))$'s, the GBC procedure outperforms by ranking with $\mathcal{A}(X_i, s_i)$'s, i.e., the assessor of $\text{Clfdr}_{s_i}(X_i)$'s, aligning with the optimal theory that suggests a Clfdr-based procedure. For illustration, we briefly introduce the three-component beta-mixture model used in Leung and Sun (2022), formulated as

$$\Phi(X_i) | s_i \sim \{1 - \pi_l(s_i) - \pi_r(s_i)\} h_0(\cdot) + \pi_l(s_i) h_l(\cdot) + \pi_r(s_i) h_r(\cdot), \quad \forall i \in \mathcal{M},$$

where $\Phi : \mathbb{R} \mapsto [0, 1]$ is a regulation function, the conditional non-null proportions $\pi_l(\cdot)$, $\pi_r(\cdot)$ are characterized by softmax distributions and the density functions $h_l(\cdot)$, $h_r(\cdot)$ follow beta distributions.Specifically, let $\widetilde{s}_i = (1, s_i)$ and the non-null proportion is then modeled as

$$\pi_l(s_i) = \frac{\exp(\theta_l^\top \widetilde{s}_i)}{1 + \exp(\theta_l^\top \widetilde{s}_i) + \exp(\theta_r^\top \widetilde{s}_i)}, \quad \pi_r(s_i) = \frac{\exp(\theta_l^\top \widetilde{s}_i)}{1 + \exp(\theta_r^\top \widetilde{s}_i) + \exp(\theta_r^\top \widetilde{s}_i)},$$

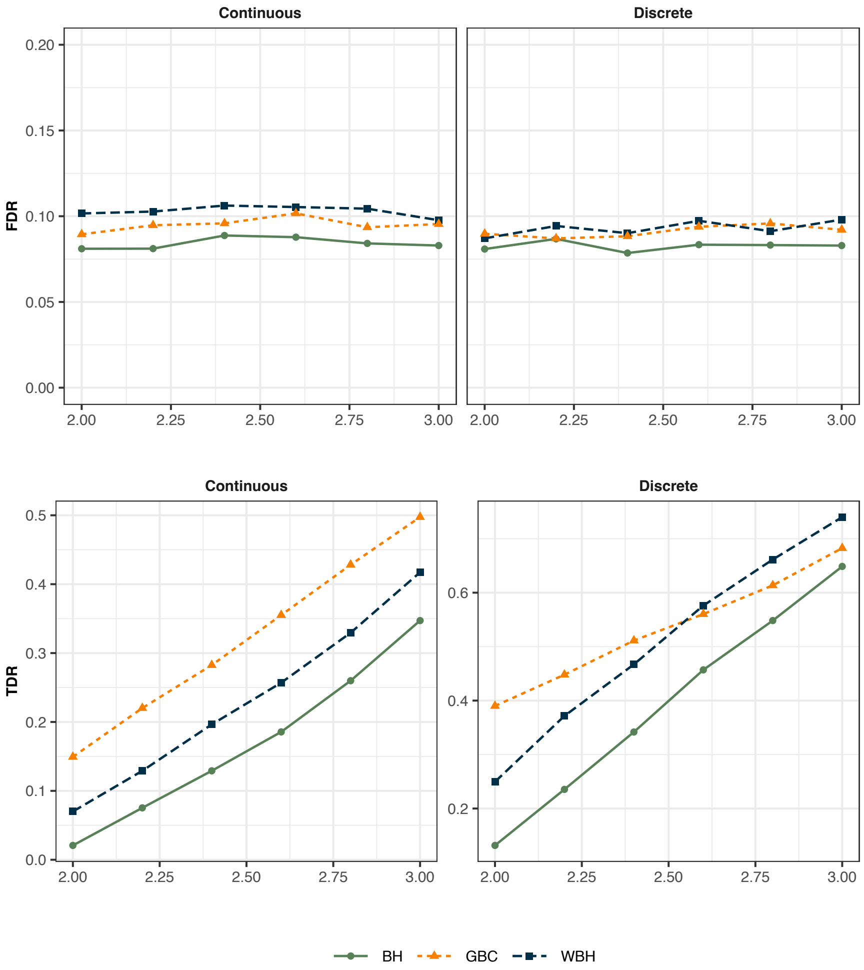

Figure 4: Numerical results of the covariate-adjusted procedures, with `AdaPT` proposed by Lei and Fithian (2018) used as an example of GBC procedures: continous cases (left) follows normal mixture model $Z_i \sim (1-\pi_{s_i})\cdot\mathcal{N}(0,1)+\pi_{s_i}\cdot\mathcal{N}(\mu_{s_i},1)$ with $s_i \sim \text{Unif}[0,1]$, $\pi_{s_i} = s_i^2/2$, and $\mu_{s_i} = s_i\mu$; discrete cases (right) follows $X_i \sim (1-\pi_{s_i})\cdot\mathcal{N}(0,1)+\pi_{s_i}\cdot\mathcal{N}(\mu,1)$ with $s_i = i$ and π_{s_i} is set as the ordering case in Figure 2. Here, μ varies from 2 to 4, and $P_i = 2\Phi(-|Z_i|)$ is normalized in correspondence.

and the conditional density function follows

$$h_l(u) = B(k_l(s_i),\gamma_l)^{-1}u^{k_l(s_i)-1}(1-u)^{\gamma_l}, \quad h_l = r(u) = B(k_r(s_i),\gamma_r)^{-1}u^{k_r(s_i)-1}(1-u)^{\gamma_r},$$

where $k_l(s) = \{1+\exp(-\beta_l^\top\widetilde{s})\}^{-1}$ and $k_r(s) = \{1+\exp(-\beta_r^\top\widetilde{s})\}^{-1}$. Following this, the practitioners can estimate the parameters using EM algorithms. The beta mixture model has long been identified as a flexible modeling tool for variables taking values in the unit interval (Parker and Rothenberg, 1988; Ferrari and Cribari-Neto, 2004; Markitsis and Lai, 2010), and similar models are adopted in Lei and Fithian (2018); Zhang and Chen (2022); Leung and Sun (2022) for covariate-adjusted testings. However, such methodology still has limitations: (i) as noted earlier, the power of testing procedure relies heavily on the choice of assessor model, which is manually decided; (ii) while GBC ensures a finite-sample FDR control with a pre-determined assessor, in practice such an assessor is estimated from data, resulting in degradation to asymptotic control. This is because estimated

$\hat{\mathcal{A}}$ is not independent of $(X_i, s_i)_{i\in\mathcal{M}}$, and Theorem 4.2 no longer holds. To avoid dependence, the data-driven conclusion is achieved by showing that: firstly, by using an oracle assessor, i.e., $\mathcal{A}^*$ that maximizes the risk of the EM algorithms, finite-sample is ensured following Theorem 4.2. Secondly, show that the data-driven $\hat{\mathcal{A}}$ is a consistent estimation of oracle $\mathcal{A}^*$ based on the property of EM estimators (McLachlan and Krishnan, 2007). See Zhang and Chen (2022); Leung and Sun (2022) for detailed proof of data-driven procedures. Moreover, Lei and Fithian (2018); Leung and Sun (2022) have developed an iterative approach with slight modifications to achieve finite-sample control. Very recently, Gang et al. (2023a); Li and Zhang (2023) have proposed generalized BH (GBH) procedures, which share a similar idea introduced in this section. Please see these papers for further discussions.

5 Multiple Hypotheses Testing under Dependence

Observations from extensive testing scenarios often exhibit dependence. However, the traditional FDR methodologies heavily hinge on the independent assumption, often overlooking the correlation among hypotheses. Conventional methods like BH procedure have demonstrated FDR control validity under certain regularity conditions (Benjamini and Yekutieli, 2001; Finner et al., 2009; Ramdas et al., 2017a). However, in various contexts within the field of economics, finance, and genetics, these assumptions may not hold, necessitating diligent verifications by practitioners. Furthermore, such dependencies frequently compromise the statistical accuracy of estimation and testing (Efron, 2007; Schwartzman and Lin, 2011), resulting in heightened variability in outcomes and a potential lack of reproducibility in scientific findings (Owen, 2005; Finner et al., 2009). In this section, we introduce the dependence-robust conditions, methodologies, and statistics respectively in detail.

5.1 Positive Regression Dependence Set (PRDS)

In this subsection, we first introduce the notation of positive regression dependence set (PRDS), one specific type of dependence structure that the BH procedure can remain robust with the removal of the "i.i.d" assumption. The PRDS property is defined below.

Definition 5.1. A set $A \subseteq \mathbb{R}^m$ is said to be increasing if $\mathbf{x} \in A$ implies $\mathbf{y} \in A$ for all $\mathbf{y} \geq \mathbf{x}$. We say $\mathbf{X} = (X_1, \ldots, X_m)$ has a positive regression dependence on the subset $\mathcal{I}_0 \subseteq \mathcal{M}$ (PRDS) if for any $i \in \mathcal{I}_0$ and increasing set $A \subseteq \mathbb{R}^m$, the function $x \mapsto \mathbb{P}(\mathbf{X} \in A \mid X_i \leq x)$ is increasing.

We remark that the original definition proposed in Benjamini and Yekutieli (2001) requires that the function $\mathbb{P}(\mathbf{X} \in A \mid X_i = x)$ is increasing, which is stronger than the condition in Definition 5.1. This version of PRDS first is used by Finner et al. (2009) and please refer to Lemma 1 in Ramdas et al. (2017a) for detailed proof. We present an illustrative example of PRDS scenarios below.

Example 5.2. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be a multivariate gaussian vector with distribution $\mathcal{N}(\mu, \Sigma)$. $\mathbf{X}$ is PRDS on subset $\mathcal{I}_0 \subseteq \mathcal{M}$ if and only if $\Sigma_{ij} > 0$ for all $i \in \mathcal{H}_0$ or $j \in \mathcal{H}_0$.

The argument in Example 5.2 can be easily verified using the conditional distribution of multivariate Gaussian distribution and employing a proof by contraction. In an informal sense, it implies that X_i exhibits a positive correlation with each entry of the random vector if $i \in \mathcal{H}_0$. Following this, we will demonstrate the robustness of the BH procedure with PRDS property.

Theorem 5.3. If p-values $(P_i)_{i\in\mathcal{M}}$ is PRDS on the set of true nulls hypotheses $\mathcal{H}_0$, BH procedure at level $\alpha \in (0, 1)$ controls FDR at $\alpha|\mathcal{H}_0|/m$ in finite sample.

Proof of Theorem 5.3. Write $\alpha_j = \alpha j/m$ and $\beta_{i,j} = \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j \mid P_i \leq \alpha_j)$ for all $(i,j) \in \mathcal{H}_0 \times \mathcal{M}$, and $\beta_{i,m+1} = 0$ for all $i \in \mathcal{H}_0$. Note that $|\mathcal{R}|$ is a decreasing function of the p-values. Thus,

$$\mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j+1 \mid P_i \leq \alpha_j) \geq \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j+1 \mid P_i \leq \alpha_{j+1}), \quad \forall\ (i,j) \in \mathcal{H}_0 \times \mathcal{M}, \tag{5.1}$$

based on the PRDS property in Defintion 5.1. Following (5.1), it holds that

$$\begin{aligned}\beta_{i,j} - \beta_{i,j+1} &= \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j \mid P_i \leq \alpha_j) - \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j+1 \mid P_i \leq \alpha_{j+1}) \\ &\geq \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j \mid P_i \leq \alpha_j) - \mathbb{P}(|\mathcal{R}_{t_\alpha}| \geq j+1 \mid P_i \leq \alpha_j) = \mathbb{P}(|\mathcal{R}_{t_\alpha}| = r \mid P_i \leq \alpha_j). \end{aligned} \tag{5.2}$$

Note that the decision rule of the BH procedure can be written as $\tau^* = \max\left\{\tau \in \mathcal{M} : P_{(\tau)} \leq \frac{\tau\alpha}{m}\right\}$. Following this, the candidate threshold set can be considered discrete such that $\mathcal{T} = \{\alpha_j\}_{j \in \mathcal{M}}$. By combining (5.2) and the arguments above, we can get

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1}\right] &= \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \frac{1}{j} \cdot \mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j) \cdot \mathbb{1}(P_i \leq \alpha_j)\right] \\ &= \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \frac{1}{j} \cdot \mathbb{P}(P_i \leq \alpha_j) \cdot \mathbb{P}(|\mathcal{R}_{t_\alpha}| = j \mid P_i \leq \alpha_j) \\ &\leq \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \frac{\alpha}{m} \cdot \mathbb{P}(|\mathcal{R}_{t_\alpha}| = j \mid P_i \leq \alpha_j) \\ &\leq \sum_{i \in \mathcal{H}_0} \sum_{j=1}^{m} \frac{\alpha}{m} \cdot (\beta_{i,j} - \beta_{i,j+1}) = \sum_{i \in \mathcal{H}_0} \frac{\alpha}{m} \beta_{i,1} = \frac{\alpha|\mathcal{H}_0|}{m},\end{aligned}$$

where the first inequality results from the uniformity of p-values such that $\mathbb{P}(P_i \leq \alpha_j) \leq \alpha_j$, and the last equality arises from $\beta_{i,1} = 1$ based on the rejection rule of the BH procedure. □

For instance, consider testing $H_{0,i} : \mu_i = 0$ versus $H_{0,i} : \mu_i > 0$ for $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma_{i,j} \geq 0$ for all $i, j \in \mathcal{M}$. Following Example 5.2, one-sided p-values $P_i = \Phi(X_i)$ are PRDS on $\mathcal{H}_0$ as $\Phi(x)$ is monotone increasing, thus allowing for direct application of BH procedure. However, for two-sided hypotheses, $P_i = 2\Phi(-|X_i|)$ does not form a co-monotone transformation and p-values may not be PRDS, which remains an open question. There have been few generic models that generate PRDS p-values. Despite the Gaussian model with non-negative correlations, in recent years researchers have shown that recursive order statistics (Loper et al., 2022) and conformal p-values (Bates et al., 2023) are also PRDS. Please refer to §A.4 for a brief introduction to conformal p-values.

5.2 Benjamini-Yekutieli (BY) Procedure

In this subsection, our focus is on FDR control for p-values with arbitrary dependence structures. As demonstrated by Benjamini and Yekutieli (2001), in the most adversarial scenario, the BH procedure needs to pay an additional price of $S_m = \sum_{i=1}^{m} \frac{1}{i} \approx \log m$ for a uniform validity. Following this, we can employ the BH procedure at level α/S_m for arbitrarily dependent p-values, summarized as:

$$\textbf{BY:} \quad T_i = P_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{mt \cdot S_m}{\sum_{i=1}^{m} \mathbb{1}(P_i \leq t)}. \tag{5.3}$$

The following theorem shows that BY procedure ensures a finite-sample FDR control.

Theorem 5.4. The BY procedure at level $\alpha \in (0,1)$ controls FDR at $\alpha|\mathcal{H}_0|/m$ in finite sample.

Proof of Theorem 5.4. Let $\alpha_j = \alpha j/m$. Note that the decision rule of the BH procedure can be written as $\tau^* = \max\left\{\tau \in \mathcal{M} : P_{(\tau)} \leq \frac{\tau\alpha}{m}\right\}$. Following this, with BY adjustment on dependence, the candidate threshold set can be considered discrete such that $\mathcal{T} = \{\alpha_j/S_m\}_{j\in\mathcal{M}}$. Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1}\right] &= \sum_{i\in\mathcal{H}_0}\sum_{j=1}^{m}\frac{1}{j}\cdot\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j)\cdot\mathbb{1}\left(P_i \leq \frac{\alpha_j}{S_m}\right)\right] \\
&= \sum_{i\in\mathcal{H}_0}\sum_{j=1}^{m}\sum_{\ell=1}^{j}\frac{1}{j}\cdot\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j)\cdot\mathbb{1}\left(\frac{\alpha_{\ell-1}}{S_m} \leq P_i \leq \frac{\alpha_\ell}{S_m}\right)\right] \\
&= \sum_{i\in\mathcal{H}_0}\sum_{\ell=1}^{m}\sum_{j=\ell}^{m}\frac{1}{j}\cdot\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| = j)\cdot\mathbb{1}\left(\frac{\alpha_{\ell-1}}{S_m} \leq P_i \leq \frac{\alpha_\ell}{S_m}\right)\right] \\
&\leq \sum_{i\in\mathcal{H}_0}\sum_{\ell=1}^{m}\frac{1}{\ell}\cdot\mathbb{E}\left[\mathbb{1}(|\mathcal{R}_{t_\alpha}| \geq \ell)\cdot\mathbb{1}\left(\frac{\alpha_{\ell-1}}{S_m} \leq P_i \leq \frac{\alpha_\ell}{S_m}\right)\right] \\
&\leq \sum_{i\in\mathcal{H}_0}\sum_{\ell=1}^{m}\frac{1}{\ell}\cdot\mathbb{P}\left(\frac{\alpha_{\ell-1}}{S_m} \leq P_i \leq \frac{\alpha_\ell}{S_m}\right) = \frac{|\mathcal{H}_0|}{m}\alpha,
\end{aligned}$$

where the third inequality results from the exchange of summation order and the last equation follows that $\mathbb{P}\left(\frac{\alpha_{\ell-1}}{S_m} \leq P_i \leq \frac{\alpha_\ell}{S_m}\right) = \alpha/nS_m$ due to the uniformity of p-values under the null. □

While the BY procedure manages to ensure FDR control under arbitrary dependence, such adjustments are over-conservative, particularly in large-scale testing, and often unnecessary in practice. Studies by Owen (2005); Finner et al. (2009) have revealed that high correlation leads to increased variability in testing outcomes, resulting in irreproducibility of scientific findings. To address this, Leek and Storey (2008); Friguet et al. (2009) investigated multiple testing under factor models and demonstrated that subtracting common factors can substantially weaken the dependence structure. Additionally, Efron (2007); Fan et al. (2012) discuss methodologies for incorporating the dependence structure to obtain more accurate FDR estimates for a given p-value threshold.

5.3 False Discovery Control with E-Values

In this subsection, we focus on false discovery control with e-values, an approach aimed at addressing dependency by leveraging a more robust statistical property. As a natural counterpart to the popular p-variable in statistical inference, the e-variable exhibits greater robustness to model *misspecification* and, notably, to *dependence* between the summary statistics. Following the definitions in Vovk and Wang (2021)[1], a p-variable is a random variable satisfies that $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$ under the null, a generalization of conventional uniformity. In comparison, an e-variable is a non-negative random variable satisfying $\mathbb{E}[E] \leq 1$ under the null, which is generated from betting scores, likelihood ratios, and stopped supermartingales (Wasserman et al., 2020; Howard et al., 2021; Shafer, 2021). Under the framework of multiple testing, we could further extend the definition of e-value, as demonstrated in Wang and Ramdas (2022); Ren and Barber (2022).

Definition 5.5. We say $(E_i)_{i\in\mathcal{M}}$ is a set of generalized e-values if $\sum_{i\in\mathcal{H}_0}\mathbb{E}[E_i] \leq m$.

[1] Realized values of e-variables or p-variables are called e-values or p-values, respectively.

Note that for any valid e-value E, $1/E$ is a p-value since $\mathbb{P}(1/E \leq \alpha) \leq \alpha$ based on the Markov's inequality with $\mathbb{E}[E] \leq 1$ under the null. Following this, we introduce the e-BH procedure (Wang and Ramdas, 2022) that substitutes P_i with $1/E_i$ in standard BH procedure and is summarized as:

$$\textbf{e-BH:} \quad T_i = 1/E_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{mt}{\sum_{i=1}^m \mathbb{1}(1/E_i \leq t)}. \tag{5.4}$$

The following theorem demonstrates that the e-BH procedure provides finite-sample FDR guarantee for any generalized e-values (see Definition 5.5), irrespective of the dependence structure.

Theorem 5.6. The e-BH procedure at level $\alpha \in (0,1)$ controls FDR at α in finite sample.

Proof of Theorem 5.6. Note that the decision rule of the e-BH procedure rejects all hypotheses with e-value satisfying that $E_i \geq m/\alpha|\mathcal{R}_{t_\alpha}|$. Following this, it holds that

$$\frac{|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0|}{|\mathcal{R}_{t_\alpha}| \vee 1} = \sum_{i\in\mathcal{H}_0} \frac{\mathbb{1}(E_i \geq m/\alpha|\mathcal{R}_{t_\alpha}|)}{|\mathcal{R}_{t_\alpha}| \vee 1} \leq \sum_{i\in\mathcal{H}_0} \frac{E_i}{|\mathcal{R}_{t_\alpha}| \vee 1} \cdot \frac{\alpha|\mathcal{R}_{t_\alpha}|}{m} = \sum_{i\in\mathcal{H}_0} \frac{\alpha E_i}{m} \cdot \mathbb{1}(|\mathcal{R}_{t_\alpha}| > 0),$$

where the inequality arises from the fact that $\mathbb{1}(E \geq t) \leq E/t$ holds for all non-negative random variable E and $t \in \mathbb{R}^+$. Thus, we have

$$\mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right] \leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{\alpha E_i}{m} \cdot \mathbb{1}(|\mathcal{R}| > 0)\right] \leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{\alpha E_i}{m}\right] \leq \alpha,$$

where the last inequality results from the definition of generalized e-values in Definition 5.5. □

We remark that the most surprising property is that *the e-BH procedure controls FDR at level α even under unknown arbitrary dependence between the e-values.* Besides, there are several reasons to work with e-values: they emerge naturally in the sequential settings; we know methodologies for e-value construction in scenarios where p-value construction is challenging; e-values exhibit greater robustness to misspecification or uncertain asymptotics in high-dimensional settings; and e-values are more amenable to evidence aggregation (Vovk and Wang, 2021; Wang and Ramdas, 2022).

Very recently, a series of studies have been conducted to exploit the power of e-values in the domain of multiple testing. Ren and Barber (2022); Bashari et al. (2023); Banerjee et al. (2023); Li and Zhang (2023) have demonstrated that e-values can be constructed to aggregate testing results for derandomization or conduct meta-analysis, e.g., $E_i = m\,\mathbb{1}(i \in \mathcal{R})/\alpha|\mathcal{R}|$. Besides, Xu and Ramdas (2023) has shown that for any arbitrary e-values, e-BH can be further enhanced via randomization to improve the power of testing, e.g., $E_i' = E_i/U$ with $U \sim \mathrm{Unif}[0,1]$. The field of e-values is rapidly evolving, and readers are encouraged to consult the latest literature for detailed discussions.

6 Discussions and Other Topics

This article explores a set of newly devised techniques for multiple testing. However, it's essential to note that this overview is selective and does not encompass the entire spectrum of recent advancements in this domain. Due to space constraints, we must omit discussions on various related testing challenges in high-dimensional inference and regression models. Specifically, FDR control provides a powerful regularization method for estimation of sparse vectors (Abramovich et al., 2005), large covariance matrices (Bailey et al., 2019), Gaussian graphical models (Liu, 2013), and covariance structure (Cai, 2017) in high-dimensional settings. Besides, simultaneous inference and variable selection for high-dimensional regression model has also received much recent attention (van de Geer et al., 2013; Barber and Candès, 2015, 2019; Dai et al., 2022a; Xing et al., 2023). Additionally, a series of recent papers have pioneered the application of multiple testing schemes in machine learning problems such as multi-label classification (Angelopoulos et al., 2021; Marandon et al., 2022).

Traditionally, multiple testing is conducted offline, where all summary observations are received at once, and all decisions must be made simultaneously. However, it is often at odds with modern data-driven decision-making processes requiring sequential decision-making. This has led to the development of online multiple testing procedures, which have been extensively studied in recent years (Foster and Stine, 2008; Aharoni and Rosset, 2014; Ramdas et al., 2017b, 2018; Gang et al., 2023b). Motivated by applications in mediation and replicability analyses, another natural variant is to consider the scenario where a multivariate summary statistics is provided. The problem can framed as testing the joint significance (JS) with composite null hypotheses or partial conjunction null hypothesis (PCH), and its multiplicity adjustment is a topic of heated discussion (Dai et al., 2022b; Wang et al., 2022; Dickhaus et al., 2021; Deng et al., 2023).

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2005). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34:584–653.

Aharoni, E. and Rosset, S. (2014). Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(4):771–794.

Andreou, E. and Ghysels, E. (2006). Monitoring disruptions in financial markets. *Journal of Econometrics*, 135(1-2):77–124.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2021). Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Bailey, N., Pesaran, M. H., and Smith, L. V. (2019). A multiple testing approach to the regularisation of large sample correlation matrices. *Journal of Econometrics*, 208(2):507–534.

Banerjee, T., Gang, B., and He, J. (2023). Harnessing the collective wisdom: Fusion learning using decision sequences from diverse sources. *arXiv preprint arXiv:2308.11026*.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, pages 2055–2085.

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5):2504–2537.

Bashari, M., Epstein, A., Romano, Y., and Sesia, M. (2023). Derandomized novelty detection with fdr control via conformal e-values. *arXiv preprint arXiv:2302.07294*.

Basu, P., Cai, T. T., Das, K., and Sun, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523):1172–1183.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721.

Benjamini, Y. and Hochberg, Y. (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.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.

Benjamini, Y., Krieger, A., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93(3):491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Cai, T., Sun, W., and Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):187–234.

Cai, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application*, 4:423–446.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481.

Cai, T. T., Sun, W., and Xia, Y. (2022). Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 117(539):1370–1383.

Cao, H., Chen, J., and Zhang, X. (2022). Optimal false discovery rate control for large scale multiple testing with auxiliary information. *The Annals of Statistics*, 50(2):807–857.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022a). False discovery rate control via data splitting. *Journal of the American Statistical Association*, pages 1–18.

Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2022b). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213.

Deng, L., He, K., and Zhang, X. (2023). Joint mirror procedure: Controlling false discovery rate for identifying simultaneous signals. *arXiv preprint arXiv:2304.10866*.

Dickhaus, T., Heller, R., and Hoang, A.-T. (2021). Multiple testing of partial conjunction null hypotheses, with application to replicability analysis of high dimensional studies. *arXiv preprint arXiv:2110.06692*.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, pages 93–103.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160.

Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

Finner, H., Dickhaus, T., and Roters, M. (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Annals of Statistics*, 37:596–618.

Fortney, K., Dobriban, E., Garagnani, P., Pirazzini, C., Monti, D., Mari, D., Atzmon, G., Barzilai, N., Franceschi, C., Owen, A. B., et al. (2015). Genome-wide scan informed by age-related disease identifies loci for exceptional human longevity. *PLoS genetics*, 11(12):e1005728.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(2):429–444.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *arXiv: Statistics Theory*.

Fu, L., Gang, B., James, G. M., and Sun, W. (2022). Heteroscedasticity-adjusted ranking and thresholding for large-scale multiple testing. *Journal of the American Statistical Association*, 117(538):1028–1040.

Gang, B., Qin, S., and Xia, Y. (2023a). A unified and optimal multiple testing framework based on rho-values. *arXiv preprint arXiv:2310.17845*.

Gang, B., Sun, W., and Wang, W. (2023b). Structure–adaptive sequential testing for online false discovery rate control. *Journal of the American Statistical Association*, 118(541):732–745.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Genovese, C. R. and Wasserman, L. A. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, 32:1035–1061.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12.

Grimmett, G. and Stirzaker, D. (2020). *Probability and random processes*. Oxford university press.

Harvey, C. R. and Liu, Y. (2015). Backtesting. *The Journal of Portfolio Management*, 42(1):13–28.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106.

Hochberg, Y. (1987). Multiple comparison procedures. *Wiley Series in Probability and Statistics*.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences.

Hsu, J. (1996). *Multiple comparisons: theory and methods*. CRC Press.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.

Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(4):555–572.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Lei, L. and Fithian, W. (2016). Power of ordered hypothesis testing. In *International conference on machine learning*, pages 2924–2932. PMLR.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Leung, D. and Sun, W. (2022). Zap: z-value adaptive procedures for false discovery rate control with side information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1886–1946.

Li, A. and Barber, R. F. (2017). Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849.

Li, G. and Zhang, X. (2023). E-values, multiple testing and beyond. *arXiv preprint arXiv:2312.02905*.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Annals of Statistics*, 41:2948–2978.

Liu, W. (2014). Incorporation of sparsity information in large-scale multiple two-sample t tests. *arXiv preprint arXiv:1410.4282*.

Loper, J. H., Lei, L., Fithian, W., and Tansey, W. (2022). Smoothed nested testing on directed acyclic graphs. *Biometrika*, 109(2):457–471.

Lumsdaine, R. L. and Papell, D. H. (1997). Multiple trend breaks and the unit-root hypothesis. *Review of economics and Statistics*, 79(2):212–218.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Machine learning meets false discovery rate. *arXiv preprint arXiv:2208.06685*.

Markitsis, A. and Lai, Y. (2010). A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

Miller, C. J., Genovese, C., Nichol, R. C., Wasserman, L., Connolly, A., Reichart, D., Hopkins, A., Schneider, J., and Moore, A. (2001). Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6):3492.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3):411–426.

Parker, R. and Rothenberg, R. (1988). Identifying important results from multiple statistical tests. *Statistics in medicine*, 7(10):1031–1043.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):381–385.

Ramdas, A., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2017a). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*.

Ramdas, A., Yang, F., Wainwright, M. J., and Jordan, M. I. (2017b). Online control of the false discovery rate with decaying memory. *Advances in neural information processing systems*, 30.

Ramdas, A., Zrnic, T., Wainwright, M., and Jordan, M. (2018). Saffron: an adaptive algorithm for online control of the false discovery rate. In *International conference on machine learning*, pages 4286–4294. PMLR.

Ren, Z. and Barber, R. F. (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

Romano, J. P. and Lehmann, E. (2005). Testing statistical hypotheses.

Sarkar, S. K. (2007). Stepup procedures controlling generalized fwer and generalized fdr. *Annals of Statistics*, 35:2405–2420.

Sarkar, S. K. and Guo, W. (2009). On a generalized false discovery rate. *Annals of Statistics*, 37:1545–1565.

Schwartzman, A. and Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(2):407–431.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *The annals of statistics*, 31(6):2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.

Sun, W. and McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107(498):673–687.

Sun, W. and Wei, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *Journal of the American Statistical Association*, 106(493):73–88.

Tony, C., Jessie Jeng, X., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5):629–662.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

van de Geer, S. A., Buhlmann, P., Ritov, Y., and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202.

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wang, J., Gui, L., Su, W. J., Sabatti, C., and Owen, A. B. (2022). Detecting multiple replicating signals using adaptive filtering procedures. *The Annals of Statistics*, 50(4):1890–1909.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Xing, X., Zhao, Z., and Liu, J. S. (2023). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241.

Xu, Z. and Ramdas, A. (2023). More powerful multiple testing under dependence via randomization. *arXiv preprint arXiv:2305.11126*.

Zeevi, Y., Astashenko, S., and Benjamini, Y. (2020). Ignored evident multiplicity harms replicability–adjusting for it offers a remedy. *arXiv preprint arXiv:2006.11585*.

Zhang, X. and Chen, J. (2022). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association*, 117(537):411–427.

Zhao, Z., De Stefani, L., Zgraggen, E., Binnig, C., Upfal, E., and Kraska, T. (2017). Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 acm international conference on management of data*, pages 527–540.

Appendix for "False Discovery Control in Multiple Testing: A Selective Overview of Theories and Methodologies"

A Supplementary Background

A.1 Multiplicity-Related Error Rates

Recap of Error Rates in Single Testing. When conducting a hypothesis test, there are two kinds of errors: reject a null hypothesis when it is true (Type I error) or fail to reject a null hypothesis when it is false (Type II error). Type I error occurs when a non-existent pattern is identified in the data (false discovery), while Type II error occurs when an actual pattern goes unnoticed (missed discovery). In practice, it is impossible to eliminate the risk of making decision errors. However, the consequences of these errors usually differ, with Type I error considered a more serious mistake. The rates of Type I and Type II errors are defined as probabilities of making these errors. In the classical single hypothesis testing, the goal is to control the Type I error rate at a pre-specified level α while minimizing the Type II error rate, i.e., maximizing the testing power.

Error Rates in Multiple Testing. In the multiple testing setting, it is desirable to assess the overall performance of a testing procedure by combining all decisions. For instance, FDR and TDR, as defined in equations (2.1) and (2.3), respectively, provide one kind of measurement for Type I and Type II errors in some sense. The selection of an error rate in applications hinges on the specific purpose of the inference (Benjamini, 2010). Please refer to Table 1 for a detailed comparison.

Error Rates	Abbr.	Expression
Per-Comparison Error Rate	FCER	$\mathbb{E}[\vert\mathcal{R}\cap\mathcal{H}_0\vert/m]\le\alpha$
Per-Family Error Rate	PFER	$\mathbb{E}[\vert\mathcal{R}\cap\mathcal{H}_0\vert]\le k$
Familywise Error Rate	FWER	$\mathbb{P}(\vert\mathcal{R}\cap\mathcal{H}_0\vert\ge 1)\le\alpha$
k-Familywise Error Rate	k-FWER	$\mathbb{P}(\vert\mathcal{R}\cap\mathcal{H}_0\vert\ge k)\le\alpha$
False Discovery Rate	FDR	$\mathbb{E}[\vert\mathcal{R}\cap\mathcal{H}_0\vert/\vert\mathcal{R}\vert\vee 1]\le\alpha$
k-False Discovery Rate	k-FDR	$\mathbb{E}[(\vert\mathcal{R}\cap\mathcal{H}_0\vert-k)_+/\vert\mathcal{R}\vert\vee 1]\le\alpha$
Marginal FDR	mFDR	$\mathbb{E}[\vert\mathcal{R}\cap\mathcal{H}_0\vert]/\mathbb{E}[\vert\mathcal{R}\vert\vee 1]\le\alpha$
Positive FDR	pFDR	$\mathbb{E}[\vert\mathcal{R}\cap\mathcal{H}_0\vert/\vert\mathcal{R}\vert\vert\vert\mathcal{R}\vert>0]\le\alpha$
False Discovery Exceedance	FDX$_\tau$	$\mathbb{P}(\vert\mathcal{R}\cap\mathcal{H}_0\vert/\vert\mathcal{R}\vert\ge\tau)\le\alpha$
Weighted FDR	wFDR	$\mathbb{E}\big[\sum_{i\in\mathcal{H}_0}\omega_i\,\mathbb{1}(i\in\mathcal{R})/\sum_i\omega_i\,\mathbb{1}(i\in\mathcal{R})\vee 1\big]\le\alpha$

Table 1: List of multiplicity-related error rates for simultaneous and selective inference.

Specifically, PCER and PFER are two unadjusted error measures, measuring the expectation of Type I errors. FWER (*strong control*, see §A.2 for discussion), a well-established concept, is defined as the probability of making at least one Type I error in the family. It provides a solution for various statistical scenarios and purposes, including regulation, policy-making, and scientific reporting. However, the drawback is that FWER control significantly limits the power of statistical methods, making it only advisable to use when *simultaneous inference* is essential. An extension of FWER is the k-FWER (Romano and Lehmann, 2005), characterizing the probability of committing k or more Type I errors within the family. In comparison, controlling FDR (Benjamini and Hochberg, 1995) safeguards against selection effects but lacks simultaneous inference. Due to its effective error measure adjusted on false rejections and the significant power of correspondence methods, FDR has emerged as one of the most widely adopted error rates. Besides, together the with False Coverage Rate (FCR), it can provide advanced testing methods and reliable confidence intervals. This makes it a suitable objective for gene filtration and feature selection. Similar to k-FWER, the extension of k-FDR (Sarkar and Guo, 2009) allows personal assessment of the implications of relaxing the requirements of FDR. Besides, mFDR and pFDR are variations on FDR, serving as important intermediate statistics (see Theorem 2.1). Genovese and Wasserman (2004) introduced FDX as a more robust version of FDR, especially when the FDPs are highly variable. wFDR (Benjamini and Hochberg, 1997) is a promising, yet less-explored approach. With external weights indicating the monetary value, it allows selective inference that can integrate more prior knowledge. Additionally, it serves regulatory purposes, particularly for controlling the selection effects of secondary endpoints within clinical trials. To summarize, the practitioners should match error rates to inference needs.

A.2 Controlling Methodologies for FWER

An extensive review of the FWER and k-FWER methodologies can be respectively found in Shaffer (1995) and Sarkar (2007). In this subsection, we present the two most celebrated methodologies — Bonferonni's method and Holm's procedure. Before delving into the details, we provide clarification on the definition of different control levels of FWER. We say a testing procedure controls FWER *weakly* if it controls the FWER under the *global null*, i.e., $\mathcal{H}_0 = \mathcal{M}$. However, discussions among statisticians primarily revolve around FWER control in the strong sense, i.e., $\mathcal{H}_0 \subseteq \mathcal{M}$. In the following, we first provide some arguments about the relationship between FDR and FWER.

Proposition A.1. (i) $\mathrm{FDR}(\mathcal{R}) = \mathrm{FWER}(\mathcal{R})$ if $\mathcal{H}_0 = \mathcal{M}$; (ii) $\mathrm{FDR}(\mathcal{R}) \leq \mathrm{FWER}(\mathcal{R})$.

Proof of Proposition A.1. Under the global null where $\mathcal{H}_0 = \mathcal{M}$, any rejection is a false rejection. Following this, we have $\mathrm{FDP}(\mathcal{R}) \in \{0, 1\}$. Thus, it holds that

$$\mathrm{FDR}(\mathcal{R}) = \mathbb{P}(\mathrm{FDP}(\mathcal{R}) = 1) = \mathbb{P}(|\mathcal{R} \cap \mathcal{H}_0| \geq 1) = \mathrm{FWER}(\mathcal{R}).$$

Similarly, we have

$$\mathrm{FDR}(\mathcal{R}) = \mathbb{E}[\mathrm{FDP}(\mathcal{R})\, \mathbb{1}(|\mathcal{R} \cap \mathcal{H}_0| \geq 1)] \leq \mathbb{P}(|\mathcal{R} \cap \mathcal{H}_0| \geq 1) = \mathrm{FWER}(\mathcal{R}),$$

where the first equation follows $\mathrm{FDP}(\mathcal{R}) = \mathrm{FDP}(\mathcal{R})\, \mathbb{1}(|\mathcal{R}| \geq 1) = \mathrm{FDP}(\mathcal{R})\, \mathbb{1}(|\mathcal{R} \cap \mathcal{H}_0| \geq 1)$. □

Proposition A.1 yields two noteworthy conclusions: (i) all FDR-controlled methodologies exert a weak control over FWER, and (ii) FDR control is less stringent, meaning that if the metric aligns with objectives, opting to control FDR rather than FWER will offer greater statistical power. Next, we introduce the two most celebrated FWER-controlled procedures.

Bonferonni's Method. Bonferonni's method for multiple hypotheses testing rejects all hypotheses with p-value below threshold $\frac{\alpha}{m}$. The method ensures FWER control in a strong sense, following:

$$\mathbb{P}(|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0| \geq 1) \leq \mathbb{E}|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0| \leq \sum_{i \in \mathcal{H}_0} \mathbb{P}\left(P_i \leq \frac{\alpha}{m}\right) = \frac{|\mathcal{H}_0|}{m}\alpha.$$

We remark that Bonferonni's method is valid for dependent p-values. If we assume independence, then we could employ the Sidak's procedure by choosing threshold as $\alpha_m = 1 - (1-\alpha)^{\frac{1}{m}}$, since

$$\mathbb{P}(|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0| \geq 1) = 1 - \mathbb{P}(|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0| = 0) = 1 - (1-\alpha_m)^{|\mathcal{H}_0|} \leq \alpha.$$

Holm's Procedure. The intuition behind Holm's procedure is that p-values can at least jointly provide information about $|\mathcal{H}_0|$ and thus the procedure can be adjusted on the number of rejections to improve the statistical power. In short, Holm's procedure follows that

$$\textbf{Holm's:} \quad T_i = P_i, \quad t_\alpha = \min\left\{m + 1 - \sum_{i \in \mathcal{M}} \mathbb{1}(P_i \le t) > \frac{\alpha}{t}\right\}, \tag{A.1}$$

and choose rejection set as $\mathcal{R}_{t_\alpha}$. Then, we provide the theoretical guarantee of Holm's procedure.

Theorem A.2. Holm's procedure at level $\alpha \in (0,1)$ controls FWER at α in finite sample.

Proof of Theorem A.2. Suppose that Holm's procedure commits at least one false rejection. Following the threshold choice of Holm's procedure in (A.2), then we have

$$t_\alpha \le \frac{\alpha}{m + 1 - \sum_{j \in \mathcal{M}} \mathbb{1}(P_j \le t_\alpha)} \le \frac{\alpha}{m + 1 - \sum_{j \in \mathcal{M}} \mathbb{1}(P_j \le \min_{i \in \mathcal{H}_0} P_i)} \le \frac{\alpha}{|\mathcal{H}_0|}, \tag{A.2}$$

where the second inequality results from $\min_{i \in \mathcal{H}_0} P_i \le t_\alpha$ due to the existence of the false rejections, and the second inequality arises from

$$\sum_{j \in \mathcal{M}} \mathbb{1}\left(P_j \le \min_{i \in \mathcal{H}_0} P_i\right) = \sum_{j \in \mathcal{H}_1} \mathbb{1}\left(P_j \le \min_{i \in \mathcal{H}_0} P_i\right) + 1 \le m - |\mathcal{H}_0| + 1.$$

Following (A.2) and taking a union bound over the set of null hypotheses $\mathcal{H}_0$, it holds that

$$\mathrm{FWER}(\mathcal{R}_{t_\alpha}) = \mathbb{P}(|\mathcal{R}_{t_\alpha} \cap \mathcal{H}_0| \ge 1) = \mathbb{P}\left(\min_{i \in \mathcal{H}_0} P_i \le t_\alpha\right) \le \mathbb{P}\left(\min_{i \in \mathcal{H}_0} P_i \le \frac{\alpha}{|\mathcal{H}_0|}\right) \le \alpha.$$

□

Holm's procedure does not require independence of the p-values and strictly dominates the Bonferroni procedure. Please refer to Hochberg (1987) for an extensive review of FWER methodologies.

A.3 Storey's Adaptive Procedure

In §3.2, we demonstrated that Storey's $\hat{\pi}_0$ effectively addresses the over-conservativeness of the BH procedure. Here, we make a supplement argument about the Storey's $\hat{\pi}_0(\lambda)$-adaptive procedure. As shown in Storey et al. (2004), finite-sample control guarantee is attainable with slight modification:

$$\textbf{Storey's:} \quad T_i = P_i, \quad \widehat{\mathrm{FDP}}(t) = \frac{\hat{\pi}_0(\lambda) \cdot mt}{\sum_{i=1}^m \mathbb{1}(P_i \le t)}, \tag{A.3}$$

where Storey's $\hat{\pi}_0$ is defined in (3.6) and the slight modification lies in the choice of candidate threshold set such $t_\alpha = \sup\{t \in (0, 1-\lambda] : \widehat{\mathrm{FDP}}(t) \le \alpha\}$. The following theorem demonstrates that Storey's procedure ensures a finite-sample FDR control for all fixed $\lambda \in (0,1)$.

Theorem A.3. Storey's procedure at level $\alpha \in (0,1)$ controls FDR at α in finite-sample.

Proof of Theorem A.3. For $t \in [0,1]$, denote $F(t) = |\mathcal{R}_t \cap \mathcal{H}_0|$. Note that

$$\mathrm{FDR}(\mathcal{R}_{t_\alpha}) = \mathbb{E}\left[\frac{\mathrm{FDP}(t_\alpha)}{\widehat{\mathrm{FDP}}(t_\alpha)} \cdot \widehat{\mathrm{FDP}}(t_\alpha)\right] \le \alpha(1-\lambda) \cdot \mathbb{E}\left[\frac{F(t_\alpha)}{t_\alpha\left(1 + \sum_{i \in \mathcal{M}} \mathbb{1}\left(P_i \ge \lambda\right)\right)}\right], \tag{A.4}$$

where the inequality results from threshold choice such that $\widehat{\mathrm{FDP}}(t_\alpha) \leq \alpha$. Write $S_\lambda = \{\mathbb{1}(P_i \geq \lambda)\}_{i\in\mathcal{M}}$ given $\lambda \in (0,1)$. Furthermore, based on the double expectation theorem, we can get

$$\mathbb{E}\left[\frac{F(t_\alpha)}{t_\alpha\left(1+\sum_{i\in\mathcal{M}}\mathbb{1}\left(P_i \geq \lambda\right)\right)}\right] = \mathbb{E}\left[\frac{1}{1+\sum_{i\in\mathcal{M}}\mathbb{1}\left(P_i \geq \lambda\right)} \cdot \mathbb{E}\left[\frac{F(t_\alpha)}{t_\alpha} \;\middle|\; S_\lambda\right]\right]. \tag{A.5}$$

For any threshold $t \in [0, 1-\lambda]$, we define the filtration as $\mathcal{F}_t = \sigma\{(\mathbb{1}(P_1 \leq \tau), \ldots, \mathbb{1}(P_m \leq \tau)) : \tau \in [t, 1-\lambda]\}$. Following this, for all $\tau \leq t$, it holds that

$$\begin{aligned}\mathbb{E}\left[F(\tau) \mid \mathcal{F}_t, S_\lambda\right] &= \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}(P_i \leq \tau) \mid \mathcal{F}_t, S_\lambda\right] \\ &= \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{1}(P_i \leq \tau) \mid \mathbb{1}(P_i \leq t), \mathbb{1}\left(P_i \geq \lambda\right)\right] = \sum_{i\in\mathcal{H}_0} \frac{\tau}{t}\mathbb{1}(P_i \leq t) = \frac{\tau}{t}F(t). \end{aligned} \tag{A.6}$$

Notes that (A.6) indicates that $\mathbb{E}[F(\tau)/\tau | \mathcal{F}_t] = F(t)/t$ and thus $t \mapsto F(t)/t$ is still a backward martingale *conditioned on* S_λ (compared with the martingale arguments in Theorem 3.1). Furthermore, t_α is a stopping time with respect to the filtration $(\mathcal{F}_t)_{t\in[0,1-\lambda]}$, and the optional stopping theorem (Grimmett and Stirzaker, 2020) gives that $\mathbb{E}[F(t_\alpha)/t_\alpha] = F(1-\lambda)$. Following this, we have

$$\begin{aligned}\mathrm{FDR}(\mathcal{R}_{t_\alpha}) &\leq \frac{\alpha(1-\lambda)}{\lambda} \cdot \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \leq 1-\lambda)}{1+\sum_{i\in\mathcal{H}_0}\mathbb{1}\left(P_i \geq \lambda\right)}\right] \\ &= \frac{\alpha(1-\lambda)}{\lambda} \cdot \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(P_i \leq 1-\lambda)}{1+|\mathcal{H}_0| - \sum_{i\in\mathcal{H}_0}\mathbb{1}\left(P_i \leq 1-\lambda\right)}\right] \leq (1-\lambda^{|\mathcal{H}_0|}) \cdot \alpha,\end{aligned}$$

where the last inequality results from $\sum_{i\in\mathcal{H}_0}\mathbb{1}\left(P_i \leq 1-\lambda\right) \sim \mathrm{Binom}(|\mathcal{H}_0|, 1-\lambda)$. □

We remark that while Storey's procedure effectively mitigates the overconservativeness of the BH procedure under general scenarios, similar to the BC procedure, it does not consistently outperform the BH procedure, particularly in cases of weak feature signals. Additionally, the proof of Theorem A.3 can be seen as a simplified version of the proof for the BC procedure with fixed λ.

A.4 Testing with Conformal P-Values

Following Bates et al. (2023), we discuss the conformal p-values under the outlier detection framework. Suppose that the practitioner observes a dataset $\mathcal{D} = \{X_i\}_{i\in\mathcal{N}}$ that contains n independent and independent and identically distributed points $X_i \in \mathbb{R}^d$ drawn from an unknown distribution $\mathcal{P}_X$, and conformity score function $s : \mathbb{R}^d \mapsto \mathbb{R}$ is decided beforehand to measure how much a new observation conforms to previous data. For instance, a smaller value of $s(X)$ may provide stronger evidence that X is an outlier. Typically, such a conformity score function is trained by partitioning the observations into training and calibration data $\mathcal{D} = \mathcal{D}_{\text{test}} \cup \mathcal{D}_{\text{cal}}$, and train $s = \hat{s}(\mathcal{D}_{\text{test}})$. Given testing data $X_{n+1}, \ldots, X_{n+m}$, the objective is to test $H_{0,i} : X_{n+1} \sim \mathcal{P}_X$ for all $i \in \{n+1, \ldots, n+m\}$. For any $X \in \mathbb{R}^d$, the conformal p-value is defined as below:

$$u(X; \mathcal{D}) = \frac{1 + |\{i \in \mathcal{N} : s(X_i) \leq s(X)\}|}{n+1}. \tag{A.7}$$

For notational simplicity, we write $P_i = u(X_i)$ for all $i \in \{n+1, \ldots, n+m\}$. Vovk et al. (1999) has shown that for all $X \sim \mathcal{P}_X$, $u(X)$ is a *marginally* super-uniform p-value such that

$$\mathbb{P}\left(u(X; \mathcal{D}) \leq t\right) \leq t, \quad \forall t \in (0,1), \tag{A.8}$$

where the randomness is introduced by X and observations $\mathcal{D}$. Equibalently, $u(X, \mathcal{D})$ is uniformly distributed on $\left\{\frac{1}{n+1}, \ldots, \frac{n}{n+1}, 1\right\}$. Given the definition of marginal p-values in (A.7), it should be intuitive that larger scores in $\mathcal{D}$ make the p-values for all test data simultaneously smaller, and vice-versa. Thus, conformal p-values shall share some kind of positive correlations, and this intuitive idea is formalized by the following result proving conformal p-values are PRDS.

Theorem A.4. Assume that $u(\cdot)$ is continuously distributed. Then, the conformal p-values $(P_i)_{i\in\mathcal{M}}$ in correspondence to $X_{n+1}, \ldots, X_{n+m}$ are PRDS on the set of true nulls hypotheses $\mathcal{H}_0$.

Proof of Theorem A.4. Let $S_i = s(X_i)$ be the conformality scores for all $i \in [n+m]$ and $\mathbf{P} = (P_1, \ldots, P_m)$ be the conformal p-values evaluated on observations $\mathcal{D}$. Furthermore, we define $Z_i = (\mathbf{S}'_{(i)}, \mathbf{S}_{-i})$, where $\mathbf{S}_{-i} = (S_{n+1}, \ldots, S_{n+i-1}, S_{n+i+1}, \ldots, S_{n+m})$ and $\mathbf{S}'_{(i)} = (S'_{(1)}, \ldots, S'_{(n+1)})$ denotes the order statistics of $(S_1, \ldots, S_n, S_{n+i})$. Note that

$$\left\{\left(S_{(1)}, \ldots, S_{(n)}\right) \mid P_i, \mathbf{S}'_{(i)}\right\} = \left\{\left(S_{(1)}, \ldots, S_{(n)}\right) \mid R_i, \mathbf{S}'_{(i)}\right\} \overset{d}{=} \left(S'_{(1)}, \ldots, S'_{(R_i-1)}, S'_{(R_i+1)} \ldots, S'_{(n+1)}\right),$$

where R_i is the rank of S_{n+i} among $(S_1, \ldots, S_n, S_{n+i})$ suggested by P_i. Intuitively, the preceding argument suggests that P_i and Z_i suffice to provide the order statistics $(S_{(1)}, \ldots, S_{(n)})$ of $(S_1, \ldots, S_n)$, implying that P_j is jointly *determined* by P_i, Z_i, and S_{n+j}. Thus, $\mathbf{P} = (P_1, \ldots, P_m)$ is a deterministic function of P_i and Z_i, denoted by $\mathbf{P} = G(P_i, Z_i)$. Given Z_i, for $j \neq i$ it holds that

$$\begin{aligned} G_j(P_i, Z_i) &= \frac{1}{n+1}\left(1 + \sum_{k\neq R_i, k\leq n+1} \mathbb{1}\left(S_{n+j} \geq S'_{(k)}\right)\right) \\ &= \frac{1}{n+1}\left(1 - \mathbb{1}\left(S_{n+j} \geq S'_{(R_i)}\right) + \sum_{k=1}^{n+1} \mathbb{1}\left(S_{n+j} \geq S'_{(k)}\right)\right). \end{aligned}$$

Following this, we have each entry $G_j(P_i, Z_i)$ is increasing in P_i for all $j \neq i$ as R_i is increasing in P_i. If $i \in \mathcal{H}_0$, then $S_{n+i} \mid \mathbf{S}'_{(i)} \sim \text{Unif}(\mathbf{S}'_{(i)})$ and P_i has the same distribution after conditioning. Thus, $P_i \perp \mathbf{S}'_{(i)}$. Furthermore, since $Z_{n+i} \perp Z_{n+j}$ for all $i \neq j$ by assumption, then $P_i \perp \mathbf{S}_{-i}$. Combine the argument above, we have $P_i \perp Z_i$. For any increasing set A, it holds that

$$\begin{aligned} \mathbb{P}(\mathbf{P} \in A \mid P_i = p) &= \mathbb{P}\left(G(p, Z_i) \in A \mid P_i = p\right) \\ &= \mathbb{E}_{Z_i}[\mathbb{P}\left(G(p, z) \in A \mid P_i = p, Z_i = z\right)] = \mathbb{E}_{Z_i}[\mathbb{1}\left(G(p, z) \in A\right)]. \end{aligned}$$

Since A is an increasing set, $G_i(p, z) = p$ is non-decresing and $G_j(p, z)$ for all $j \neq i$ are increasing in p, then $\mathbb{1}\left(G(p, z) \in A\right)$ is increasing in p, which implies the PRDS property. □

We remark that similar to the argument in (A.8), in the proof of Theorem A.4, we treat observations $\mathcal{D}$, equivalently $\mathbf{S}'_{(i)}$ in Z_i, as random variables. Thus, the conformal p-values in (A.7) only conform to the PRDS property *marginally*. Unfortunately, such a guarantee may be insufficient for a practitioner who needs to compute p-values for a large number of test points but is constrained to working with single observations $\mathcal{D}$. Fortunately, it has been shown that calibration-conditional conformal p-values are attainable. Please refer to Section 3 in Bates et al. (2023) for details.

B Omitted Proofs in Main Article

B.1 Proof of Theorem 3.4

Proof of Theorem 3.4: Suppose that threshold t_α of $\widehat{\text{Lfdr}}$-based procedure exists. Then,

$$\text{FDR}(\mathcal{R}_{t_\alpha}) = \mathbb{E}\left[\frac{\text{FDP}(\mathcal{R}_{t_\alpha})}{\widehat{\text{FDP}}(\mathcal{R}_{t_\alpha})} \cdot \widehat{\text{FDP}}(\mathcal{R}_{t_\alpha})\right] \leq \alpha \cdot \mathbb{E}\left[\frac{\text{FDP}(\mathcal{R}_{t_\alpha})}{\widehat{\text{FDP}}(\mathcal{R}_{t_\alpha})}\right], \tag{B.1}$$

where the inequality results from the threshold choice such that $t_\alpha = \max\{t \in [0,1] : \widehat{\text{FDP}}(\mathcal{R}_t) \leq \alpha\}$. Thus, in the following proof we will respectively show that: (i) $\mathbb{E}\left[\frac{\text{FDP}(\mathcal{R}_{t_\alpha})}{\widehat{\text{FDP}}(\mathcal{R}_{t_\alpha})}\right] \leq 1$ and (ii) t_α exists.

Step 1: Note that we can decompose the ratio as

$$\begin{aligned}\frac{\text{FDP}(\mathcal{R}_t)}{\widehat{\text{FDP}}(\mathcal{R}_t)} = &\frac{\sum_{i=1}^m \mathbb{1}\left(\widehat{\text{Lfdr}}(X_i) \leq t\right)(1-\theta_i)}{\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i)} \cdot \frac{\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i)}{\sum_{i=1}^m \text{Lfdr}(X_i)\,\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)} \\ &\cdot \frac{\sum_{i=1}^m \text{Lfdr}(X_i)\,\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)}{\sum_{i=1}^m \mathbb{1}\left(\widehat{\text{Lfdr}}(X_i) \leq t\right)\widehat{\text{Lfdr}}(X_i)}.\end{aligned}$$

Following this, we define

$$\begin{aligned}
D_{m,0}(t) &= \frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right), \quad \widehat{D}_{m,0}(t) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\widehat{\text{Lfdr}}(X_i) \leq t\right), \\
D_{m,1}(t) &= \frac{1}{m}\sum_{i=1}^m \text{Lfdr}(X_i)\,\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right), \quad \widehat{D}_{m,1}(t) = \frac{1}{m}\sum_{i=1}^m \widehat{\text{Lfdr}}(X_i)\,\mathbb{1}\left(\widehat{\text{Lfdr}}(X_i) \leq t\right) \\
D_{m,2}(t) &= \frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i), \quad \widehat{D}_{m,2}(t) = \frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\widehat{\text{Lfdr}}(X_i) \leq t\right)(1-\theta_i)
\end{aligned}$$

Recall that for each $i \in [m]$ the observation follows model in (3.9):

$$\theta_i \sim \text{Bernoulli}(\pi), \quad X_i \sim (1-\theta_i)f_0 + \theta_i f_1,$$

Following this, based on the Baye's theorem, for all $i \in \mathcal{M}$, it holds that

$$\begin{aligned}
\mathbb{E}_{X_i,\theta_i}\left[\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i)\right] &= \mathbb{E}_{X_i}\left[\mathbb{E}_{\theta_i|X_i}\left[\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i)\right]\right] \\
&= \mathbb{E}_{X_i}\left[\mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right) \cdot \mathbb{P}(\theta_i = 0|X_i)\right] = \mathbb{E}_{X_i}\left[\mathbb{1}\left(\text{Lfdr}(X) \leq t\right)\text{Lfdr}(X)\right].
\end{aligned} \tag{B.2}$$

Based on the weak law of large numbers, it implies that

$$\frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)(1-\theta_i) \xrightarrow{p} D_1(t), \quad \frac{1}{m}\sum_{i=1}^m \mathbb{1}\left(\text{Lfdr}(X_i) \leq t\right)\text{Lfdr}(X_i) \xrightarrow{p} D_1(t),$$

where $D_1(t)$ is a continuous function over $[0,1]$. Since $|D_{m,2}-D_{m,1}| \leq |D_{m,2}-D_1|+|D_1-D_{m,1}|$, then $|D_{m,2}-D_{m,1}| \xrightarrow{p} 0$. Now we are going to show that $|\widehat{D}_{m,1}-D_{m,1}| \xrightarrow{p} 0$ and $|\widehat{D}_{m,2}-D_{m,2}| \xrightarrow{p} 0$.

Note that the difference $|\widehat{D}_{m,1} - D_{m,1}|$ can be decomposed as

$$\begin{aligned}
|\widehat{D}_{m,1} - D_{m,1}| &\leq \left| \frac{1}{m} \sum_{i=1}^{m} \widehat{\mathrm{Lfdr}}(X_i)\, \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \frac{1}{m} \sum_{i=1}^{m} \mathrm{Lfdr}(X_i)\, \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) \right| \\
&\quad + \left| \frac{1}{m} \sum_{i=1}^{m} \mathrm{Lfdr}(X_i)\, \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \frac{1}{m} \sum_{i=1}^{m} \mathrm{Lfdr}(X_i)\, \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right| \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \left| \widehat{\mathrm{Lfdr}}(X_i) - \mathrm{Lfdr}(X_i) \right| + \frac{1}{m} \sum_{i=1}^{m} \left| \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right|,
\end{aligned}$$

where the first term converges to 0 in probability directly follows the weak consistency of $\widehat{\mathrm{Lfdr}}(X_i)$. And similarly, $|\widehat{D}_{m,2} - D_{m,2}|$ can be written as

$$\begin{aligned}
|\widehat{D}_{m,2} - D_{m,2}| &= \left| \frac{1}{m} \sum_{i=1}^{m} \left[\mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right] (1 - \theta_i) \right| \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \left| \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right|.
\end{aligned}$$

All we left to show is that $\frac{1}{m} \sum_{i=1}^{m} \left| \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right|$ converges to 0 in probability. Note that for any $\epsilon \in (0,1)$, the term can be bounded by

$$\begin{aligned}
&\frac{1}{m} \sum_{i=1}^{m} \left| \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right) \right| \\
&= \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t, \widehat{\mathrm{Lfdr}}(X_i) > t\right) + \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t, \mathrm{Lfdr}(X_i) > t\right) \\
&= \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(t - \epsilon < \mathrm{Lfdr}(X_i) \leq t, \widehat{\mathrm{Lfdr}}(X_i) > t\right) + \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t - \epsilon, \widehat{\mathrm{Lfdr}}(X_i) > t\right) \right] \\
&\quad + \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(t < \mathrm{Lfdr}(X_i) \leq t + \epsilon, \widehat{\mathrm{Lfdr}}(X_i) \leq t\right) + \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(\mathrm{Lfdr}(X_i) > t + \epsilon, \widehat{\mathrm{Lfdr}}(X_i) \leq t\right) \right] \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(t - \epsilon < \mathrm{Lfdr}(X_i) \leq t + \epsilon\right) + \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left(\left| \mathrm{Lfdr}(X_i) - \widehat{\mathrm{Lfdr}}(X_i) \right| > \epsilon \right) \\
&\leq \left| D_{m,0}(t + \epsilon) - D_{m,0}(t - \epsilon) \right| + \frac{1}{m\epsilon} \sum_{i=1}^{m} \left| \widehat{\mathrm{Lfdr}}(X_i) - \mathrm{Lfdr}(X_i) \right|. \qquad (\mathbb{1}(X \geq t) \leq X/t)
\end{aligned}$$

As $\frac{1}{m} \sum_{i=1}^{m} \left| \widehat{\mathrm{Lfdr}}(X_i) - \mathrm{Lfdr}(X_i) \right| \xrightarrow{p} 0$ based on the weak consistency and ϵ can be arbitrarily small, take limitation at both side and the second term converges to 0 in probability. Then finishes the proof of $|\widehat{D}_{m,1} - D_{m,1}| \xrightarrow{p} 0$ and $|\widehat{D}_{m,2} - D_{m,2}| \xrightarrow{p} 0$. Thus, it holds that

$$\frac{\mathrm{FDP}(\mathcal{R}_t)}{\widehat{\mathrm{FDP}}(\mathcal{R}_t)} = \frac{\sum_{i=1}^{m} \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right)(1 - \theta_i)}{\sum_{i=1}^{m} \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) \widehat{\mathrm{Lfdr}}(X_i)} = \frac{\widehat{D}_{m,2}(t)}{D_{m,2}(t)} \cdot \frac{D_{m,2}(t)}{D_{m,1}(t)} \cdot \frac{D_{m,1}(t)}{\widehat{D}_{m,1}(t)} \xrightarrow{p} 1,$$

for all $t \in [0,1]$ based on the Slusky's theorem.

Step 2: Recall that there exists constant $t_\infty \in (0,1]$ such that $D_1(t_\infty)/D_0(t_\infty) \leq \alpha$. Following similar arguments as above, we can show that $D_{m,1}(t_\infty)/D_{m,0}(t_\infty) \leq \alpha + o(1)$ as

$$\begin{aligned} |\widehat{D}_{m,0} - D_{m,0}| &= \frac{1}{m}\sum_{i=1}^{m}\Big| \mathbb{1}\left(\widehat{\mathrm{Lfdr}}(X_i) \leq t\right) - \mathbb{1}\left(\mathrm{Lfdr}(X_i) \leq t\right)\Big| \\ &\leq \Big| D_{m,0}(t+\epsilon) - D_{m,0}(t-\epsilon)\Big| + \frac{1}{m\epsilon}\sum_{i=1}^{m}\Big|\widehat{\mathrm{Lfdr}}(X_i) - \mathrm{Lfdr}(X_i)\Big|, \end{aligned} \tag{B.3}$$

and thus $|\widehat{D}_{m,0} - D_{m,0}| \xrightarrow{p} 0$. Furthermore, we have

$$\begin{aligned} \left|\frac{D_{m,1}(t_\infty)}{D_{m,0}(t_\infty)} - \frac{\widehat{D}_{m,1}(t_\infty)}{\widehat{D}_{m,0}(t_\infty)}\right| &\leq \frac{|\widehat{D}_{m,0}(t_\infty) - D_{m,0}(t_\infty)| + |D_{m,1}(t_\infty) - \widehat{D}_{m,0}(t_\infty)|}{|D_{m,0}(t_\infty)\widehat{D}_{m,0}(t_\infty)|} \\ &\leq \frac{|\widehat{D}_{m,0}(t_\infty) - D_{m,0}(t_\infty)| + |D_{m,1}(t_\infty) - \widehat{D}_{m,0}(t_\infty)|}{D_{m,0}(t_\infty)(D_{m,0}(t_\infty) - |\widehat{D}_{m,0}(t_\infty) - D_{m,0}(t_\infty)|)}. \end{aligned} \tag{B.4}$$

Since $D_{m,0}(t_\infty) > 0$ as $D_{m,0}(0) = 0$, $t_\infty > 0$ and $D_{m,0}$ is monotone increasing, then $\widehat{D}_{m,1}(t_\infty)/\widehat{D}_{m,0}(t_\infty)$ $\xrightarrow{p} D_{m,1}(t_\infty)/D_{m,0}(t_\infty)$. Recall that $\widehat{\mathrm{FDP}}(t_\infty) = \widehat{D}_{m,1}(t_\infty)/\widehat{D}_{m,0}(t_\infty) \leq \alpha + o(1)$, thus $t_\alpha = \max\{t \in [0,1] : \widehat{\mathrm{FDP}}(\mathcal{R}_t) \leq \alpha\}$ exists in probability. Combine Step 1 and Step 2, then we finish the proof. □

B.2 Proof of Theorem 4.2

Proof of Theorem 4.2. The proof is akin to that of Theorem 3.2 with an explicit construction. For notational simplicity, we write $T_i = \mathcal{A}(X_i, s_i)$, function $c_i := c_{s_i}$ and $S_i = c_i(T_i)$, then we have

$$\begin{aligned} \mathrm{FDR}(\mathcal{R}_{t_\alpha}) &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(T_i \leq t_\alpha)}{\sum_{i=1}^{m}\mathbb{1}(T_i \leq t_\alpha)\vee 1}\right] \\ &\leq \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(T_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha))}\cdot\frac{1 + \sum_{i=1}^{m}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha))}{\sum_{i=1}^{m}\mathbb{1}(T_i \leq t_\alpha)\vee 1}\right] \\ &\leq \alpha\cdot\mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(T_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha))}\right], \end{aligned}$$

where the first inequality is based on $\sum_{i\in\mathcal{H}_0}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha)) \leq \sum_{i=1}^{m}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha))\}$ and the last inequality follows from the choice of t_α. Then, we only need to show

$$\mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}(T_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0}\mathbb{1}(S_i \geq 1 - c_i(t_\alpha))}\right] \leq 1. \tag{B.5}$$

The inequality can be proved using Lemma C.1 through a stopping time argument. Define

$$\breve{T}_i = \begin{cases} c_i^{-1}(S_i) & \text{if } S_i \leq 0.5 \\ c_i^{-1}(1 - S_i) & \text{if } S_i > 0.5\ . \end{cases}$$

Let $\breve{\mathcal{T}} = \{\hat{T}_i : i \in \mathcal{H}_0\}$. Define the order statistics as $\breve{T}_{(1)} \leq \cdots \leq \breve{T}_{(m_0)}$, where $m_0 = |\mathcal{H}_0|$. Without loss of generality, assume that the first $|\mathcal{H}_0|$ hypotheses are null, i.e., $\mathcal{H}_0 = \{1, \ldots, |\mathcal{H}_0|\}$. Consider optional stopping time $J = \max\{j \in \mathcal{H}_0 : \breve{T}_{(j)} \leq t_\alpha\}$, where J must exist since $c_i(\breve{T}_i) \leq 0.5$ for all i

based on the definition, and $t_\alpha \leq t_{\max} = \max\{t : c_i(t) \leq 0.5 \text{ for all i}\}$ such that $c_i(t_\alpha) \leq 0.5$ holds. Following this, we can get

$$\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(T_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0} \mathbb{1}(S_i \geq 1 - c_i(t_\alpha))} = \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(c_i^{-1}(S_i) \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0} \mathbb{1}(c_i^{-1}(1 - S_i) \leq t_\alpha)}. \tag{B.6}$$

Let $B_i = \mathbb{1}(S_{(i)} > 0.5)$ for all $i \in \mathcal{H}_0$ and the order of $S_{(i)}$'s is inherited from $\breve{T}_{(i)}$'s, rather than the magnitude of S_i's. Note that for all $i \in \mathcal{H}_0$, $c_i^{-1}(S_i) \leq t_\alpha$ if and only if $\breve{T}_i \leq t_\alpha$ and $S_i \leq 0.5$, i.e., $i \leq J$ and $B_i = 0$. Similarly, $c_i^{-1}(1 - S_i) \leq t_\alpha$ if and only if $\breve{T}_i \leq t_\alpha$ and $S_i > 0.5$, i.e., $i \leq J$ and $B_i = 1$. Combine the arguments above, (B.6) can be equivalently written as

$$\frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}(T_i \leq t_\alpha)}{1 + \sum_{i\in\mathcal{H}_0} \mathbb{1}(S_i \geq 1 - c_i(t_\alpha))} = \frac{(1 - B_1) + \cdots + (1 - B_J)}{1 + B_1 + \cdots + B_J} = \frac{1 + J}{1 + B_1 + \cdots + B_J} - 1,$$

Since B_i are independent random variables satisfying $B_i \sim \text{Bernoulli}(0.5)$, by the optional stopping lemma in Lemma C.1, we have $\mathbb{E}\left[\frac{1+J}{1+B_1+\cdots+B_J}\right] \leq 2$, which completes the proof. □

C Technical Lemmas

Lemma C.1. (Barber and Candès, 2019, Lemma 1) Suppose that $B_1, \ldots, B_n$ are independent variables with $B_i \sim \text{Bernoulli}(\rho_i)$ for each i where $\min_i \rho_i \geq \rho > 0$. Let J be a stopping time in reverse time in respect to the filtration $\{\mathcal{F}_j\}$ where

$$\mathcal{F}_j = \{B_1 + \cdots + B_j, B_{j+1} + \cdots + B_n\}.$$

Then

$$\mathbb{E}\left[\frac{1 + J}{1 + B_1 + \cdots + B_n}\right] \leq \rho^{-1}.$$