An Empirical Bayes Approach to Controlling the False Discovery Exceedance

Pallavi Basu\textsuperscript{a,b}, Luella Fu\textsuperscript{c}, Alessio Saretto\textsuperscript{d}, and Wenguang Sun\textsuperscript{e}

\textsuperscript{a}Operations Management, Indian School of Business, Hyderabad, India; \textsuperscript{b}Statistics & OR, Tel Aviv University, Tel Aviv, Israel; \textsuperscript{c}San Francisco State University, San Francisco, CA; \textsuperscript{d}Federal Reserve Bank of Dallas, Dallas, TX; \textsuperscript{e}School of Management and Center for Data Science, Zhejiang University, Hangzhou, China

\section*{ABSTRACT}
In large-scale multiple hypothesis testing problems, the false discovery exceedance (FDX) provides a desirable alternative to the widely used false discovery rate (FDR) when the false discovery proportion (FDP) is highly variable. We develop an empirical Bayes approach to control the FDX. We show that, for independent hypotheses from a two-group model and dependent hypotheses from a Gaussian model fulfilling the exchangeability condition, an oracle decision rule based on ranking and thresholding the local false discovery rate (lfdr) is optimal in the sense that the power is maximized subject to the FDX constraint. We propose a data-driven FDX procedure that uses carefully designed computational shortcuts to emulate the oracle rule. We investigate the empirical performance of the proposed method using both simulated and real data and study the merits of FDX control through an application for identifying abnormal stock trading strategies.

\section*{1. Introduction}
\subsection*{1.1. False Discovery Proportion}
Multiple hypothesis testing provides a useful and powerful technique for identifying sparse signals in massive data. To account for the multiplicity in large-scale testing problems, the false discovery rate (FDR; Benjamini and Hochberg 1995) has been widely used, and a plethora of FDR procedures, exemplified by the Benjamini-Hochberg (BH) procedure, have been developed to control the FDR. Control of the FDR guarantees that the expected value of the false discovery proportion (FDP) over repeated trials is below a pre-specified level. However, the simple and elegant FDR notion still allows for large variability in the FDPs. Specifically, the effective control of the FDR at the nominal level $\alpha$ does not imply the effective control of the FDP in a particular experiment. This is concerning since scientific findings are often reported based on experiments performed only a few times or even just once, exacerbating the replicability crisis. The issue can be particularly worrying in applications where the signals are sparse and weak, or where the tests are highly dependent—in such settings classical methods such as BH often exhibit unstable behaviors, leading to highly variable and possibly skewed FDPs across different experiments (Korn et al. 2004; Delattre and Roquain 2015). As an extreme example, consider a multiple testing procedure that produces FDPs of 0.2 in 50\% of the trials and makes no rejection in the other trials (by convention the FDP in these trials is set 0). Suppose that the nominal FDR level is 0.1. Then, this procedure successfully produces an FDR at exactly the nominal level, but it does a poor job in every single trial, where in half the trials, the procedure produces too many false rejections and in the other half, the procedure fails to find any signals. This is clearly an undesirable situation, where more stringent FDP control is needed.

\subsection*{1.2. False Discovery Exceedance}
We propose to directly control the tail probability of the FDP, also known as the false discovery exceedance (FDX). The FDX provides a more appropriate alternative to the FDR in practical situations where the FDPs are highly variable. Concretely, an FDX procedure at level $(\gamma, \alpha)$ aims to keep the probability that the FDP exceeds a tolerable proportion of false rejections, $\gamma$, at less than a pre-specified small number, $\alpha$. For example, an FDX procedure at level (10\%, 5\%) guarantees that the probability of having 10\% or more false discoveries among all rejections is less than 5\%. In contrast with the FDR that has no control of large FDPs in specific trials, the FDX criterion guarantees that a single implementation of an FDX method has at least a 95\% chance of producing no more than 10\% false rejections.

The FDX can be understood via the notion of $k$-familywise error rate ($k$-FWER), which generalizes the FWER by incorporating a tolerable number of false rejections, $k$, into its definition. Both the $k$-FWER and FDX procedures aim to control the probability of an undesirable event to be small, with the key difference being whether the event is characterized by a count (FWER) or a proportion (FDX) of false positives. The merits of controlling the FDX have been discussed in Genovese and Wasserman (2004), Genovese and Wasserman (2006), Guo and Romano (2007), Chi and Tan (2008), Gordon and Salzman (2008), and Delattre and Roquain (2015), among others. A clear advantage for the characterization of the tolerance level via a
proportion is that the FDX methods can easily scale up with the number of tests, as FDR methods do. Various FDX procedures, some of which will be discussed and compared in later sections, have been proposed in the literature (Lehmann and Romano 2005; Chi and Tan 2008; Roquain and Villers 2011; Döhler and Roquain 2020).

### 1.3. Empirical Bayes (EB) FDX Control

We propose an empirical Bayes procedure for FDX control that relies on first ranking the local false discovery rates (lfdr) and second using the Poisson binomial distribution to compute the lfdr cutoff. By contrast, most existing FDX procedures are based on $p$-values. The hypotheses are first ordered according to their respective $p$-values and then a $p$-value cutoff is determined based on the pre-specified FDX level $(\gamma, \alpha)$. The choice of cutoff is the key difference that distinguishes existing $p$-value methods from one another. We prove that the lfdr ranking is optimal (see also, Fu 2018) and that the procedure controls the FDX at the pre-specified level (see also, Basu 2016). Useful techniques are developed such as the Poisson binomial characterization of the false discovery process and efficient computational cutoffs. We demonstrate that the new techniques lead to substantial power gain over conventional $p$-value based methods. The proposed FDX procedure is simple and fast to implement, and is capable of handling millions of tests. We give an overview of the algorithm in the prototype below.

**The proposed EB-FDX procedure: A prototype algorithm**

Step 1. Compute the $\text{lfdr}$, adjusting for empirical null if needed. Sort all the hypotheses by increasing order of the estimated $\text{lfdr}$ statistics.

Step 2. Compute the probability of the undesirable event that the cumulative failure proportion is greater than $\gamma$ using Poisson binomial distribution (PBD); the Bernoulli probabilities in the PBD are the lfdr values as computed in Step 1.

Step 3. Determine the maximum number of rejections via carefully constructed computational shortcuts (Section 3.4) to ensure that the probability of the undesirable event is smaller than or equal to $\alpha$.

### 1.4. Our Contributions

Under the widely used two-group model (Efron et al. 2001), we formulate the FDX control problem as a constrained optimization problem where the goal is to maximize the expected number of true positives subject to the FDX constraint characterized by the pair $(\gamma, \alpha)$. Our work makes several contributions to theory, methodology, and practice.

- We propose a new empirical Bayes approach to FDX control based on $\text{lfdr}$ and illustrate its efficiency gain over existing frequentists $p$-value methods such as Lehmann and Romano (2005) and Guo and Romano (2007).
- We establish an optimality theory for FDX control by showing that the $\text{lfdr}$ ranking is optimal in the sense that the thresholding rule based on $\text{lfdr}$ has the largest power among all FDX procedures that fulfill the FDX constraint. Although there is a large body of work on FDX, the important optimality issue has not previously been well understood.
- We develop an efficient computational algorithm for determining a data-driven cutoff along the $\text{lfdr}$ ranking; this greatly reduces the computational complexity and enables the broad applicability of our method to large-scale problems with millions of tests. We provide theoretical support to justify the shortcut and illustrate that it significantly improves the computational efficiency using numerical examples.
- We demonstrate the strong empirical performance of the proposed method via both simulated and real datasets. The key strengths of our methodology include simplicity, power gain, and fast computation.

### 1.5. Connection to Existing Work

The proposed procedure builds upon and contributes to the FDX literature. The very notion of controlling the exceedance of the FDP was first defined in Genovese and Wasserman (2006). The authors constructed a confidence envelope for FDP by inverting a set of uniformity tests. The approach can handle correlated tests from random fields (Perone Pacifico et al. 2004). The other approach to the FDX problem is through augmentation, which involves expanding the rejection region from FWER procedures (van der Laan, Dudoit, and Pollard 2004; Farcomeni 2009). van der Laan, Birkner, and Hubbard (2005) proposed using bootstrap-based Monte Carlo methods to generate the states of hypotheses conditional on observed data, and showed that the method is more powerful than inversion and augmentation methods. However, all above mentioned methods involve ranking hypotheses using $p$-values or adjusted $p$-values. For earlier works on this topic, see Farcomeni (2008) for a detailed review. A bootstrap-based heuristics is developed in Romano and Wolf (2007), later formally justified by Delattre and Roquain (2015).

A more recent and closely related line of research aims to provide confidence bounds for the FDP with respect to a user-specified (aka post hoc) rejection sets, see, for example, Hemerik, Solari, and Goeman (2019). Some follow-up works to Hemerik, Solari, and Goeman (2019) include Blanchard, Neuvial, and Roquain (2020), Katsevich and Ramdas (2020), and Goeman, Hemerik, and Solari (2021), among others. However, the efficient ranking of the hypotheses is not discussed, and the tightness of the FDP bounds/thresholds remains unknown in these works. This article addresses both issues by developing a new FDX procedure which (a) ranks hypotheses using the $\text{lfdr}$ and (b) determines the $\text{lfdr}$ cutoff using an efficient and powerful data-driven algorithm.

Our work is closely related to the method in Döhler and Roquain (2020), where the Poisson binomial distribution is employed to determine the $p$-value threshold under the frequentist setting. From a theoretical standpoint, the derivation of our FDX procedure resembles the techniques in Heller and Rosset (2021) for optimal multiple testing under an empirical Bayes setting preceded by the thesis works in (Basu 2016) and (Fu 2018) on topics in weighted FDR and FDX control. However, the optimal ranking under the FDX formulation, the
connection between Poisson binomial distribution and \( \lfdr \), and the empirical Bayes approach to FDX control are new to the literature. Finally, we demonstrate the power gain over other FDX procedures in simulations and provide a real-life application for uncovering interesting financial trading strategies.

1.6. Example and Illustrations

We provide in Figure 1 an illustration of the proposed FDX controlling procedure and highlight a few differences between FDR and FDX control. We compare two \( \lfdr \)-based methods; the adaptive \( \lfdr \) procedure, which controls FDR at level 0.1 (Sun and Cai 2007), and the proposed procedure which controls FDX at level 0.1 and 0.05, (i.e., the probability is kept below 0.05 that the FDP will exceed 0.1). The methods are performed for 5000 trials. In each trial, because both methods use the \( \lfdr \), the rankings of the hypotheses from the two methods are identical, but the different false discovery control mechanisms create different cutoffs along the \( \lfdr \) ranking. Figure 1 contrasts the 5000 FDPs resulting from both methods. In contrast with the FDR procedure (left) that only controls the average of the FDPs over many experiments, resulting in roughly half of trials exceeding the target level (black line), the FDX procedure controls the probability of obtaining a high FDP for any trial, resulting in only a small proportion of trials exceeding the target level. Consequently, the FDP distribution in the proposed FDX procedure is shifted to the left compared to that of the FDR procedure. This illustrates one important merit of the FDX criterion: FDX methods can effectively reduce the chance of having high FDPs, which may increase the replicability of scientific discoveries.

1.7. Organization

The rest of the article is organized as follows. Section 2 introduces notation and sets up the problem. Section 3 describes our proposed solutions. Section 4 discusses further practical concerns about the implementation and \( \lfdr \), which give a greater understanding of the proposed methodology. Section 5 provides numerical simulations. Section 6 provides an application to real data. Section 7 concludes with some future propositions. All code for the procedure and experiments can be requested from the first author. Selected proofs and further numerical experiments are described in the supplementary material.

2. Problem Formulation

2.1. Model and Notation

Our analysis is based on the premise of a widely used two-group model first described by Efron et al. (2001). Suppose we are interested in testing \( m \) hypotheses: \( \mathcal{H} = (H^0_i, H^1_i)_{i \in [m]} \), where \( i \) represents the index of a study unit; \( H^0_i : \theta_i = 0 \) and \( H^1_i : \theta_i = 1 \) are respectively the null and alternative hypotheses corresponding to study unit \( i \); \( [m] := \{1, \ldots, m\} \) denotes the index set of all hypotheses; \( \theta_i \in [0, 1] \) is the true state of nature with \( \theta_i = 0 \) representing a true null hypothesis and \( \theta_i = 1 \) a true alternative. We assume that \( \theta_i \) are iid variables, obeying a Bernoulli distribution with success probability \( \pi = \mathbb{P}(\theta_i = 1) \). The observed data are summarized as a vector of \( z \)-values \( Z = (Z_i)_{i \in [m]} \), whose distribution can be described using the following (hierarchical) two-group model:

\[
\theta_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi), \quad Z_i | \theta_i \sim (1 - \theta_i)F_0 + \theta_i F_1, \quad i \in [m],
\]

where \( F_0 \) and \( F_1 \) are the null and alternative distributions, respectively. Let

\[
F = (1 - \pi)F_0 + \pi F_1
\]

be the mixture distribution. Let \( f_0, f_1, \) and \( f \) be the corresponding densities, which are assumed to exist. \( F_0 \) is assumed to be
known and taken as the cumulative distribution function (CDF) of an \( \mathcal{N}(0, 1) \) variable. However, Efron (2004) and Jin and Cai (2007) argued that the empirical null should be employed in practice and provided methods for estimating the empirical null distribution. For the present, we assume that \( \pi, f_0, \) and \( f \) are known. We discuss the estimation of \( \pi, f_0, \) and \( f \) and investigate issues related to the empirical null in Sections 4 and 6.

A multiple testing rule, which involves making \( m \) simultaneous decisions, can be represented by a binary vector \( \delta = (\delta_i)_{i\in[m]} \in \{0, 1\}^m \). The rule is based on observed data and reflects our belief about the unknown \( \theta = (\theta_i)_{i\in[m]} \). The decision \( \delta_i = 1 \) indicates the rejection of the null hypothesis (aka “a statistical discovery”), whereas \( \delta_i = 0 \) indicates failure to reject the null. The data produces \( z \)-values, which can then be converted to a significance index that indicates the strength of evidence against the null. Decisions can be made by thresholding the significance index. The two most widely used significance indices are the (two-sided) \( p \)-value and the local false discovery rate (lFDR), which are respectively defined as

\[
P_i = 2F_0(−|Z_i|),
\]

\[
lFDR(z) = P(\theta_i = 0|Z_i = z) = (1 − \pi)f_0(z)/f(z).
\]

Sun and Cai (2007) showed that \( p \)-value based methods can be uniformly improved by the adaptive \( z \)-value (AZ) procedure that is based on ranking and thresholding the lFDR. In this article, we aim to develop an lFDR-based decision rule that is provably valid for FDX control. Such an approach has been taken by Sun and Cai (2007) and more recently by Basu et al. (2018) and Heller and Rosset (2021). Further, we want to ensure that the methodology is computationally efficient and can be seamlessly used in applications with millions of tests, analogous to BH (Benjamini and Hochberg 1995) and AZ (Sun and Cai 2007) for FDR control.

### 2.2. False Discovery eXceedance (FDX) and Power

Consider a decision rule \( \delta = (\delta_1, \ldots, \delta_m) \in \{0, 1\}^m \). Using the notations in Section 2.1, the FDP can be defined as

\[
\text{FDP} := \frac{\sum_i (1 − \theta_i)\delta_i}{\sum_i \delta_i} \vee 1,
\]

where \( a \lor b = \max(a, b) \). Let \( \gamma \) represent a tolerance level on the FDP and \( \alpha \) a small probability. An FDX procedure at level \( (\gamma, \alpha) \) satisfies

\[
\text{FDX} := P(\text{FDP} > \gamma) \leq \alpha.
\]

The efficiency or power of an FDX procedure is evaluated using the expected number of true positives:

\[
\text{ETP} := \mathbb{E} \left\{ \sum_i \theta_i \delta_i \right\}.
\]

We call an FDX procedure optimal if it maximizes the ETP subject to the constraint (5).

### 3. Oracle Procedure for FDX Control

This section considers an idealized setup where the distributional quantities \( \pi, f_0, \) and \( f \) are known. Using these, we propose an oracle rule that controls the FDX at level \( (\gamma, \alpha) \) (Section 3.1) and then establish its validity and optimality (Section 3.3). The connection to the Poisson binomial distribution (PBD) is drawn in Section 3.2, and computational shortcuts for fast implementation are developed in Section 3.4. The estimation of \( \pi, f_0, \) and \( f \); and other implementation issues are discussed in Section 4.

#### 3.1. Oracle Procedure

Consider a decision rule that rejects \( k \) hypotheses. Denote \( \mathcal{R}_k \) the set of rejected hypotheses. The number of false rejections is given by \( \sum_{i \in \mathcal{R}_k} (1 − \theta_i) \). Our derivation of the FDX procedure involves calculating the following tail probability

\[
P_{\theta|Z}(\text{FDP} > \gamma) = P_{\theta|Z} \left\{ \sum_{i \in \mathcal{R}_k} (1 − \theta_i) > k\gamma \right\}.
\]

We shall see that (7) can be found using the PBD, which generalizes the binomial distribution to the case when each trial has a different probability of success.

Let \( \text{PBD}(k, p) \) be a random variable following the Poisson binomial distribution with \( k \) being the total number of trials and \( p = (p_1 : i = 1, \ldots, k) \) the vector of success probabilities. To reflect that we consider the idealized setup, the notation \( T_{\theta}^{\text{OR}} \) is used to denote the oracle lFDR associated with study unit \( i \):

\[
T_{i}^{\text{OR}} := P(\theta_i = 0|Z_i = z_i) = \frac{(1 − \pi)f_0(z_i)}{f(z_i)}.
\]

**Procedure 1** describes our proposed oracle FDX procedure at level \((\gamma, \alpha)\).

**Procedure 1.**

1. Consider the lFDR statistic \( (T_{i}^{\text{OR}})_{i\in[m]} \) defined in (8), and denote \( (T_{i}^{\text{OR}})_{i\in[m]} \) the ranked statistic in ascending order. The corresponding null hypotheses are denoted \( (H_{i}^{0})_{i\in[m]} \).

2. Denote \( p^{(k)} = \{T_{(1)}^{\text{OR}}, \ldots, T_{(k)}^{\text{OR}}\} \). Let

\[
K := \max \left\{ k : \mathbb{P}(\text{PBD}(k, p^{(k)}) > \gamma k) \leq \alpha \right\},
\]

and reject the top \( K \) hypotheses \( (H_{(1)}^{0}, \ldots, H_{(K)}^{0}) \) along the lFDR ranking.

3. Denote \( \mathcal{R}_K \) the set of rejected hypotheses from Step 2. Reject \( H_{(K+1)}^{0} \) with the following probability

\[
\alpha − P_{\theta|Z} \left\{ \sum_{i \in \mathcal{R}_K} (1 − \theta_i) > \gamma K \right\} \geq P_{\theta|Z} \left\{ \sum_{i \in \mathcal{R}_{K+1}} (1 − \theta_i) > \gamma (K + 1) \right\}.
\]

Similar to Benjamini and Hochberg (1995), **Procedure 1** is a step-up procedure, in the sense that it starts from the least significant hypothesis (i.e., the one with the largest lFDR) and...
moves up at each step to a more significant one. The procedure stops when it finds the first null hypothesis, $H_{0}^{(k)}$, for which the tail probability is less than $\alpha$. It then rejects all null hypotheses $\{H_{i}^{(0)}, \ldots, H_{i}^{(k)}\}$. The randomization of the decision at the last step ensures that we achieve exact FDX control at the nominal level $(\gamma, \alpha)$. This randomization technique was employed in the weighted FDR procedure proposed by Basu et al. (2018) and later in Gu and Koenker (2020).

3.2. Poisson Binomial Distribution (PBD) and its Connection to the lfdr

Poisson's binomial distribution or PBD refers to the sum of independent Bernoulli random variables, with not necessarily equal expectations. In the special case that the expectations are all equal, a PBD simplifies to a binomial distribution. The PBD provides a useful tool for probability calculations in a range of statistical applications (Chen and Liu 1997). More recently, the PBD has been employed by Döhler and Roquain (2020) to develop FDX-controlling procedures for multiple testing with heterogeneous units. This section discusses the connection of the PBD to the lfdr procedure.

Consider the two-group model (1). If we do not have any prior knowledge, then each null hypothesis is true (or false) with the same probability $1 - \pi$ (or $\pi$). However, conditional on the observables $Z = (Z_i)_{i \in [m]}$, the unknown states of the hypotheses $\theta_i$ marginally follow Bernoulli distributions with heterogeneous success probabilities, namely, $\pi_i = \mathbb{P}(\theta_i = 0 | Z)$. In the situation where the joint density $f(Z|\theta_1, \ldots, \theta_m)$ can be factorized into the product of marginal densities. That is,

$$f(Z|\theta_1, \ldots, \theta_m) = \prod_{i \in [m]} f(Z_i|\theta_i),$$

and $\mathbb{P}(\theta_i = 0 | Z)$ reduces to the lfdr statistic $\mathbb{P}(\theta_i = 0 | Z)$. Furthermore, if $\theta_i$ are independent, then the partial sum $\sum_{i \in S} (1 - \theta_i)$, where $S \subset [m]$ is an arbitrary nonempty subset of $[m]$, is a PBD random variable conditional on $Z$.

In the oracle procedure (Procedure 1), we rank hypotheses by the lfdr; hence, the subset $S$, which consists of the hypotheses we reject, is data-dependent. Then, the index set of the partial sum should be denoted by $S_Z$, which critically depends on the thresholding (aka “selection”) step of Procedure 1. An important concern is that this may lead to a selection bias, but the next lemma proves that the selection based on lfdr does not distort the Poisson binomial distribution of the partial sum $\sum_{i \in S_Z} (1 - \theta_i) I_{i \in S_Z}$. Our proof essentially corroborates the general principle that selective inference based on Bayes rule is unbiased conditional on a selection event. The high-level idea is explained in earlier works (e.g., Dawid 1994).

**Lemma 1.** (No Selection Bias) Ranking by the lfdr does not alter the conditional distribution of the partial sums $\sum_{i \in S_Z} (1 - \theta_i) I_{i \in S_Z}$, where $S_Z$ denotes any index set under consideration after viewing the data $Z$.

**Proof.** Define random variables $R_i := (1 - \theta_i) I_{i \in S_Z}$, where $S_Z \subset [m]$ is determined by the observables $Z = (Z_i)_{i \in [m]}$. Consider any nonempty fixed subset of indices denoted by $S_P$. Consider two situations:

- (i) If $S_P \not\subseteq S_Z$, then there exists an index $i_0 \in S_P \cap S_Z$. We have
  $$\mathbb{E} \left\{ \prod_{i \in S_P} R_i | Z \right\} = 0 = \mathbb{E} \left( R_{i_0} | Z \right) \cdot \mathbb{E} \left( \prod_{i \in S_P \setminus \{i_0\}} R_i | Z \right).$$

- (ii) If $S_P \subseteq S_Z$, then we have
  $$\mathbb{E} \left[ \prod_{i \in S_P} R_i | Z \right] = \mathbb{E} \left[ \prod_{i \in S_P} (1 - \theta_i) | Z \right] = \prod_{i \in S_P} \mathbb{E}[(1 - \theta_i) | Z] = \prod_{i \in S_P} \mathbb{E}[R_i | Z].$$

where the middle equality is due to factorization of the joint density and independence of the $\theta_i$s. Therefore, conditional on $Z$, the $R_i$’s are identically zero if $i \not\in S_Z$ and independently distributed as Bernoulli random variables with the expectation of $\text{lfdr}(Z_i)$ if $i \in S_Z$.

3.3. Properties of the Oracle Procedure

The oracle procedure can be split into two steps, ranking and thresholding, which shape the procedure’s properties. For the ranking step, which orders the hypotheses from the most significant to the least significant according to some significant index, a desirable property is optimal ranking. The optimal ranking with respect to a significant index $T$ indicates that the thresholding rule along the $T$ ranking has equal or larger power than any other rule at the same error rate.

For the thresholding step, which chooses a cutoff along the ranking to control the desired error rate, a desirable property is error-rate control at the target level. In this section, assuming that lfdr ranking does not introduce any selection bias (see Lemma 1), we first show that the FDX level $(\gamma, \alpha)$ is exhausted by our proposed oracle procedure (thereby achieving the second property), and then establish that the lfdr ranking is optimal (thereby achieving the first property).

**Proposition 1.** (Exact Controlof FDX) Procedure 1 controls the FDX at level $(\gamma, \alpha)$.

**Proof.** By design, Procedure 1 ensures $\mathbb{P}(\theta | Z) \left\{ \sum_{i \in R_K} (1 - \theta_i) > \gamma K \right\} \leq \alpha$. Consider an independent arbiter $U$ (i.e., a weighted coin flip), which is chosen to favor one more rejection with probability given in (9). This arbitrary randomization at Step 3 in Procedure 1 ensures that

$$\mathbb{E}[\theta(U) | Z] \left[ \sum_{i \in U} (1 - \theta_i) \delta_i > \gamma \sum_{i \in U} \delta_i \right] = \alpha,$$

where $\delta_i$ is the decision rule determined by the data $Z$ and the independent arbiter. Taking a further expectation with respect to $Z$ completes the proof.

**Proposition 2.** (Optimal Ranking) In the iid two-group model, Procedure 1 has the best ranking almost surely in the sense that for any other decision rule at FDX-level $(\gamma, \alpha)$, we can always find an lfdr-based thresholding rule at the same level that has a higher or equal ETP.

The proof of Proposition 2 is presented in the supplementary material.
3.4. Computational Shortcuts

Procedure 1 is a step-up procedure, which starts by computing the tail probability of the PBD for all tests under consideration. At each progressive step, the set of tests under consideration decreases. However, if we have a massive number of tests to begin with, then Procedure 1 can be computationally intensive. To facilitate fast implementation in large-scale testing problems, we develop some computational shortcuts and modify Procedure 1 as follows.

Procedure 2. Consider the Ifdr statistics \( (T_{i}^{OR})_{i \in [m]} \) defined in (8). The modified FDX procedure consists of 4 steps that successively narrow down the focus of the search.

1. Order the Ifdr statistics in ascending order. Denote the ordered statistics by \( (T_{i}^{OR})_{i \in [m]} \).
2. Reject up to \( K_1 := \max \left\{ k \in [m] : \mathbb{E}_{\theta | Z} \left\{ \sum_{i=1}^{k} (1 - \theta_i) \right\} \leq k \cdot (\alpha + \gamma (1 - \alpha)) \right\} \).
3. Reject up to \( K_2 := \max \left\{ k \in [K_1] : \mathbb{P}(Y > \gamma k) \leq \alpha \right\} \), where \( Y \sim \text{Binomial} \left( k, \prod_{i=1}^{k} T_{i}^{1}/k \right) \).
4. Denote \( p^{(k)} = (T_{1}^{OR}, \ldots, T_{k}^{OR}) \). Reject only up to
   \[ K := \max \left\{ k \in [K_2] : \mathbb{P}(\text{PBD}(k, p^{(k)}) > \gamma k) \leq \alpha \right\} \]
5. Randomize the decision to reject the \( K + 1 \) null hypothesis as in Procedure 1.

First, note that relative to Procedure 1, Procedure 2 has two additional steps (i.e., Step 2 and Step 3). We progressively reduce the number of hypotheses under consideration so that the computationally intensive Step 4 can be conducted on a much smaller subset of hypotheses. This greatly reduces the computational cost because Step 2 only involves computing cumulative average Ifdr over progressively smaller sets of hypotheses, and Step 3 involves binomial calculations rather than Poisson binomial calculations over sets of hypotheses which are further whittled down.

Next we justify the computational shortcuts by showing that Procedures 1 and 2 are equivalent. The main idea is that both Steps 2 and 3 are step-up procedures in themselves: a testing unit’s failure to meet their criteria guarantees that the testing unit will also fail the tail probability criteria in Step 4.

Step 2 is equivalent to the adaptive z-value procedure for FDR control in Sun and Cai (2007) at the level \( \alpha + \gamma (1 - \alpha) \). We claim that \( \mathbb{P}_{\theta | Z} \left\{ \sum_{i=1}^{k} (1 - \theta_i) > \gamma k \right\} \leq \alpha \) implies

\[
\mathbb{E}_{\theta | Z} \left\{ \sum_{i=1}^{k} (1 - \theta_i) \right\} \leq k \cdot (\alpha + \gamma (1 - \alpha)).
\]

To see this, rewrite the above expectation by partitioning the event into the sub-event where the number of false positives is greater than \( \gamma k \) and its complement event:

\[
\mathbb{E}_{\theta | Z} \left( \sum_{i=1}^{k} (1 - \theta_i) > \gamma k \right)\sum_{i=1}^{k} (1 - \theta_i),
\]

The sum in the first expectation is bounded by \( k \), and the sum in the second expectation, the number of false positives, is bounded by \( \gamma k \). Assume that there is an \( \alpha’ \), such that \( P_{\theta | Z_i} \left\{ \sum_{i=1}^{k} (1 - \theta_i) > \gamma k \right\} = \alpha’ \), then the sum of the two expectations is bounded by \( k \cdot \alpha’ + \gamma (1 - \alpha’) \). Since \( (k - \gamma k) \geq 0 \) and \( \alpha’ \leq \alpha \), because \( P_{\theta | Z_i} \left\{ \sum_{i=1}^{k} (1 - \theta_i) > \gamma k \right\} \leq \alpha \), then \( k \cdot (\alpha’ + \gamma (1 - \alpha’)) \leq k \cdot (\alpha + \gamma (1 - \alpha)) \), which becomes the upper bound. Thus, if the condition in Step 4 holds, then the condition in Step 2 also holds. It follows that if the condition in Step 2 fails, then the condition in Step 4 also fails. Hence, the reduction of the set of hypotheses produced by Step 2 is legitimate, as it only eliminates cases in which the condition of Step 4 would fail. Step 2 ends when we find the largest index for which the condition does not fail, and we pass \( \{H_{11}, \ldots, H_{K_2}\} \) to the next step in the procedure.

In Step 3, we apply a useful result from Shaked and Shanthikumar (2007). Concretely, consider \( n \) independent Binomial random variables \( X_i \sim B(1, \pi_i) \) with \( i \in [1, \ldots, n] \), and

\[
Y \sim \text{Binomial} \left( n, (\prod_{i=1}^{n} \pi_i)_{1/n} \right).
\]

Then \( Y \) is stochastically smaller than \( \sum_{i=1}^{n} X_i \). This implies that if the condition of Step 3 fails, that is,

\[
\mathbb{P} \left\{ \text{Binomial} \left( k, \left( \prod_{i=1}^{k} T_{i}^{1}/k \right) \right) > \gamma k \right\} > \alpha,
\]

then the condition in Step 4 also fails (i.e., \( \mathbb{P}(\text{PBD}(k, p^{(k)}) > \gamma k) > \alpha \)), where \( p^{(k)} = (T_{1}^{OR}, \ldots, T_{k}^{OR}) \). Step 3 is also a step-up search, which will end at the first index for which the condition does not fail, denoted by \( K_2 \).

Finally, Step 4 of Procedure 2 is equivalent to Step 2 of Procedure 1, but it is applied to a set with \( K_2 \) hypotheses instead of the initial set of \( m \) hypotheses. The equivalence between Procedures 1 and 2 is thus established. The computational advantage of Procedure 2 is illustrated in the supplementary material.

4. Implementation and Related Issues

4.1. Estimation of the Ifdr

So far, we have assumed that the Ifdr is known. However, it is unknown in practice and must be estimated from data. There are several approaches to estimating the Ifdr, which differ in how the null density \( f_o \), the null proportion \( (1 - \pi) \), and the mixture density \( f \) are computed. We use the model-based clustering approach of Fraley and Raftery (2002) to estimate the Ifdr in our simulation experiments. The approach is available in the R package mclust (version of November 20, 2020). The package first clusters the data into \( G \) groups then computes for each \( z \)-value the probability that it belongs to a specific group \( (\pi_g, g \in [G]) \), and finally estimate the corresponding clusterwise densities \( (f_g : g \in [G]) \). The group with the highest \( \pi_g \) will be defined as the null group, and we denote the corresponding proportion and density as \( \pi_0 \) and \( f_0 \). In our numerical studies, \( f_0 \) always corresponds to the cluster whose center is close to zero. The Ifdr is then computed as

\[
\text{Ifdr} = \frac{\pi_0 f_0}{\sum_{g \in [G]} \pi_g f_g}.
\]
Alternatively, the denominator can also be computed as the mixture density \( f(\cdot) \) using a nonparametric kernel density estimator.

Another method for computing the unknown \( \text{lfdr} \) statistic, particularly in applications where the data generating process is blurry or unknown, is through the R package \( \text{locfdr} \) based on Efron (2004, 2008, 2009). We use this approach in the real data application presented in Section 6. Other methods include the adaptive \( z \)-value approach discussed in Sun and Cai (2007), for which code is available at [http://stat.wharton.upenn.edu/~tcai/paper/html/FDR.html](http://stat.wharton.upenn.edu/~tcai/paper/html/FDR.html), and the robust error-correction method recently proposed by Roquain and Verzelen (2021).

### 4.2. Dependencies and Exchangeability

The proof of the optimality of the \( \text{lfdr} \) ranking (Proposition 2) requires that the observables \( (Z = Z_{\theta[m]}) \) be independent. This section investigates the optimality issue under dependence. We first show that ranking by \( \text{lfdr} \) is still optimal when hypotheses are jointly Gaussian and exchangeable (i.e., equal-variables and equal-covariances): \( Z|\theta \sim \mathcal{N}(\mu, \Sigma) \), where \( \mu \) is the common mean, and the correlation matrix corresponding to \( \Sigma \) is equi-correlated:

\[
\Sigma = \begin{bmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \rho \\
\rho & \cdots & \cdots & 1
\end{bmatrix}
\]

where \( \rho \in [0, 1) \). This model can be rewritten as (Carpentier et al. 2021)

\[
Z_i = \mu \theta_i + \rho^{1/2} W + (1 - \rho)^{1/2} \xi_i,
\]

for \( 1 \leq i \leq m \) and \( W, \xi_i \) are independent \( \mathcal{N}(0, 1) \). Then conditional on \( W \) the individual \( Z_i \) are independent and the null mean and variance are \( \rho^{1/2} W \) and \( 1 - \rho > 0 \). In supplementary Section E, by contrast, if hypotheses are non-exchangeable, we provide a counter-example in which the optimality of the \( \text{lfdr} \) ranking does not hold (the numerical results are provided in Table 5).

### 4.3. Asymptotic Guarantees

This section states the conditions and statements of asymptotic validity and optimality. The framework, or definitions of validity and optimality, are developed along the proofs. The results are shown under the mildest conditions, the convergence of the \( \text{lfdr} \)
estimates, without requiring stringent assumptions. We restate a common assumption in the lfdr literature.

**Assumption 1.** (a) [Assumption 4.3 of Basu et al. 2018]. Let \( \hat{lfdr} \) denote the estimated lfdr statistic. Then \( \hat{lfdr}_1 - \hat{lfdr}_1 = o_p(1) \). Also, \( \hat{lfdr}_1 \xrightarrow{d} lfdr \), where \( lfdr \) is an independent copy of \( \hat{lfdr}_1 \). (b) [Support of the lfdr statistic.] For a choice of \( \epsilon' > 0 \), depending only on \( \gamma \) and \( \alpha \), assume \( P(\lfdr \leq \epsilon') \geq \rho \epsilon' > 0 \).

**Proposition 4.** (Asymptotic Validity.) It holds that \( P(FDP \geq \gamma + \epsilon) \leq \alpha + o(1) \), for any \( \epsilon > 0 \), under Assumption 1.

**Proposition 5.** (Asymptotic Optimal Ranking.) Under Assumption 1, the proposed procedure using the estimated lfdr has an asymptotically optimal ranking.

The above results complete the asymptotic picture of our proposed methodology. Their proofs are completed in Appendix D. Next; we present numerical experiments to illustrate the finite-sample performances.

## 5. Numerical Experiments

We present several numerical examples to highlight the properties of Procedure 2 and illustrate how it compares to other competing methods. We start with a setup based on the two-group model (1) that assumes independence, then turn to a setup that incorporates dependency. In the supplementary material, we simulate a setup that mirrors the real-life application presented in Section 6. In the implementation of our procedures we compute PBDs following the method described in Biscarri, Zhao, and Brunner (2018) as implemented in the R Package “PoissonBinomial”.

### 5.1. Independent Hypotheses

We consider a setup where the observed data are generated from the two-group model (1). The non-null proportion is given by \( \pi \in \{0.1, 0.2, 0.3\} \), the null distribution is \( \mathcal{N}(0, 1) \), and the non-null distribution is \( \mathcal{N}(\mu, 1) \), with \( \mu \in \{-1.5, -2, -2.5\} \). This is a standard setting; for example, the case with \( \pi = 0.2, \mu = -2 \) corresponds to the example in Table 1 of Heller and Rosset (2021).

We simulate 10,000 datasets, each with \( m = 5000 \) tests. We evaluate Procedure 2 for \( (\gamma = 0.05, \alpha = 0.05) \), and compare it to a set of representative procedures: Sun and Cai (2007) (SC), Benjamini and Hochberg (1995) (BH), Guo and Romano (2007) (GR), and Lehmann and Romano (2005) (LR). All SC procedures reported here, as a well-justified contender to the BH method, use the estimated lfdr statistic. For Procedure 2 we consider three versions: the oracle, which knows the non-null distribution and the non-null proportion (Oracle), the version where the distributional parameters are unknown and need to be estimated (lfdr), and the version where the distributional parameters are unknown but a strong prior of 1 is imposed on the proportion of nulls (lfdr(\(\hat{\pi} = 0\), which can be viewed as a conservative version of the lfdr procedure. The FDX and ETP levels of different methods are obtained by averaging the testing results from the 10,000 datasets. We summarize the empirical performances of different methods in Table 1. In the FDX row, the numbers represent the relatively frequencies among the 10,000 experiments where the realized FDP exceeds \( \gamma \).

The following three general patterns can be observed across various simulation scenarios:

1. The BH and SC procedures, which are designed to control FDR, are effective for FDR control, with SC close to the desired level \( \gamma \), and BH close to \( (1 - \pi)\gamma \). Both procedures have highly inflated FDX especially for weak and sparse scenarios: A key motivation for our work.
2. The procedures that are designed to control FDX have varying degrees of success. The FDX level of Procedure 2 yields an \( \alpha \) that is generally close to the 5% target. By contrast, GR and LR become more conservative as sparsity decreases (i.e., \( \pi \) increases) and as the absolute value of the non-null effect becomes larger.
3. The FDX methods are in general less powerful than the FDR procedures. Procedure 2 is more powerful than GR and LR in most settings, and is relatively closer to BH in power than the other two FDX methods are in some settings.

Finally, we compare the performance of the three versions of Procedure 2. When the \( \hat{lfdr} \) is estimated, the procedure generally delivers higher levels of FDX, particularly when signals are sparse and weak. To address the FDX inflation in situations where the signals are sparse and weak, we propose a conservative approach which imposes the assumption that the null proportion is approximately 1 [lfdr(\(\hat{\pi} = 0\)]. We see that this approach leads to conservative FDX control but is still more powerful than GR and LR.

### 5.2. Impact of Dependence

As mentioned in Section 4.2, Procedure 2 is not guaranteed to control the FDX when the hypotheses are nonexchangeable. However, in particular cases which are useful in practice, Procedure 2 can still achieve excellent performance by using its empirical Bayes framework to estimate the \( \hat{lfdr} \) statistic in order to learn the structure of the null distribution.

First, we illustrate the impact of dependence on the distribution of Student t-statistics using scenarios which may reasonably reflect dependence structures in real life. Figure 2 shows a scatterplot of 100 pairs of two null test statistics under three settings. In the left panel, the two null test statistics are independently generated. In the middle panel, the data are dependent, because they are generated as the sum of an independent \( \mathcal{N}(0, 1) \) variable and a noise variable following an equi-correlated structure with \( \rho = 0.25 \), and marginal variances scaled to 1. In the right panel, the observed data are again dependent but follow a hierarchical setup. In this setup \( \pi = 0.1 \), and non-null \( z \)-values can be divided into two groups: group 1 follows the marginal distribution \( \mathcal{N}(0.25, 1) \) with probability \( \pi^+ = 0.05 \) and group 2 follows the marginal distribution \( \mathcal{N}(-0.25, 1) \) with probability \( \pi^- = 0.05 \). Under the null, observations arise from a perturbed variation of the standard normal \( \mathcal{N}(\mu, 1) \), where \( \mu \sim \text{Unif}[-1.0, 1.0] \). Note that conditional on the realized value of \( \mu \), the observations are independent; unconditionally, the
null observations are not. Specifically, the correlation among the Student test-statistics is about 0.25. Figure 2 shows a distinctly different distribution for the independent statistics from the distribution for dependent statistics. However, aside from the third scatterplot being more spread out due to the location shift, the patterns in the last two plots look strikingly similar.

In our numerical experiments, we simulate data for $m = 5000$ tests. The goal is to control the FDX at $\gamma = 0.1$ with confidence level $1 - \alpha = 0.95$. We compare our proposed procedure to the method in Guo and Romano (2007) (GR), and the bootstrap-based procedure in Romano and Wolf (2007) and Romano, Shaikh, and Wolf (2008) (RSW). Notably, RSW is shown to asymptotically control the FDX in the presence of arbitrary dependencies.

Table 2 reports the realized FDX, average power (correctly rejected hypotheses as a percentage of the number of non-nulls), and the average of the realized FDPs over 10,000 experiments. In Columns 4–5, we tabulate results for GR and RSW, both of which assume the theoretical null (i.e., $\mathcal{N}(0, 1)$). Columns 6–7 provide corresponding statistics when the null distribution is estimated from the data (i.e., the empirical null). We correct for the unknown null location by re-centering all the test statistics using the mean of all observations. Note that Procedure 2 is designed to adapt to the scenario automatically so that there is no essential difference when re-centering the data. The $\hat{lfdr}$ statistic is estimated using the locfdr package (Efron 2004).

We can see that both GR and RSW yield FDX levels that are much higher than the nominal level. Both procedures do better in FDX control when the estimated empirical null is used (columns 6–7), although RSW becomes very conservative. By contrast, Procedure 2 is capable of controlling the FDX at the desired level while maintaining good power.

6. Application: Financial Trading Strategies

We apply Procedure 2 to a real-life example where the goal is to identify interesting trading strategies among over two million candidate strategies, as in Chordia, Goyal, and Saretto (2020). The construction of trading strategies reflects exactly the simulation setup of Section G in the supplementary materials. Each trading strategy is constructed by sorting stocks into deciles based on a trading signal at the end of June of each year. Stocks in the top decile are sold short. Portfolio compositions are held for 12 months, but the weights are rebalanced monthly to reflect value-weighted exposures (i.e., stocks weights are proportional to relative market capitalizations). As trading signals, we consider every variable in the combined COMPSTAT/CRSP datasets: we take the level, the growth rate, the ratio between computational risk-premium, an idiosyncratic mean-zero, and a time-invariant component ($\alpha_i$). Under the null, the returns are

![Figure 2](image-url)
entirely due to compensation for exposure to systematic risk factors; the time-invariant component is precisely zero (i.e., $\alpha_t = 0$). We test whether the portfolio $\alpha_t$ is zero (i.e., this is a two-tail test) for 2,396,456 strategies. Thus, we have a standard multiple testing problem.

The rationale behind adopting a combinatorial approach to constructing trading signals is essentially twofold: it accounts for both variables used in common practice and variables left out by common practice. On the one hand, there is a long tradition among finance scholars and practitioners to relate stock returns to accounting variables: quantities such as the equity market value of a firm (i.e., a level), the profitability of the assets (i.e., a ratio of two), and the ratio of assets minus equity, divided by assets (i.e., a transformation of three) have all been studied as predictors of future stock returns. See, for example, Chen and Zimmermann (2021) who construct a large laboratory dataset of such variables. On the other hand, only predictors that worked and were discovered by academics or those that no longer worked and were found by industry practitioners are known. That leaves a large set of possible trading signals: those that were tried by academics or practitioners but did not work, those still used in the industry but that are not widely publicized (for obvious reasons), and those that were never tried in the first place. In other words, there is a significant file drawer problem by considering many trading signals of the same functional form as those that have likely been studied. The combinatorial approach aids in providing an exhaustive set that can be analyzed through the lens of a multiple testing problem. We expect the proportion of signals that are true predictors to be tiny. Thus, we would expect that a proper multiple testing procedure fails to select the very great majority of the strategies. However, because our data contains many trading strategies that have already been reported as profitable, we expect some amount of divergence in the tails of the respective distributions (i.e., few genuinely nonzero alphas). How many non-nulls may be in the data is a question that can only be answered by correcting for multiple hypotheses.

As mentioned above, this set of trading strategies has been studied in Chordia, Goyal, and Saretto (2020). The authors apply several multiple hypotheses procedures and still “find” many profitable strategies before applying some economic restriction. We expect the proportion of signals that are true predictors to be tiny. Thus, we would expect that a proper multiple testing procedure fails to select the very great majority of the strategies. Probably a problematic aspect of that study is the failure to incorporate the information gained from the data about the null distribution into the procedures. Thus, this particular dataset seems perfect to evaluate Procedure 2, which heavily relies on ranking hypotheses by the data-driven $\ell_0$-norm (i.e., the one that relies on the empirical null).

In Figure 3 we compare the histogram of the distribution of 2,396,456 alpha $t$-statistics, with the theoretical null (i.e., a normal with mean zero and standard deviation equal to 1) and with the empirical null. We estimate the empirical null distribution parameters and null proportion using the analytical method described in Section 4 of Efron (2008).

The data is more spread out than the theoretical null but relatively close to the empirical null. The cross-sectional distribution of estimated alphas is dependent because some signals are correlated, and alpha is conditional on a set of common returns. In that sense, conceptually, the Efron empirical null is a much better approximation to the data-generating model. However, because our data contains many trading strategies that have already been reported as profitable, we expect some amount of divergence in the tails of the respective distributions (i.e., few genuinely nonzero alphas). How many non-nulls may be in the data is a question that can only be answered by correcting for multiple hypotheses.

We implement Procedure 2 and report in Table 3 the number of strategies that are selected for different levels of $\gamma$ and $\alpha$, where $\gamma$ denotes the maximum allowable proportion of false discoveries (FDP), and $\alpha$ refers to the allowable tail probability. Similar to what we do for the simulation exercise described in a previous section, we apply both Procedures 2 and 2 GR to the set of 2,396,456 trading strategies and construct a sample of possible trading strategies from accounting and stock price information for the period between 1972 and 2015.
in Section G in the supplementary materials, we compare the results obtained from applying Procedure 2 to those obtained by applying the Guo and Romano (2007) procedure (GR), the FDP-StepM procedure of Romano and Wolf (2007) and Romano, Shaikh, and Wolf (2008) (RSW), and the FDR procedure of Sun and Cai (2007) (SC).

In general, the number of findings increases with how many false discoveries are allowed. For example, for a choice of $\gamma = 0.10$ and $\alpha = 0.05$, we pick out 20 strategies, while for $\gamma = 0.2$ the procedure selects 47 strategies. The number of selected strategies also varies considerably with the assumption about the shape of the null hypothesis: if one relies on the theoretical null (Panel B) instead of Efron’s empirical null, the number of discoveries grows dramatically.

Procedure 2 selects more strategies than the standard frequentist procedure of Guo and Romano (2007), which emerges from our simulation as the most powerful solution in the frequentist’s paradigm. This is not surprising as the result of the simulation presented in the previous section confirms GR to be less powerful. For example, in the case where the empirical null is used, $\gamma = 0.10$, and $\alpha = 0.10$, Procedure 2 selects 24 strategies while GR selects 3. When the theoretical null is used for the same parameters, Procedure 2 selects 417,000 strategies, while GR selects 352,000. This is understandable as Efron’s method relaxes the null specification, allowing more of the data to fall under the null distribution. This ultimately restricts the number of strategies that can be selectable by any procedure.

Finally, compared to procedures based on entirely different assumptions, the number of selected strategies by Procedure 2 is between RSW, and SC. RSW aims to control FDP at the same levels of $\gamma$ and $\alpha$. Still, it is very conservative while SC controls FDR at a level equal to $\alpha + (1 - \alpha) \gamma$. The comparison reinforces the cautionary message about applying a multiple comparisons procedure to a vast number of tests when all the information in the data is not taken into account, resulting in questionable answers. Learning the parameters of the data-generating process is, therefore, a valuable effort, especially in the context of the procedure that we propose in the article, which relies on the local false discovery rate as one of its primary inputs.

7. Discussion

Controlling the FDX provides an instrumental framework for applications where experiments are carried out only once or a few times. Unlike FDR methods, which only offer a long-run average guarantee, FDX methods provide a high-confidence control of the FDP for individual experiments. Under an empirical Bayes framework, we employ the PBD and its connection to $lfdr$ to develop an ($\gamma, \alpha$)-level FDX procedure and demonstrate its superiority over existing methods. There are several open issues as the scope of this work progresses.

One such issue is that several other error rates could be more attractive to specific researchers. For example, maximizing power only on the “nicer” realizations, that is, on those realizations where the FDP is indeed controlled at $\gamma$. Such exciting and highly relevant error rate notions are left for future explorations.

Also, an essential question for FDX and FDR methods based on the empirical Bayes framework is how to estimate the $lfdr$ statistic in practice. This article only considers a few available estimates to provide an implementable FDX procedure. Although valuable, a careful study of $lfdr$ estimation is outside the scope of this work.

Supplementary Materials

The supplementary materials contain the proofs of several propositions omitted in the main text and, in particular, establish the optimal ranking and the asymptotic guarantees. We also provide several illustrations, such as the computational advantage for procedure 2, a counterexample of ranking by $lfdr$ when the tests are not exchangeable, and a numerical study to illustrate the performance of the proposed procedure for the stock trading strategies via simulation.

Acknowledgments

We thank the two anonymous reviewers, Associate Editor and the Editor, whose comments have significantly improved the work and its exposition. Pallavi Basu is grateful to Profs. Sebastian Döhler, Etienne Roquain, and Daniel Yekutieli for insightful discussions and to Mr. Gunashekhar Nandiboyina for parallel computing support. All errors are our own. The views expressed in this article do not necessarily reflect those of the Federal Reserve Bank of Dallas or its staff.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Part of the work was done while PB was at Tel Aviv University and has received funding from the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 294519-PSARPS. Her research is partially supported by a MATRICS grant (MTR/2022/000073) from DST-SERB, India.

References

Basu, P. (2016). “Model Selection Principles and False Discovery Rate Control,” Ph. D. thesis. University of Southern California. [1042]

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, 1172–1183. DOI:10.1080/01621459.2017.1336443. [1044,1045,1048]

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, 57, 289–300. DOI:10.1111/j.2517-6161.1995.tb02031.x. [1041,1044,1047,1048]

Biscarri, W., Zhao, S. D., and Brunner, R. (2018). “A Simple and Fast Method for Computing the Poisson Binomial Distribution Function,” Computational Statistics and Data Analysis, 122, 92–100. DOI:10.1016/j.csda.2018.01.007. [1048]

Blanchard, G., Neuviel, P., and Roquin, E. (2020). “Post Hoc Confidence Bounds on False Positives Using Reference Families,” The Annals of Statistics, 48, 1281–1303. DOI:10.1214/19-aos1847. [1042]

Carpentier, A., Delattre, S., Roquin, E., and Verzele, N. (2021). “Estimating Minimum Effect with Outlier Selection,” The Annals of Statistics, 49, 272–294. [1047]

Chen, A., and Zimmermann, T. (2021). “Open-Source Cross-Sectional Asset Pricing,” Critical Finance Review, forthcoming. DOI:10.2139/ssrn.3604626. [1050]

Chen, S. X., and Liu, J. S. (1997). “Statistical Applications of the Poisson-Binomial and Conditional Bernoulli Distributions,” Statistica Sinica, 7, 875–892. [1045]

Chi, Z., and Tan, Z. (2008). “Positive False Discovery Proportions: Intrinsic Bounds and Adaptive Control,” Statistica Sinica, 18, 857–860. [1041,1042]
Chordia, T., Goyal, A., and Saretto, A. (2020), "Anomalies and False Rejections," *The Review of Financial Studies*, 33, 2134–2179. DOI:10.1093/rfs/hhaa018. [1049,1050]

Dawid, A. (1994), "Selection Paradoxes of Bayesian Inference," Lecture Notes—Monograph Series, 24, 211–220. DOI:10.1214/lmns/1215463797. [1045]

Delattre, S., and Roquain, E. (2015), "New Procedures Controlling the False Discovery Proportion via Romano-Wolf’s Heuristic," *The Annals of Statistics*, 43, 1141–1177. DOI:10.1214/14-aos1302. [1041,1042]

Döbler, S., and Roquain, E. (2020), "Controlling the False Discovery Exceedance for Heterogeneous Tests," *Electronic Journal of Statistics*, 14, 4244–4272. DOI:10.1214/20-ejs1771. [1042,1045]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. DOI:10.1198/01621450400000089. [1044,1047,1049]

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Estimation of False Discovery Rates," *Journal of the American Statistical Association*, 96, 1151–1160. DOI:10.1198/016214501753382129. [1042,1043]

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, 1151–1160. DOI:10.1198/016214501753382129. [1042,1043]

Farcomeni, A. (2008), "A Review of Modern Multiple Hypothesis Testing, with Particular Attention to the False Discovery Proportion," *Statistical Methods in Medical Research*, 17, 347–388. DOI:10.1177/0962280206079046. [1042]

——— (2009), "Generalized Augmentation to Control the False Discovery Exceedance in Multiple Testing," *Scandinavian Journal of Statistics*, 36, 501–517. DOI:10.1111/j.1467-9469.2008.00633.x. [1042]

Fraley, C., and Raftery, A. E. (2002), "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. DOI:10.1198/01621450276004731. [1046]

Fu, L. (2018), "Nonparametric Empirical Bayes Methods for Large-Scale Inference Under Heteroscedasticity," Ph. D. thesis. University of Southern California. [1042]

Genovese, C., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061. DOI:10.1214/009053604000000283. [1041]

Genovese, C. R., and Wasserman, L. (2006), "Exceedance Control of the False Discovery Proportion," *Journal of the American Statistical Association*, 101, 1408–1417. DOI:10.1198/016214506000000339. [1041,1042]

Goeman, J. J., Hemerik, J., and Solari, A. (2021), "Only Closed Testing Procedures are Admissible for Controlling False Discovery Proportions," *The Annals of Statistics*, 49, 1218–1238. DOI:10.1214/20-aos1999. [1042]

Gordon, A. Y., and Salzman, P. (2008), "Optimality of the Holm Procedure among General Step-Down Multiple Testing Procedures," *Statistics & Probability Letters*, 78, 1878–1884. DOI:10.1016/j.spl.2008.01.055. [1041]

Gu, J., and Koenker, R. (2020), "Invidious Comparisons: Ranking and Selection as Compound Decisions," arXiv preprint arXiv:2012.12550. [1045]

Guo, W., and Romano, J. (2007), "A Generalized Sidak-Holm Procedure and Control of Generalized Error Rates Under Independence," *Statistical Applications in Genetics and Molecular Biology*, 6, 1–35. DOI:10.2202/1544-6115.1247. [1041,1042,1047,1048,1049,1050,1051]

Heller, R., and Rosset, S. (2021), "Optimal Control of False Discovery Criteria in the Two-Group Model," *Journal of the Royal Statistical Society, Series B*, 83, 133–155. DOI:10.1111/rssb.12403. [1042,1044,1048]

Hetemer, J., Solari, A., and Goeman, J. J. (2019), "Permutation-based Simultaneous Confidence Bounds for the False Discovery Proportion," *Biometrika*, 106, 635–649. DOI:10.1093/biomet/asz021. [1042]

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, 495–506. DOI:10.1198/016214507000000167. [1044]

Katsevich, E., and Ramdas, A. (2020), "Simultaneous High-Probability Bounds on the False Discovery Proportion in Structured, Regression and Online Settings,” *The Annals of Statistics*, 48, 3465–3487. DOI:10.1214/19-aos1938. [1042]

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004), "Controlling the Number of False Discoveries: Application to High-Dimensional Genomic Data," *Journal of Statistical Planning and Inference*, 124, 379–398. DOI:10.1016/s0378-3758(03)00211-8. [1041]

Lehmann, E. L., and Romano, J. P. (2005), "Generalizations of the Familywise Error Rate," *The Annals of Statistics*, 33, 1138–1154. DOI:10.1214/009053605000000084. [1042,1047,1048]

Perone Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004), "False Discovery Control for Random Fields," *Journal of the American Statistical Association*, 99, 1002–1014. DOI:10.1198/01621450001655. [1042]

Romano, J. P., Shaikh, A. M., and Wolf, M. (2008), "Formalized Data Snooping based on Generalized Error Rates," *Econometric Theory*, 24, 404–447. DOI:10.1017/s0266466608080171. [1049,1050,1051]

———, J. P., and Wolf, M. (2007), "Control of Generalized Error Rates in Multiple Testing," *The Annals of Statistics*, 35, 1378–1408. DOI:10.1214/009053606000000162. [1042,1049,1050,1051]

Roquain, E., and Verzelen, N. (2021), "False Discovery Rate Control with Unknown Null Distribution: Is It Possible to Mimic the Oracle?" *The Annals of Statistics*, Forthcoming. [1047]

Roquain, E., and Villers, F. (2011), "Exact Calculations for False Discovery Proportion with Application to Least Favorable Configurations," *The Annals of Statistics*, 39, 584–612. DOI:10.1214/10-aos1143. [1042]

Shaked, M., and Shanthikumar, J. G. (2007), *Stochastic Orders*, New York: Springer. [1046]

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, 901–912. DOI:10.1198/016214507000000545. [1043,1044,1046,1047,1048,1050,1051]

van der Laan, M. J., Birkner, M. D., and Hubbard, A. E. (2005), "Empirical Bayes and Resampling based Multiple Testing Procedure Controlling Tail Probability of the Proportion of False Positives," *Statistical Applications in Genetics and Molecular Biology*, 4, 1–32. DOI:10.2202/1544-6115.1143. [1042]

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004), "Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives," *Statistical Applications in Genetics and Molecular Biology*, 3, 1–27. DOI:10.2202/1544-6115.1042. [1042]