Federal Reserve
Bank of Dallas

Empirical Bayes Control of the False Discovery Exceedance

Pallavi Basu, Luella Fu, Alessio Saretto and Wenguang Sun

Working Paper 2115 **November 2021**

Research Department

https://doi.org/10.24149/wp2115

Empirical Bayes Control of the False Discovery Exceedance[*]

Pallavi Basu[†], Luella Fu[‡], Alessio Saretto[§] and Wenguang Sun[±]

November 5, 2021

Abstract

In sparse large-scale testing problems where the false discovery proportion (FDP) is highly variable, the false discovery exceedance (FDX) provides a valuable alternative to the widely used false discovery rate (FDR). We develop an empirical Bayes approach to controlling 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 FDX constraint. We propose a data-driven FDX procedure that emulates the oracle via carefully designed computational shortcuts. We investigate the empirical performance of the proposed method using simulations and illustrate the merits of FDX control through an application for identifying abnormal stock trading strategies.

Keywords: Cautious Data Mining; False Discovery Exceedance Control; Local False Discovery Rates; Multiple Hypotheses Testing; Poisson Binomial Distribution; Trading Strategies.

JEL Codes: C11, C12, C15

---

[*]The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Dallas, its staff, or the Federal Reserve System. 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. Part of the work was done while Pallavi Basu was at the 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. All errors are our own.

[†]Pallavi Basu, Indian School of Business, Tel Aviv University, pallavi_basu@isb.edu.

[‡]Luella Fu, San Francisco State University, luella@sfsu.edu.

[§]Alessio Saretto, Federal Reserve Bank of Dallas, alessio.saretto@dal.frb.org.

[±]Wenguang Sun, USC Marshall School of Business, wenguans@marshall.usc.edu.

1 Introduction

Scientific findings often come from experiments performed only a few times or even just once, in which multiple hypotheses are simultaneously tested. Many statistical procedures have been developed to account for the multiple hypothesis testing problem. The most widely used methods aim to control the false discovery rate (FDR). Control of the FDR guarantees that the expected value of the false discovery proportion (FDP) in each trial is below a certain threshold. While procedures that control the FDR are generally easier to implement, they expose researchers to tail events. In particular, focussing one's efforts in maintaining FDR control can be problematic when the variability in the FDP is high in sparse or weak signal settings. For example, even perfectly controlling the FDR at 10% allows for fifty out of a hundred repeated experiments to produce FDPs of 20% as long as the other fifty have FDPs of 0%. Thus if the particular trial in question falls into the first group, it might lead to non easily reproducible findings.

In practice, popular FDR procedures such as the Benjamini–Hochberg procedure (BH) (Benjamini and Hochberg, 1995), display high variability and skewness in the FDP across implementations (Korn et al., 2004; Delattre and Roquain, 2015). Therefore, direct control of the tail probability of the FDP, also known as false discovery exceedance (FDX), is desirable. In general, FDX keeps a low probability, α, if the FDP exceeds an acceptable proportion, γ. For example, a 5% FDX procedure for an FDP of 10% guarantees that the probability of having 10% or more false discoveries is less than 5%. FDX is most similar to the k–FWER, itself an extension of the Family-Wise Error Rate (FWER), except that it controls the probability of an undesirable event defined not by several false rejections k but by a proportion of false rejections γ. This allows FDX methods to scale up with the number of correct rejections as FDR methods do. The merits of controlling the FDX have been discussed in Genovese and Wasserman (2004, 2006); Guo and Romano (2007); Chi and Tan (2008); Gordon and Salzman (2008); Delattre and Roquain (2015). Different FDX methods have been proposed in Lehmann and Romano (2005), Chi and Tan (2008), Roquain and Villers (2011), and Döhler and Roquain (2020). These procedures differ in choosing their respective cut-offs for rejected tests but are all similar in using p-values to rank hypotheses.

Differently, we propose an empirical Bayes procedure for FDX control that relies on ranking local false discovery rates (*lfdr*) and using the Poisson binomial distribution to compute the cumulative failure proportion (i.e., the quantity that is controlled in the FDX definition). We prove that the ranking is optimal (see also, Fu, 2018) and the procedure controls the FDX at the pre-specified level (see also, Basu, 2016). Our work is most similar to Döhler and Roquain (2020) who use the Poisson Binomial distribution to threshold p-values under the frequentist setting. From a theoretical standpoint, the derivation of our procedure is also similar to the theoretical work of Heller and Rosset (2021) who study FDR control in an empirical Bayes framework. We demonstrate the power gain over other FDX procedures in simulations and

provide a real-life application to the problem of isolating interesting financial trading strategies.

Procedure: First, compute the local false discovery rate test statistic (*lfdr*), adjusting for an empirical null if needed. Next, sort all the hypotheses by increasing order of *lfdr*. A posteriori, the unknown states of the hypotheses are Bernoulli with the *lfdr* as failure probabilities, where failure indicates that the null hypothesis is true. Compute the probability of the cumulative failure proportion greater than γ as the probability of a Poisson binomial random variable. Reject the maximum number of hypotheses K, allowing for the cumulative failure proportion greater than γ to be lower than or equal to α. Define hypothesis $\mathcal{H}_{K+1}$ as the first hypothesis that is not rejected (i.e., the one with *lfdr* greater than the threshold *lfdr*). To achieve the exact level α, randomize the decision to accept or reject $\mathcal{H}_{K+1}$ by an independent coin toss with appropriate success probability. In practical terms, the last step might be avoided (i.e., the size of the procedure could be smaller than α).

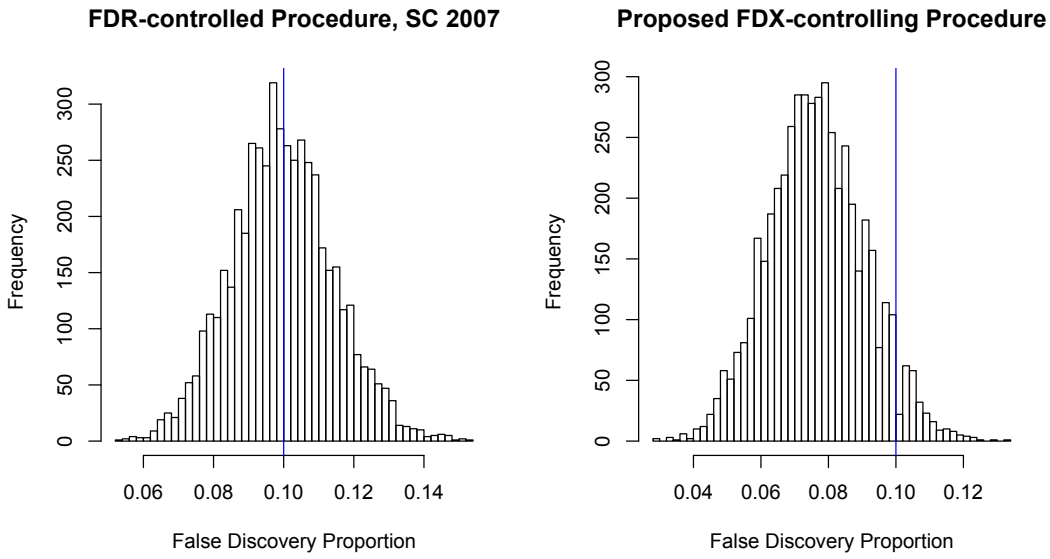
An illustration of the implementation of our FDX controlling procedure is provided in Figure 1. By keeping a low probability on undesirably high levels of FDP each time the testing procedure is carried out rather than keeping an average FDP level low over many hypothetical experiments on which the same testing procedure is applied, FDX procedures reduce the variability of the FDP's distribution and increase the reproducibility of scientific discoveries.

1.1 Contribution

We view an FDX procedure in a two-group model as maximizing the power subject to a constraint on the tail probability of the FDP. We adopt an empirical Bayes viewpoint and suggest a procedure that we use to test millions of hypotheses. We rely on computational efficiency, practical uses, and theoretical understandings. Our work differs from other works on FDP control (see, Farcomeni (2008) for a detailed review). Genovese and Wasserman (2006) first defined the notion of the *exceedance* control of the FDP where the authors suggested tests of uniformity for all subsets of tests. This class of tests is defined as inversion for determining the thresholds and typically uses p-values to rank the hypotheses. Their approach was extended to random fields in Perone Pacifico et al. (2004). The other approach to the FDX question is develop an augmentation-based approach from FWER, developed in van der Laan et al. (2004), and later extended in Farcomeni (2009). A more powerful bootstrap-based Monte Carlo approach was developed in van der Laan et al. (2005). Their procedure uses Monte Carlo simulations to generate the states of hypotheses conditional on observing the data. However, the hypotheses are still ranked using (adjusted) p-values. Another line of work by Delattre and Roquain (2015) formally justifies the bootstrap-based heuristics developed in Romano and Wolf (2007). A more recent line of work involves providing the confidence bounds for FDP

Figure 1: Contrasting FDR and FDX methods

Realized FDPs of 5000 replications for 5000 tests from the following Gaussian mixture model: $0.8\mathcal{N}(0,1)+0.2\mathcal{N}(-2,1)$. This illustration contrasts our proposed FDX procedure with the lfdr-based FDR procedure proposed by Sun and Cai (2007). Both procedures use the oracle version of the *lfdr* test statistic for demonstration purposes. The Sun and Cai (2007) procedure aims at controlling FDR at 0.10, while the FDX procedure aims to keep the probability that FDP is larger than 0.10 at below 0.05.



for a user-specified (aka post hoc) rejection sets, see, for example, Hemerik et al. (2019) and follow up works such as Blanchard et al. (2020), Katsevich and Ramdas (2020), and Goeman et al. (2021). The ranking of the hypotheses is not discussed, and the tightness of the FDP bounds is not well established in these works.

Relative to this large body of work, and keeping in mind that our goal is to develop an efficient FDP controlling procedure that can be applied for millions of tests, we make several contributions:

- We propose a new empirical Bayes approach to FDX control and illustrate the efficiency gain over existing frequentists methods such as Lehmann and Romano (2005) and Guo and Romano (2007).
- We establish the optimality theory for FDX control by showing that the *lfdr* ranking is optimal because the thresholding rule based on *lfdr* has the largest power subject to the constraint on the FDX.
- We develop an efficient computational algorithm for determining a data-driven cutoff along the *lfdr* ranking. We provide supporting results justifying the algorithm for FDX control.
- We demonstrate the strong empirical performance of the proposed method via both simulated and real data sets.

We note that a procedure proposal, where the ranking of the hypotheses is thoroughly justified and valid thresholding is offered, has been incomplete in the literature until the current work. A particular merit of our work is that we have paid attention to simplicity and efficiency in order to enable broad applicability.

1.2 Motivating Example

Our primary motivating example is from a recent work by Chordia et al. (2020) published at Review of Financial Studies. We analyze two million trading strategies based on publicly available signals to isolate some with potentially attractive returns. Strategies are first benchmarked against factors that reflect aggregate market conditions to determine an abnormal return measure (i.e., an alpha). We apply several FDX controlling procedures to determine a cut-off for rejection of the null of zero alpha. We find that our approach identifies more trading strategies than existing state-of-the-art FDX methods. At the same time, it identifies far less compared to FDR controlling methods, alarming applied research to conduct data mining exercises way more cautiously than existing practices.

1.3 Organization

The rest of the paper is organized as follows. Sections 2 and 3 provide more details on the setup and our proposed solutions. Section 4 discusses some concerns and ideas that help enhance our understanding of the question. Section 5 provides numerical simulations, with particular stock returns experiments matching closely to our motivating data question. Section 6 analyzes our motivating example in depth. Section 7 concludes with some future propositions. All codes for the procedure and the experiments can be requested from the authors.

2 Problem Formulation

2.1 Model and Notation

The premise of our analysis is motivated by the two-group model of Efron et al. (2001). Let π denote the probability that the alternative hypothesis is true and $1-\pi$ denote the probability that the null hypothesis is true. Let $[m]$ denote the index set $\{1,\ldots,m\}$ of hypotheses. Given observations $\boldsymbol{Z} = (Z_i)_{i\in[m]}$ we want to test the hypotheses $\mathcal{H} = (H_i^0, H_i^1)_{i\in[m]}$, where $H_i^0 : \theta_i = 0$ and $H_i^1 : \theta_i = 1$, where θ_i denotes the true state of nature with $\theta_i = 0$ representing the null hypothesis and $\theta_i = 1$ the alternative. In a hierarchical two group model:

$$\begin{aligned} \theta_i &\overset{iid}{\sim} Bernoulli(\pi) \\ Z_i|\theta_i &\sim (1-\theta_i)F_0 + \theta_i F_1, \end{aligned} \tag{1}$$

where F_0 is the distribution of Z_i under the null hypothesis; and F_1 is the distribution of Z_i under the alternative hypothesis. We must make a decision, $\delta_i \in \{0,1\}$ indicating our belief about θ_i. A decision of $\delta_i = 1$ indicates the rejection of the null (aka 'a statistical discovery'), where $\delta_i = 0$ indicates failure to reject the null hypothesis. The key ranking test statistic for our methodology is the (marginal) local false discovery rate, *lfdr*, defined as

$$lfdr(Z_i) = (1-\pi)f_0(Z_i)/f(Z_i), \tag{2}$$

where $f_0(\cdot)$ and $f(\cdot)$ denote the null and the mixture probability density functions respectively corresponding to the null distribution function $F_0(\cdot)$ and the mixture distribution function $F(\cdot)$ respectively.

2.2 False Discovery eXceedance (FDX) and Power

We are interested in developing a decision rule $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_m)$ such that the expected number of true positives (ETP) is *efficiently* maximized subject to FDX control. Formally we formulate

power as

$$ETP := E\sum_i \theta_i \cdot \delta_i, \tag{3}$$

the expected number of correctly rejected hypotheses. The error-rate that we aim to control:

$$FDX := P(FDP > \gamma) \leq \alpha, \tag{4}$$

defines FDX-control for a given $(\gamma, \alpha) \in (0, 1)$, where γ represents a tolerance level on the FDP and α a low probability event, where the false discovery proportion is

$$FDP := \frac{\sum_i (1 - \theta_i) \cdot \delta_i}{\sum_i \delta_i \vee 1}, \tag{5}$$

where $\sum_i \delta_i \vee 1$ denotes the maximum of $\sum_i \delta_i$ and 1. Hence, our methodological approach follows a decision-theoretic approach to multiple-hypothesis testing problems. We aim to develop a z-value 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 Heller and Rosset (2021). Further, we want to ensure that the methodology is computationally efficient and can be seamlessly used for millions of tests analogous to Benjamini and Hochberg (1995) and Sun and Cai (2007) for FDR-control to enable wide-applicability and eventually uncomplicated adoption.

3 Oracle Procedure for FDX Control

This section proposes an oracle solution to control the (γ, α)–level FDX. We provide the properties of the oracle procedure and some indicative computational shortcuts. Further, we provide a data-driven approximation.

3.1 Oracle Procedure

For a decision rule rejecting k hypotheses, $\sum_{i=1}^{k}(1 - \theta_i)$ is the number of false rejections. For FDX control, we are interested in calculating the conditional control of the tail probability $P_{\theta|\boldsymbol{Z}}(FDP > \gamma) = P_{\theta|\boldsymbol{Z}}(\sum_{i=1}^{k}(1 - \theta_i) > k\gamma)$. This probability can be found using the Poisson binomial distribution, which generalizes the binomial distribution to the case when each trial has a different probability of success, p_i.

Denote the Poisson binomial distribution by $PBD(k, \boldsymbol{p})$, with k being the total number of trials and $\boldsymbol{p} = (p_i : i = 1, \ldots, k)$ the vector of success probabilities. We define T_i^{OR} as the local false discovery rate ($lfdr$) statistic:

$$T_i^{OR} := P(\theta_i = 0 | Z_i = z_i) = \frac{(1 - \pi) f_0(z_i)}{f(z_i)}, \tag{6}$$

where π is the proportion of non-nulls, f_0 is the null density for $\boldsymbol{Z}$, and $f(\cdot)$ is the mixture distribution for $\boldsymbol{Z}$. An oracle FDX procedure at level (γ, α) is given below.

Procedure 1. *1. Consider the lfdr test statistics $(T_i^{OR})_{i\in[m]}$ as in (6), and denote the ranked statistics $(T_{(i)}^{OR})_{i\in[m]}$, in the increasing order.*

2. Let $K := \max\{k : P\left(PBD(k, \boldsymbol{p^{(k)}}) > \gamma k\right) \leq \alpha\}$ where $\boldsymbol{p^{(k)}} = (T_{(1)}^{OR}, \ldots, T_{(k)}^{OR})$. Reject the top K hypotheses along the lfdr ranking.

Note that, 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 hypothesis, $\mathcal{H}_K$, for which the tail probability is less than α. It then rejects all hypotheses in $\{\mathcal{H}_1, \ldots, \mathcal{H}_K\}$.

Define $\mathcal{H}_{K+1}$ as the first hypothesis that the procedure does not reject. To achieve control at the exact level α, we propose to randomize the decision to accept or reject $\mathcal{H}_{K+1}$ by an independent coin toss with appropriate success probability. This randomization is in the spirit of the weighted FDR procedure proposed by Basu et al. (2018) and Gu and Koenker (2020).

3.2 Properties of the Oracle Procedure

We view the design of a multiple testing methodology as a procedure that uses a statistic to rank and threshold tests. We first show that our proposed oracle procedure controls the FDX.

Proposition 1. *(Exact Validity) Procedure 1 controls the FDX at level (γ, α).*

Proof. Procedure 1 ensures $P_{\theta|\boldsymbol{Z}}(\sum_{i=1}^{k}(1-\theta_i) > \gamma k) \leq \alpha$. Furthermore arbitrary randomization at the ultimate decision ensures

$$E_{(\theta,U)|\boldsymbol{Z}}\left[\mathbb{I}_{\sum_i(1-\theta_i)\delta_i^* > \gamma\sum_i\delta_i^*}\right] = \alpha,$$

where δ_i^* is the decision rule determined by the data $\boldsymbol{Z}$ and the independent arbiter U. Taking a further expectation with respect to $\boldsymbol{Z}$ completes the proof. The independent arbiter is chosen to favor one more rejection with probability $\{\alpha - P_{\theta|\boldsymbol{Z}}(\sum_{i=1}^{K}(1-\theta_i) > \gamma K)\}/\{P_{\theta|\boldsymbol{Z}}(\sum_{i=1}^{K+1}(1-\theta_i) > \gamma\{K+1\})\} - P_{\theta|\boldsymbol{Z}}(\sum_{i=1}^{K}(1-\theta_i) > \gamma K)\}$. □

Proposition 2. *(Optimal Ranking) In the i.i.d. two-group model, Procedure 1 has the best ranking for almost all sample points: For any decision rule with FDX-level (γ, α) we can find an lfdr-based thresholding rule at the same level that has a higher or equal ETP.*

Proof. Suppose Proposition 2 is not true. Then for any claimed "optimal" decision rule there must exist a subset in the sample space $\mathbb{Z}$, which depends on the decision rule, such that $\mu\{w : \boldsymbol{Z}(w) \in \mathbb{Z}\} > 0$, where the decisions obtained by ranking *lfdr* are not preserved. Let us refer to this claimed decision rule as $\boldsymbol{\delta}^*$ where $lfdr_j < lfdr_\ell$, but $\delta_j = 0$ and $\delta_\ell = 1$, where j, ℓ are random indices determined by $\boldsymbol{Z}$. Denote $\boldsymbol{\delta}^{new}$ as an alternative decision rule where $\delta_j^{new} = 1$ and $\delta_\ell^{new} = 0$, and every other decision is as in $\boldsymbol{\delta}^*$. We show that $E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^*\right] < E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^{new}\right]$ when $\boldsymbol{Z} \in \mathbb{Z}$. Whenever $\boldsymbol{Z} \notin \mathbb{Z}$, we have $E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^*\right] = E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^{new}\right]$. Let $\varepsilon > 0$ be our uniform measurement precision, that is, $(1 - lfdr_j) \geq \varepsilon + (1 - lfdr_\ell)$. Then we note,

$$
\begin{aligned}
E\left[\sum_i (1-\theta_i)\delta_i^*\right] =& E_{\boldsymbol{Z}}\left[E_{\theta|\boldsymbol{Z}}\left[\sum_i (1-\theta_i)\delta_i^*\right]\right] \\
=& E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^*\{I_{\boldsymbol{Z}\in\mathbb{Z}} + I_{\boldsymbol{Z}\in\mathbb{Z}^c}\}\right] \\
\leq& E_{\boldsymbol{Z}}\left[\sum_i (1 - lfdr_i)\delta_i^{new}\{I_{\boldsymbol{Z}\in\mathbb{Z}} + I_{\boldsymbol{Z}\in\mathbb{Z}^c}\}\right] - \varepsilon \cdot \mu\{w : \boldsymbol{Z}(w) \in \mathbb{Z}\} \\
<& E\left[\sum_i (1-\theta_i)\delta_i^{new}\right],
\end{aligned}
$$

where the strict inequality is due to $\mu\{w : \boldsymbol{Z}(w) \in \mathbb{Z}\} > 0$.

Next we verify that $\boldsymbol{\delta}^{new}$ is a valid decision rule. Note that the number of rejections, k_0, is the same in $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}^{new}$. Suppose we show that

$$
P_{\theta|\boldsymbol{Z}}\left(\sum_{i=1}^k \{(1-\theta_i)\delta_i^{new}\} > \gamma k_0\right) \leq P_{\theta|\boldsymbol{Z}}\left(\sum_{i=1}^k \{(1-\theta_i)\delta_i^*\} > \gamma k_0\right) \tag{7}
$$

then if $\boldsymbol{\delta}^*$ is valid so is $\boldsymbol{\delta}^{new}$. We now show that (7) holds. Define $\boldsymbol{K}$ as the set of rejected strategies common to both decision rules. Conditional on $\boldsymbol{Z}$, we have that $\sum_{i=1}^k \{(1-\theta_i)\delta_i^*\} \sim PBD(\{lfdr_{i\in\boldsymbol{K}}, lfdr_\ell\}, k_0)$, where PBD is a random variable with Poisson binomial distribution.[1] Similarly, we have that $\sum_{i=1}^k \{(1-\theta_i)\delta_i^{new}\} \sim PBD(\{lfdr_{i\in\boldsymbol{K}}, lfdr_j\}, k_0)$. Recall that we have $lfdr_j < lfdr_\ell$. We consider random variables $Y^* \sim PBD(\{p_1 \cdots p_\ell^* \cdots p_{k_0}\}, k_0)$ and $Y^{new} \sim PBD(\{p_1 \cdots p_j^{new} \cdots p_{k_0}\}, k_0)$ with $p_j^{new} < p_\ell^*$. Then to show that $P(Y^{new} > \gamma k_0) \leq P(Y^* > \gamma k_0)$.

Because the states of the hypotheses are independent, we can express $P(Y^* > \gamma k_0) =$

[1]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. PBDs have found their use in multiple applications; see, for example, Chen and Liu (1997) for a survey on applications.

$p^*_\ell P(\bar{Y}^* > \gamma k_0 - 1) + (1 - p^*_\ell) P(\bar{Y}^* > \gamma k_0)$ where $\bar{Y}^* \sim PBD(\{p_{i \in \boldsymbol{K}}\}, k_0 - 1)$. Then,

$$\begin{aligned} P(Y^* > \gamma k_0) =& P(\bar{Y}^* > \gamma k_0) + p^*_\ell \{P(\bar{Y}^* > \gamma k_0 - 1) - P(\bar{Y}^* > \gamma k_0)\} \\ =& P(\bar{Y}^* > \gamma k_0) + p^*_\ell P(\bar{Y}^* = \lceil \gamma k_0 - 1 \rceil) \\ \geq& P(\bar{Y}^* > \gamma k_0) + p^{new}_j P(\bar{Y}^* = \lceil \gamma k_0 - 1 \rceil) \\ =& P(Y^{new} > \gamma k_0), \end{aligned}$$

where $\lceil x \rceil$ refers to the nearest integer strictly larger than x. □

3.3 Poisson Binomial Distribution and Connection to *lfdr*

Poisson's Binomial Distributions or PBDs have been used in the FDR literature, more recently by Döhler and Roquain (2020) in the context of controlling the FDX for heterogeneous tests. In the two-group model, each hypothesis is true either under the null or the alternative with probability $1 - \pi$ and π. However, conditional on the observations $\boldsymbol{Z} = (Z_i)_{i \in [m]}$, the states of the hypotheses marginally follows a Bernoulli distribution with heterogeneous expectations, namely, $P(\theta_i = 0 | \boldsymbol{Z})$. In the case that the joint density of the vector $\boldsymbol{Z}$ conditional on the hypotheses states can be factorized into marginal densities, $P(\theta_i = 0 | \boldsymbol{Z})$ reduces to $P(\theta_i = 0 | Z_i)$, which is equal to the local false discovery rate evaluated at Z_i (i.e., $lfdr(Z_i)$). Furthermore, if the conditional states of hypotheses are independent, the partial sum of their state indicators (i.e., θ_i) is a random variable that follows a PBDs.

Next, we discuss the connection of the PBD to the *lfdr* procedure. In the procedure that we introduce in the previous sections, we rank hypotheses by the *lfdr*, which is determined by the data. We show here that a ranking of hypotheses determined by the data does not alter the conditional posterior distribution of their states.

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

Proof. Define random variables $\mathcal{R}_i := \theta_i \mathcal{I}_{i \in S_{\boldsymbol{Z}}}$ and $S_{\boldsymbol{Z}}$ as determined by the observations $\boldsymbol{Z} = (Z_i)_{i \in [m]}$. Consider any nonempty partial set of indices denoted by $S_{\mathcal{P}}$. Suppose if $S_{\mathcal{P}} \nsubseteq S_{\boldsymbol{Z}}$, then there exists an index $i_0 \in S_{\mathcal{P}} \cap S^c_{\boldsymbol{Z}}$. Then $E[\prod_{i \in S_{\mathcal{P}}} R_i | \boldsymbol{Z}] = 0 = E[R_{i_0} | \boldsymbol{Z}] \cdot E[\prod_{i \in S_{\mathcal{P}} \setminus i_0} R_i | \boldsymbol{Z}]$. Alternately, if $S_{\mathcal{P}} \subseteq S_{\boldsymbol{Z}}$, then we have $E[\prod_{i \in S_{\mathcal{P}}} R_i | \boldsymbol{Z}] = E[\prod_{i \in S_{\mathcal{P}}} \theta_i | \boldsymbol{Z}] = \prod_{i \in S_{\mathcal{P}}} E[\theta_i | \boldsymbol{Z}] = \prod_{i \in S_{\mathcal{P}}} E[R_i | \boldsymbol{Z}]$. Therefore conditional on $\boldsymbol{Z}$, the R_i's are identically zero if $i \notin S_{\boldsymbol{Z}}$ and independently distributed as Bernoulli random variables with an expectation of $lfdr(Z_i)$ if $i \in S_{\boldsymbol{Z}}$. □

Note that in our procedure, the index set $S_{\boldsymbol{Z}}$ is, more specifically, determined by the thresholding of the *lfdr* test statistic. However, we have proved that this does not distort the hypothesis's Poisson binomial distribution of conditional states.

3.4 Computational Shortcuts

Because Procedure 1 is a step-up procedure, it starts by computing the tail probability of the Poisson binomial distribution for all tests under consideration. At each progressive step, the set of tests under consideration decreases. If there is a massive number of tests under consideration, the procedure can be computationally intensive. To increase computational efficiency, we modify Procedure 1 as follows. We justify the modifications after we present the updated procedure.

Procedure 2. *1. Consider the lfdr test statistics* $(T_i^{OR})_{i\in[m]}$ *as in* (6)*, and denote the ranked statistics* $(T_{(i)}^{OR})_{i\in[m]}$*, in the increasing order.*

2. First reject up to $K_1 := \max\{k \in [m] : \ E_{\theta|\boldsymbol{Z}}\left(\sum_{i=1}^{k}(1-\theta_i)\right) \leq k \cdot [\alpha + \gamma(1-\alpha)]\}$*,*

3. Next reject up to $K_2 := \max\{k \in [K_1] : P\left(Y > \gamma k\right) \leq \alpha\}$*, where* $Y \sim B(k, (\prod_1^k T_{(i)})^{1/k})$*,*

4. Finally reject only up to $K := \max\{k \in [K_2] : P\left(PBD(k, \boldsymbol{p^{(k)}}) > \gamma k\right) \leq \alpha\}$*, and* $\boldsymbol{p^{(k)}} = (T_{(1)}^{OR}, \ldots, T_{(k)}^{OR})$*.*

Relative to Procedure 1, there are 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 operated on a small set of hypotheses. The idea is that both Step 2 and Step 3 are step-up procedures in themselves: failure to meet their criteria guarantees that the tail probability criteria in Step 4 also fail.

Step 2 is equivalent to the FDR control in Sun and Cai (2007) at the level $\alpha + \gamma(1-\alpha)$. In fact, note that

$$P_{\theta|\boldsymbol{Z}}\left(\sum_{i=1}^{k}(1-\theta_i) > \gamma k\right) \leq \alpha,$$

implies

$$E_{\theta|\boldsymbol{Z}}\left(\sum_{i=1}^{k}(1-\theta_i)\right) \leq k \cdot \{\alpha + \gamma(1-\alpha)\}.$$

To see this, rewrite the expectation by separating the event where the argument (i.e., the sum of false positives) is greater than γk and the complement event (i.e., sum is less or equal to γk):

$$E_{\theta|\boldsymbol{Z}}\left(\mathcal{I}_{\sum_i(1-\theta_i)>\gamma k}\sum_{i=1}^{k}(1-\theta_i)\right) + E_{\theta|\boldsymbol{Z}}\left(\mathcal{I}_{\sum_i(1-\theta_i)\leq\gamma k}\sum_{i=1}^{k}(1-\theta_i)\right).$$

The sum in the first expectation is bounded by k, and its probability is bounded by α, so the expectation is bounded by $k\alpha$. In the second expectation, the number of false positives is no greater than γk, and the associated probability is $1-\alpha$, which gives a bound of $\gamma k(1-\alpha)$. Adding the two pieces together, the expected number of false positives is bounded above by

$k \cdot \{\alpha + \gamma(1-\alpha)\}$. Thus if the condition in Step 4 holds, the condition in Step 2 will also hold. From this follows that if the condition in Step 2 fails, the condition in Step 4 also fails, and hence the reduction of the set of hypotheses produced by Step 2 is legitimate, as it only eliminates cases in which the condition in Step 4 would fail.

This is a fast step because it only involves computing cumulative average *lfdr* over progressively smaller sets of hypotheses. Step 2 ends when we find the largest index for which the condition does not fail, and we pass $\{\mathcal{H}_1, \ldots, \mathcal{H}_{K_1}\}$ to the next step in the procedure.

In Step 3, we apply a useful result from Shaked and Shanthikumar (2007), that when considering n independent binomial random variables $X_i \sim B(1, p_i)$, with $i \in \{1, \ldots, n\}$, then the random variable $Y \sim B(n, (\Pi_{i=1}^{n} p_i)^{1/n})$ is stochastically smaller than $\sum X_i$. This implies that if the condition of Step 3 fails (i.e., $P\left(B(k, (\prod_i^k T_{(i)})^{1/k}) > \gamma k\right) > \alpha$) then the condition in Step 4 will also fail (i.e., $P\left(PBD(k, \boldsymbol{T}) > \gamma k\right) > \alpha$). 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 tests instead of the initial set of m hypotheses.

3.5 Illustration of Procedures 1 and 2

We implement Procedure 2 on three independent random samples of 10,000 independent test statistics from the mixture model where 90% come from a $\mathcal{N}(0, 1)$ and 10% arise from a $\mathcal{N}(-2, 1)$. Table 1 reports the realized time for each of the runs, as well as the progressive upper limits of rejections, when $\alpha = 0.05$ and $\gamma = 0.10$.

While the two procedures reject the same number of hypotheses (as is expected) and thus produce the same realized FDP, they take very different times to do so: Procedure 1 run-times are around 3.8 minutes; Procedure 1 accomplishes the task in a fraction of a second. The difference in execution time also decreases with the proportion of true nulls, π: when π is very small, Procedure 1 execution time grows a lot more than Procedure 2's. As for the mechanics of how Procedure 2 operates, we see in Panel B that it quickly first reduces the number of tests from ten thousand to about 300 trials, denoted by column K_1. Then there are further reductions in the number of tests represented by columns K_2 and K. The realized FDPs are close to the desired 0.10, with one in the three reported runs exceeding 0.10.

4 Implementation and Related Issues

4.1 Estimation of the *lfdr*

When it comes to computing the *lfdr* there are many alternatives that differs based on how the null density and the null proportion are computed. We mainly use the model-based clustering approach of Fraley and Raftery (2002) to estimate *lfdr* in our experiments and data application.

Table 1: Computational advantages of Procedure 2

The table report illustrative results of the computational advantage of Procedure 2 (Panel B) relative to Procedure 1 (Panel A). The underlying data generating process is a mixture model where 90% come from a $\mathcal{N}(0,1)$ and 10% arise from a $\mathcal{N}(-2,1)$. We report the number of rejected nulls at each step of the oracle procedure and the total execution time for a sample number of runs (3 reported). FDX control is implemented at $\gamma = 0.05$ and with a confidence level $1-\alpha = 0.95$. We report the execution times for the experimental runs for a CPU using a 3GHz processor with 8GB RAM.

	K_1	K_2	K	Realized FDP	CPU Time
	Panel A: Procedure 1				
Run 1	-	-	150	0.10	3.83 mins
Run 2	-	-	142	0.06	3.68 mins
Run 3	-	-	130	0.07	3.84 mins
	Panel B: Procedure 2				
Run 1	337	200	150	0.10	0.16 secs
Run 2	340	193	142	0.06	0.17 secs
Run 3	294	182	130	0.07	0.22 secs

The approach is available in an R package *mclust* (version of November 20, 2020). The packages cluster the data into G groups, compute the probability that the data belongs to each group, and estimate relative density. We define the group with the highest probability, π_0, and that is clustered around zero as the null, and compute the *lfdr* as

$$lfdr = \frac{\pi_0 f_0}{\sum_{g \in G} \pi_g f_g},$$

alternatively, the denominator can also be computed as the mixture density $f(\cdot)$ using a non-parametric kernel density estimator.

Other possibilities of computing the *lfdr* statistic are the R package *locfdr* based on Efron (2004, 2008, 2009), the approach followed in Sun and Cai (2007), for which code is available at the paper's website, and the robust error-specific correction of Roquain and Verzelen (2021).

4.2 Dependencies and Exchangeability

The proof of Proposition 2 requires that the states of the hypotheses considered are independent. However, we can show that ranking by *lfdr* is still optimal when hypotheses are jointly Gaussian and exchangeable (i.e., equal-variances and equal-covariances): $\boldsymbol{Z}|\boldsymbol{\theta} \sim \mathcal{N}(\mu\ \boldsymbol{\theta}, \Sigma)$, where μ is the common mean, and the correlation matrix corresponding to Σ is equi-correlated. Contrary, if hypotheses are non-exchangeable in Table 4, we provide a counterexample to the optimality by the lfdr ranking for multivariate Gaussian.

Proposition 3. *(Optimal Ranking for Exchangeable Gaussian Observations) When* $\boldsymbol{Z}|\boldsymbol{\theta} \sim \mathcal{N}(\mu\,\boldsymbol{\theta}, \Sigma)$*, Procedure 1 has the best ranking for almost all sample points: For any decision rule with FDX-level* (γ, α) *we can find an lfdr-based thresholding rule at the same level that has a higher or equal ETP.*

Proof. Note that in the exchangeable case, μ takes a specific sign (and value), either positive or negative, for all tests. Suppose without loss of generality that $\mu < 0$. We show that, in such a case, ranking by *lfdr* is equivalent to the ranking by increasing values of z-scores, which is the same as ranking by increasing values of marginal *lfdr*. Rewrite:

$$P(\boldsymbol{Z}|\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{Z}\Sigma^{-1}\boldsymbol{Z}\right\} \cdot \exp\left\{-\frac{1}{2}\mu^2\boldsymbol{\theta}\Sigma^{-1}\boldsymbol{\theta}\right\} \cdot \exp\left\{\mu\boldsymbol{\theta}\Sigma^{-1}\boldsymbol{Z}\right\}.$$

If $z_1 > z_2$, and not altering other z-scores, then $P(\boldsymbol{Z}|\theta_1 = 0, \theta_2 = 1, \boldsymbol{\theta}_{\{-1,-2\}}) > P(\boldsymbol{Z}|\theta_1 = 1, \theta_2 = 0, \boldsymbol{\theta}_{\{-1,-2\}})$, where $\boldsymbol{\theta}_{\{-1,-2\}}$ represents the vector of all θ that are not θ_1 and θ_2. Note, in fact, that

$$\ln\{P(\boldsymbol{Z}|\theta_1 = 0, \theta_2 = 1, \boldsymbol{\theta}_{\{-1,-2\}})/P(\boldsymbol{Z}|\theta_1 = 1, \theta_2 = 0, \boldsymbol{\theta}_{\{-1,-2\}})\} \propto \mu(z_1 - z_2)(\Sigma^{-1}_{ij} - \Sigma^{-1}_{ii}), \quad (8)$$

where $i \neq j$ represent all the off-diagonal entries and $\Sigma^{-1}_{ij} - \Sigma^{-1}_{ii} = 1/(\sigma^2 \cdot (\rho - 1)) < 0$ and $\mu < 0$. $\sigma^2 > 0$ and $\rho < 1$ denotes the common variance and correlation respectively for the Gaussian multivariate density. Now to show that ranking by *lfdr* is equivalent to ranking by increasing values of z-scores: $P(\theta_1 = 0|\boldsymbol{Z}) > P(\theta_2 = 0|\boldsymbol{Z})$. The result follows since $P(\theta_1 = 0|\boldsymbol{Z}) - P(\theta_2 = 0|\boldsymbol{Z}) = \sum_{\boldsymbol{\theta}_{\{-1,-2\}} \in \{0,1\}^{m-2}} P(\theta_1 = 0, \theta_2 = 1, \boldsymbol{\theta}_{\{-1,-2\}}|\boldsymbol{Z}) - P(\theta_1 = 1, \theta_2 = 0, \boldsymbol{\theta}_{\{-1,-2\}}|\boldsymbol{Z})$ and $\theta_i \sim Ber(\pi)$ independently.

Showing that ranking by increasing values of z-scores is the same as ranking by increasing values of marginal *lfdr* is simpler: if $\mu < 0$, a higher positive value of a z-score indicates a higher marginal *lfdr* value. Similarly, the converse holds when $\mu > 0$.

It remains to be shown that ranking by the *lfdr* maximizes the power among all decisions that control the $FDX(\gamma, \alpha)$ in the exchangeable and Gaussian framework. To do so, if the hypotheses are not already ranked by the marginal *lfdr*, consider switching decisions. As we show in the proof to Proposition 2, swapping decisions leads to an increase in power. Thus, the ranking by marginal *lfdr*, will lead to maximize power. Moreover, note that when $\mu < 0$ and $z_1 > z_2$ we have $P(\theta_1 = 0, \theta_2 = 1, \boldsymbol{\theta}_{\{-1,-2\}}|\boldsymbol{Z}) > P(\theta_1 = 1, \theta_2 = 0, \boldsymbol{\theta}_{\{-1,-2\}}|\boldsymbol{Z})$ (as is shown in the previous paragraph). Hence, if the original procedure had valid FDX control, so does the modified procedure. □

Remark 1. *For generalized non-Gaussian distributions, (8) highlights the sufficient condition to guarantee both the optimality of ranking and that ranking by the marginal* lfdr *is adequate. This condition is naturally satisfied for exchangeable elliptical distributions such as the multivariate normal, t, or Laplace distributions. Similarly, for a general test statistic, such as*

the absolute values of the z-score or of the t-statistic, it is sufficient that the left side term of (8) is negative (or positive in $T(\cdot)$). That is, the sufficient condition is that if $T_j(\boldsymbol{z}) := |z_j| < |z_i| =: T_i(\boldsymbol{z})$, then $P(T(\boldsymbol{Z})|\theta_i = 0, \theta_j = 1, \boldsymbol{\theta}_{\{-i,-j\}}) < P(T(\boldsymbol{Z})|\theta_i = 1, \theta_j = 0, \boldsymbol{\theta}_{\{-i,-j\}})$ holds for the joint density. When iid, this condition factorizes to the monotone likelihood ratio (MLR) criterion. Thus this condition may be viewed as a generalized MLR condition for multivariate densities.

Once ranked, hypotheses need to be thresholded. Procedure 2 requires hypotheses to be independent, and hence if applied to a case study where there are dependencies, it is not delivering the optimal threshold. If the source of dependency is unknown, one could still compute the proper posterior probability by evaluating the probabilities of all possible combinations of $\boldsymbol{\theta}$ conditional on data, a highly burdensome computational problem. However, suppose individual z-scores are independent conditionally on the unknown location and scale of the data generating process. In that case, one could first estimate these hyper-parameters, thus obtaining an estimated "null" distribution, and continue pretending independence. We return to this in Section 5.2.

5 Numerical Experiments

We present here a few numerical examples designed to highlight the properties of Procedure 2 and how it compares to other existing procedures. We start with a similar model to that presented in Table 1 of Heller and Rosset (2021),we then introduce conditional dependencies and conclude with a setup that mirrors the real-life application presented in Section 6.

5.1 Independent Hypotheses

We consider a base setup with a non-null proportion of $\pi \in \{0.1, 0.2, 0.3\}$, the null model being generated from a standard normal, and the non-null distribution being $\mathcal{N}(\mu, 1)$ with $\mu \in \{-1.5, -2, -2.5\}$. The base case with $\pi = 0.2,\ \mu = -2$, corresponds to the example in Table 1 of Heller and Rosset (2021). Each of 10,000 simulations considers 5,000 tests. We evaluate Procedure 2 at $\gamma = 0.05$ and $\alpha = 0.05$, 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). For Procedure 2 we consider three alternatives: the oracle, which knows the non-null distribution and the non-null proportion (Oracle), the situation when the distributional parameters are unknown and need to be estimated ($lfdr$), and the situation when the distributional parameters are unknown but a strong prior of 1 is imposed in the proportion of nulls ($lfdr(\hat{\pi} = 0)$). Results are reported in Table 2.

Three themes are true in general across the various scenarios: first, procedures that are

designed to control FDR, SC and BH, get very close to doing so with SC being close to the desired level α, and the BH being close to $(1-\pi)\alpha$ while both procedures have highly inflated FDX especially for weak and sparse scenarios; second, procedures that are designed to control FDX do so, but with varying degrees of success. Our Procedure is generally close to a 5% control in every situation; however, GR and LR become less accurate as sparsity decreases (i.e., π increases) and as the average non-null effect becomes more sizable (i.e., μ becomes more negative). Third, there is an inverse relationship between the ability of a procedure to control the number of false discoveries and its power: methods that control FDX are less powerful, although, at least in this setting, Procedure 2 is more powerful than GR and LR, and relatively closer to BH in power.

Table 2: Comparison of different procedures

The table compares the performance of Procedure 2 relative to some popular methods: Sun and Cai (2007) (SC), Benjamini and Hochberg (1995) (BH), Guo and Romano (2007) (GR), and Lehmann and Romano (2005) (LR). Three version of Procedure 2 are implemented: the oracle version (Oracle), a version where the parameters of the data generating process are estimated from the data ($lfdr$), and a version where we impose the assumption that $\pi = 0$, $lfdr(\hat{\pi} = 0)$. The data generating process is a mixture model where with probability $1-\pi$ the test is drawn from a $\mathcal{N}(0,1)$ (null), and with probability π the test is drawn from $\mathcal{N}(\mu,1)$ (alternative), where $\pi \in \{0.1, 0.2, 0.3\}$ and $\mu \in \{-1.5, -2, -2.5\}$. Each simulation considers 5,000 tests. We repeat the exercise for 10,000 simulations. FDX control is implemented at $\gamma = 0.05$ and with a confidence level $1-\alpha = 0.95$, FDR control is implemented at a nominal level of $\alpha = 0.05$.

						Procedure 2		
		SC	BH	GR	LR	Oracle	$lfdr$	$lfdr(\hat{\pi}=0)$
$\pi = 0.2$	FDX	0.452	0.348	0.040	0.040	0.047	0.076	0.062
$\mu = -1.5$	FDR	0.050	0.040	0.013	0.013	0.015	0.021	0.019
	Power (%)	3.4	2.6	0.3	0.3	0.3	0.4	0.4
$\pi = 0.2$	FDX	0.484	0.242	0.039	0.037	0.052	0.066	0.025
$\mu = -2$	FDR	0.050	0.040	0.004	0.004	0.028	0.029	0.022
	Power (%)	22.6	18.8	1.5	1.2	13.5	14.0	10.6
$\pi = 0.2$	FDX	0.471	0.133	0.014	0.000	0.049	0.054	0.006
$\mu = -2.5$	FDR	0.050	0.040	0.030	0.002	0.036	0.036	0.028
	Power (%)	51.0	46.5	40.6	7.6	44.3	44.3	39.6
$\pi = 0.1$	FDX	0.458	0.399	0.044	0.042	0.047	0.068	0.059
$\mu = -2$	FDR	0.049	0.045	0.007	0.007	0.011	0.016	0.014
	Power (%)	11.2	10.3	1.2	1.2	2.3	3.1	2.5
$\pi = 0.3$	FDX	0.484	0.061	0.019	0.019	0.053	0.059	0.004
$\mu = -2$	FDR	0.050	0.035	0.011	0.002	0.035	0.035	0.023
	Power (in %)	33.2	25.8	8.8	1.4	25.7	25.9	18.7

We now discuss relative differences in performance across the three versions of Procedure

2. When $lfdr$ is estimated, the procedure generally delivers higher levels of FDX, particularly when signals are sparse and weaker. To address the FDX inflation in very sparse and low signal situations, one conservative approach is to impose the assumption that the null proportion is approximately 1, $lfdr(\hat{\pi} = 0)$. We observe a realized FDX that is conservative yet still more powerful than GR and LR when we do so.

5.2 One Example of Specific Dependencies

As mentioned in Section 4.2, the performance of Procedure 2 is not guaranteed when hypotheses are non-exchangeable. There are, however, particular and not uncommon cases where we can take advantage of the fact that our procedure tries to learn the properties of the null distribution by estimating the $lfdr$.

Consider, for example, a hierarchical setup where one hypothesis is null most of the time, with a probability of 0.90. When non-null, one-half of the time, the marginal distribution is $\mathcal{N}(0.25, 1)$, and for the other half, it is $\mathcal{N}(-0.25, 1)$. Under the null, the observations arise from a perturbed variation of the standard normal, $\mathcal{N}(\mu, 1)$ where $\mu \sim Unif[-0.1, 0.1]$. Note that while conditional on the realized value of μ, the observations are independent, unconditionally, the null observations are not: in a sample of size $n = 100$, the correlation among the studentized test-statistics is about 0.25. Figure 2 shows a scatter plot of two null test statistics first when they are independently generated, second when generated via an independent standard normal with an additive correlated noise model with a correlation of $\rho = 0.25$, and a third scenario following the above discussed hierarchical unknown location specified null. Aside from the third scatter plot being more spread out due to the location shift, we cannot visually distinguish the last two cases.

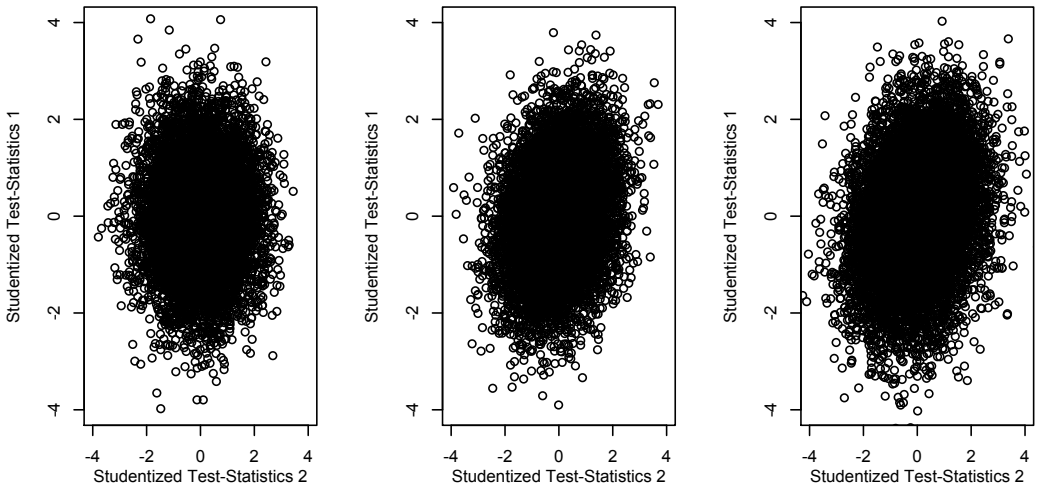
We consider 5,000 tests and control the FDX at $\gamma = 0.1$ and with a confidence level $1 - \alpha = 0.95$. For convenience, we compare our proposed procedure to Guo and Romano (2007) (GR), and to the bootstrap based procedure of Romano and Wolf (2007) and Romano et al. (2008) (RSW) that asymptotically controls FDX in presence of arbitrary dependencies.

Table 3 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 3-4, we tabulate results obtained when GR and RSW assume the theoretical null (i.e., $\mathcal{N}(0, 1)$) as the underlying distribution for the null observations. Columns 5-6 show corresponding statistics when the null distribution is estimated from the data (i.e., the empirical null). We correct for the unknown null location by centering all the test statistics by the mean of the observations. Note that Procedure 2 is designed to adapt to the scenario automatically so that there is no large difference in results. The $lfdr$ estimation in procedure 2 here uses the $locfdr$ estimates following Efron (2004).

In the first scenario, both GR and RSW report a relatively high FDX, much higher in fact than allowed by the selected values of α and γ. Both procedures do better when allowed to

Figure 2: Conditional dependencies

The figure presents a visual comparison of different distributions of test statistics: the leftmost panel corresponds to mutually independent tests; the middle panel shows tests statistics obtained from data that have a 0.25 correlation caused by a common additive noise; the rightmost panel, instead, shows test statistics obtained from conditionally independent samples, which are however conditionally correlated because of a common location shift that induces a correlation of about 0.25.



"learn" something about the background null (columns 5-6), although RSW overshoots and becomes very conservative and loses significant power. In this setup, Procedure 2 holds its ability to control FDX at the desired level while maintaining power.

Table 3: Conditional dependencies

The table reports the average FDX, FDR, and power for 10,000 simulated experiments. Each experiment considers 5,000 tests drawn in a hierarchical setup where a hypothesis is null most of the time, with a probability of 0.90. When non-null, one-half of the time, the marginal distribution is $\mathcal{N}(0.25, 1)$, and one-half of the time it is $\mathcal{N}(-0.25, 1)$. When null, the observations arise from a perturbed variation of the standard normal, $\mathcal{N}(\mu, 1)$ where $\mu \sim Unif[-0.1, 0.1]$. We compare Procedure 2 to Romano et al. (2008) (RSW) and Guo and Romano (2007) (GR). To facilitate comparison, the table shows the results from applying the original RSW and GR procedures (Theoretical Null) alongside the case where the econometrician can estimate the null distribution from the data (Empirical Null). FDX control is implemented at $\gamma = 0.1$ and with a confidence level $1 - \alpha = 0.95$.

		Theoretical Null		Empirical Null	
	Procedure 2	RSW	GR	RSW	GR
FDX	0.046	0.492	0.771	0.003	0.164
FDR	0.057	0.131	0.223	0.016	0.079
Power (%)	28.5	18.2	32.6	10.7	32.2

With the aid of a numerical experiment, we demonstrate that the lfdr test statistic ranking is not necessarily the optimal ranking for non-exchangeable study situations. We follow the setting described in Section 5.2 in Heller and Rosset to provide a counterexample for the best ranking for FDX control. Ten z-scores are generated from the two-group model with $\theta_i \sim Bernoulli(0.3)$ independently. Further $\boldsymbol{Z}|\boldsymbol{\theta} \sim N(-1.5 * \boldsymbol{\theta}, \Sigma + 0.01 * diag(\boldsymbol{\theta}))$. Because the computational burden grows exponentially, we simulate only 10 test statistics and work with $\gamma = 0.5$. Since for independent (or exchangeable) z-scores, we obtain an optimal ranking (by sorting on *lfdr*) regardless of the size of γ, the framework serves as a good counterexample.

We work with a part of the variance-covariance matrix Σ that is block diagonal with two blocks, each block being equicorrelated with a varied choice of ρ. We ran our experiment 200 times. Each time we consider the case when the indices are ranked by *lfdr*, which is the ideal choice for exchangeable tests, and we select the indices with the lowest two values. We divide the hypotheses into the two blocks in the second situation, known to the oracle as if she knew the variance-covariance matrix structure. Further, the top indices from the two blocks with the minimum lfdr values are selected. We find this the most intuitive way to counter the dependencies within blocks and potentially make a better rejection due to the blocks' positive dependencies. Table 4 reports the percentage of time the lfdr statistic's ranking provided higher tail probability values than when ranked by the lfdr values within blocks separately. The values are reported for varied values of ρ in percentage for 200 experimental runs. Consider

the case of $\rho = 0.3$. Table 4 shows that in 13% of the cases selecting the top indices from the two blocks separately produces a lower value of the tail probability than selecting the top indices from all the indices combined.

Table 4: Counterexample to the optimal *lfdr* ranking

The table reports the proportion of time the tail probability was lowered when not necessarily ranked the lfdr. The ranking may be substantially improved for moderately correlated test statistics by incorporating the correlation structure into account. This demonstrates that the lfdr ranking and thresholding procedure may not be optimal for non-exchangeable tests where the correlation structure is unknown to a practitioner.

ρ	0.01	0.1	0.3	0.5	0.7	0.9
Contradicts (in %)	0	4.5	13	16	21.5	31

One might also wonder if the differences in the tail probabilities are high enough. For example, consider $\rho = 0.5$: the tail probability is relatively higher by an average of 16.7%.

5.3 Simulation of Stock Returns Trading Strategies

In this section we present a simulated version of the empirical application presented in Section 6 that follows the set up of Chordia et al. (2020). A total of $N = 2,000$ stock returns are generated for $T = 500$ periods. Returns follow a linear factor structure

$$R_{it} = \alpha_i + \beta_i' F_t + \varepsilon_{it},$$

where the factors, F, are drawn from a multivariate normal distribution with the mean and the covariance matrix matching closely those of the five Fama and French (2015) factor model, augmented with Carhart (1997) momentum. For each stock, α represents the return that an investor could realize in excess of the risk generated by the factors, and is drawn from a $\mathcal{N}(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 15.1\%$. In each time period (i.e., month), we draw $S = 5,000$ trading signals for each stock: a fraction π of the trading signals are informative (although imperfectly) about the α of each stock: $s_{it} = \alpha_i + \eta_{it}$, where $\eta_{it} \sim \mathcal{N}(0, \sigma_\eta)$. A fraction $1 - \pi$ just contain noise: $s_{it} = \eta_{it}$. Informative and uninformative trading signals might share some common noise through the correlation coefficient ρ_η among signals.

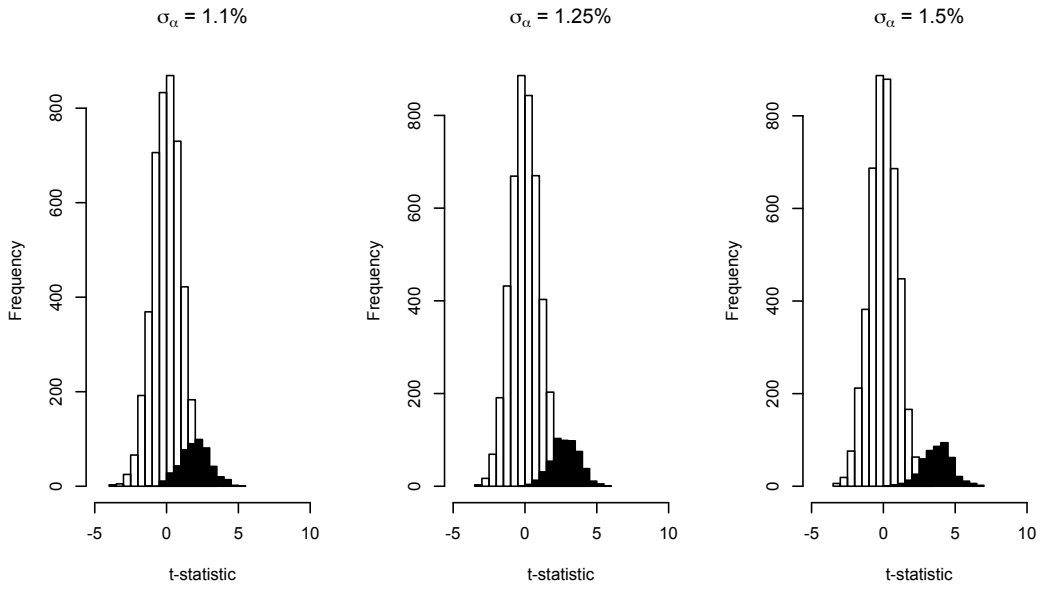
A stock trader constructs 5,000 trading strategies by sorting stocks each month based on the signals realizations. She forms 10 portfolios, buys the portfolio corresponding to the largest signals, and shorts the portfolio corresponding to the lowest. The performance of these 5,000 long-short portfolios is evaluated by regressing the time series of 500 portfolio return observations on the realizations of the factors. A t-statistics on the estimated regression intercept (i.e., the portfolio alpha/abnormal return) is used to evaluate each strategy in classical hypothe-

ses testing, and altogether in multiple hypothesis testing. The entire simulation procedure is repeated 1,000 times.

Figure 3 provides a visual representation of the null and alternative distributions in three scenarios that differ for the strength of the informative signal, $\sigma_\alpha \in \{1.1\%, 1.25\%, 1.5\%\}$. As the signal becomes stronger, the truly informative trading signals are more easily identifiable. Note, however, that even the most favorable scenario is rather complicated for a multiple comparisons procedure as there is substantial overlap between null and alternative distributions.

Figure 3: Simulation scenarios

The figure presents a visual comparison of different distributions of test statistics for the trading strategy simulation. In the base case situation: $\pi = 0.1$, $\sigma_\varepsilon = 0.15$, $\sigma_\eta = 0.2$, and $\rho_\eta = 0$. The figure presents different simulations that vary $\sigma_\alpha \in \{1.1\%, 1.25\%, 1.5\%\}$.



We analyze the simulated data under two scenarios that differ for whether the econometrician knows the true proportion of nulls and the parameters of the null distribution (Oracle) or has to estimate that from the data (Data-Driven). Note that here, differently from the case described in Section 5.1 where the null is a $\mathcal{N}(0,1)$, the exact null distribution is unknown. Thus, constructing the Oracle information set in a non-traditional way: We allow the Oracle to estimate the parameters of the null distribution by temporarily endowing her with the knowledge of which tests are null and which are not. Based on that, the Oracle can estimate the null and alternative parameters, after which she forgets to have ever known which tests are null and tries to learn that by running the various procedures. In the Data-Driven scenario, the econometrician does not know the structural parameters and estimates them all.

We report results in Table 5 and Table 6, respectively. Both tables compare average FDX,

FDR, and power of Procedure 2 to Sun and Cai (2007) (SC), Benjamini and Hochberg (1995) (BH), Guo and Romano (2007) (GR), and Lehmann and Romano (2005) (LR). Both tables consider different specifications of the standard deviation of α, $\sigma_\alpha \in \{1.1\%, 1.25\%, 1.5\%\}$, which determines the signal to noise ratio in the trading signal, and the trading signals pairwise correlation coefficient, $\rho_\eta \in \{0, 0.1, 0.2\}$. A higher σ_α makes the signals more informative. Similarly, a non-zero correlation makes the signal to noise ratio higher, by reducing the background noise, and places procedures outside of the canonical i.i.d case.

Almost in every scenario considered Procedure 2 sits between BH and GR, in terms of FDX and power. From Table 5 we see that the Oracle version is able to maintain and effective, although a bit conservative, FDX control while delivering reasonable power, especially in the very difficult scenario (i.e., $\sigma_\alpha = 1.1\%$). The Procedure outperforms GR, its most natural comparison, in all but one scenario: when $\sigma_\alpha = 1.1\%$ GR delivers a slightly higher FDX. Although it is designed to control FDR, as opposed to FDX, the most applied MHT procedure is BH. Relative to Procedure 2, BH does guarantee FDR control while maintaining a higher power, but at the cost of FDX of relatively large FDX, often over 30%. Increasing correlation among tests makes it easier to separate null and alternative, thus leading to an increase in power for all procedures: thus even in the case of simple correlation structure which maintains exchangeability, Procedure 2 is still able to accomplish its objectives of delivering FDX control and outpacing the power of other procedures designed in the same spirit.

Once we place the procedures in the real-life scenario of having to learn the data-generating process from the data, see Table 6, both Procedure 2 ($lfdr$) and GR deliver an FDX control that is above the desired threshold in every scenario, with Procedure 2 erring on the side of being less conservative. This is likely because estimating π from the simulated data is particularly challenging. A way to simplify the problem is to start with the assumption that the proportion of informative signal is zero, $\hat{\pi} = 0$. That leads to a substantial decrease in FDX, which nears the desired threshold, coming at a slight power reduction expense.

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 et al. (2020). The construction of trading strategies reflects exactly the simulation set up of Section 5.3. 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 purchased at the closing price, and stocks in the bottom decile are sold short. Portfolio compositions are held for twelve 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 COMPUSTAT/CRSP datasets: we take the level, the growth rate, the ratio

Table 5: Stock return simulation (Oracle)

The table compare average FDX, FDR, and power across 5,000 trading strategies obtained from Procedure 2 to Sun and Cai (2007) (SC), Benjamini and Hochberg (1995) (BH), Guo and Romano (2007) (GR), and Lehmann and Romano (2005) (LR). In each simulations, the return of 2,000 stocks for 500 observations is generated from a linear factor structure model, $R_{it} = \alpha_i + \beta_i' F_t + \varepsilon_{it}$, where α_i represents the return that an investor could realize in excess of the risk generated by the factors, F_t and is drawn from a $\mathcal{N}(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In each time period (i.e., month), we draw $S = 5,000$ trading strategies for each stock: a fraction π of the trading signals are informative (although imperfectly) about the α of each stock: $s_{it} = \alpha_i + \eta_{it}$, where $\eta_{it} \sim \mathcal{N}(0, \sigma_\eta)$. A fraction $1 - \pi$ just contain noise: $s_{it} = \eta_{it}$. Informative and uninformative trading signals might share some common noise through the correlation coefficient ρ_η. Each month stocks are sorted into deciles based on the trading signal realization, and a long short portfolio is obtained from buying stocks in the top deciles and shorting stocks in the bottom deciles. A t-statistics of the portfolio regression alpha serves as the relevant test statistic in evaluating each trading strategy/long-short portfolio. In the base case situation: $\pi = 0.1$, $\sigma_\varepsilon = 0.15$, $\sigma_\eta = 0.2$. The table presents different simulations that vary $\sigma_\alpha \in \{1.1\%, 1.25\%, 1.5\%\}$ and $\rho_\eta \in \{0, 0.1, 0.2\}$. Each procedure is implemented as an Oracle, who has knowledge, or is able to accurate estimate, the parameters of the data generating process. FDX control is implemented at $\gamma = 0.05$ and with a confidence level $1 - \alpha = 0.95$, FDR control is implemented at a nominal level of $\alpha = 0.05$.

		SC	BH	GR	LR	Procedure 2
$\sigma_\alpha = 1.5\%$	FDX	0.471	0.289	0.034	0.000	0.039
$\rho_\eta = 0$	FDR	0.050	0.045	0.035	0.003	0.037
	Power (%)	92.9	92.5	91.0	68.4	91.3
$\sigma_\alpha = 1.25\%$	FDX	0.493	0.340	0.032	0.003	0.049
$\rho_\eta = 0$	FDR	0.051	0.046	0.028	0.002	0.032
	Power (%)	52.7	50.9	41.8	12.7	44.4
$\sigma_\alpha = 1.10\%$	FDX	0.477	0.382	0.045	0.045	0.041
$\rho_\eta = 0$	FDR	0.050	0.045	0.006	0.006	0.018
	Power (%)	18.9	17.6	2.2	2.0	8.6
$\sigma_\alpha = 1.25\%$	FDX	0.500	0.341	0.031	0.000	0.041
$\rho_\eta = 0.1$	FDR	0.050	0.045	0.029	0.003	0.032
	Power (%)	57.5	55.9	47.8	15.8	49.8
$\sigma_\alpha = 1.25\%$	FDX	0.483	0.334	0.031	0.000	0.044
$\rho_\eta = 0.2$	FDR	0.050	0.046	0.030	0.003	0.033
	Power (%)	62.3	60.8	53.6	19.6	55.3

Table 6: Stock return simulation (Data-Driven)

The table compare average FDX, FDR, and power across 5,000 trading strategies obtained from Procedure 2 to Sun and Cai (2007) (SC), Benjamini and Hochberg (1995) (BH), Guo and Romano (2007) (GR), and Lehmann and Romano (2005) (LR). In each simulations, the return of 2,000 stocks for 500 observations is generated from a linear factor structure model, $R_{it} = \alpha_i + \beta_i' F_t + \varepsilon_{it}$, where α_i represents the return that an investor could realize in excess of the risk generated by the factors, F_t and is drawn from a $\mathcal{N}(0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. In each time period (i.e., month), we draw $S = 5,000$ trading signals for each stock: a fraction π of the trading signals are informative (although imperfectly) about the α of each stock: $s_{it} = \alpha_i + \eta_{it}$, where $\eta_{it} \sim \mathcal{N}(0, \sigma_\eta)$. A fraction $1 - \pi$ just contain noise: $s_{it} = \eta_{it}$. Informative and uninformative trading signals might share some common noise through the correlation coefficient $\rho_\eta = corr(\eta_i, \eta_j)$. Each month stocks are sorted into deciles based on the trading signal realization, and a long short portfolio is obtained from buying stocks in the top deciles and shorting stocks in the bottom deciles. A t-statistics of the portfolio regression alpha serves as the relevant test statistic in evaluating each trading strategy/long-short portfolio. In the base case situation: $\pi = 0.1$, $\sigma_\varepsilon = 0.15$, $\sigma_\eta = 0.2$. The table presents different simulations that vary $\sigma_\alpha \in \{1.1\%, 1.25\%, 1.5\%\}$ and $\rho_\eta \in \{0, 0.1, 0.2\}$. Each procedure is implemented in a Data-Driven fashion: when necessary the econometrician estimates parameters directly from the observed data without knowledge of the true data-generating process. FDX control is implemented at $\gamma = 0.05$ and with a confidence level $1 - \alpha = 0.95$, FDR control is implemented at a nominal level of $\alpha = 0.05$.

						Procedure 2	
		SC	BH	GR	LR	$lfdr$	$lfdr(\hat{\pi} = 0)$
$\sigma_\alpha = 1.5\%$	FDX	0.470	0.322	0.074	0.000	0.082	0.033
$\rho_\eta = 0$	FDR	0.050	0.046	0.036	0.003	0.037	0.034
	Power (%)	92.9	92.5	91.0	68.4	91.3	90.6
$\sigma_\alpha = 1.25\%$	FDX	0.593	0.398	0.073	0.002	0.123	0.069
$\rho_\eta = 0$	FDR	0.054	0.048	0.030	0.003	0.035	0.031
	Power (%)	53.6	51.6	42.7	13.2	45.5	43.4
$\sigma_\alpha = 1.1\%$	FDX	0.668	0.572	0.074	0.069	0.111	0.078
$\rho_\eta = 0$	FDR	0.063	0.057	0.008	0.007	0.026	0.022
	Power (%)	22.1	20.4	2.8	2.3	11.1	9.6
$\sigma_\alpha = 1.25\%$	FDX	0.569	0.403	0.072	0.000	0.109	0.051
$\rho_\eta = 0.1$	FDR	0.053	0.047	0.030	0.003	0.035	0.031
	Power (%)	58.4	56.4	48.6	16.3	50.8	48.9
$\sigma_\alpha = 1.25\%$	FDX	0.556	0.410	0.077	0.000	0.120	0.071
$\rho_\eta = 0.2$	FDR	0.053	0.048	0.032	0.003	0.035	0.032
	Power (%)	63.0	61.3	54.3	20.1	56.1	54.2

between two variables, and a transformation of three variables (i.e., (x1 - x2)/x3). When more than one variable is involved, we consider all the possible combinations. We apply filters to guarantee that strategies are well populated and that microstock returns are not overly biased. Details can be found in Chordia et al. (2020). In total, we obtain 2,396,456 trading strategies for the period between 1972 and 2015.

The data generating process provides that asset (stocks or portfolios) returns arise because of three components: a systematic risk-premium, an idiosyncratic mean-zero, and a time-invariant component (α_i). Under the null, the returns are entirely due to compensation for exposure to systematic risk factors; the time-invariant component is precisely zero (i.e., $\alpha_i = 0$). We test whether the portfolio α_i is zero (i.e., this is a two-tail test) for 2,396,456 strategies. Thus we have a standard multiple testing problem.

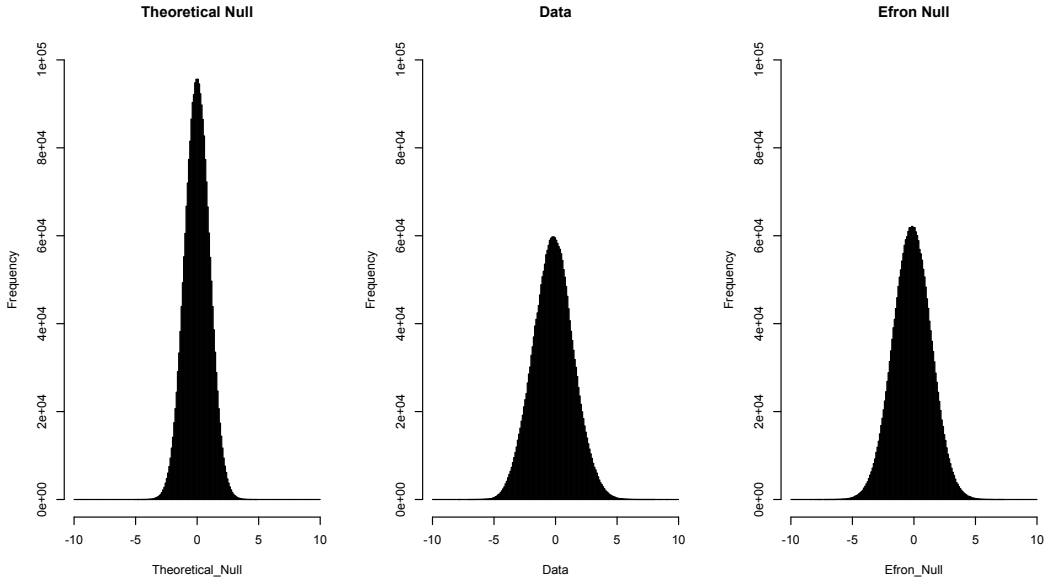
The rationale behind adopting a combinatorial approach to constructing trading signals is essentially twofold. On the one hand, there is a long tradition among finance scholars and practitioners to relate stock returns to accounting variables: quantities such as the equity market value of a firm (i.e., 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 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 procedure. Adopting the combinatorial approach to generating trading signals is not without consequences: we will uncover many trading strategies that appear to be very profitable, but which likely are also meaningless, in the sense that they are artifacts of our data mining, are not based on any reasonable economic argument, and are likely not going to be profitable out of sample.

As mentioned above, this set of trading strategies has been studied in Chordia et al. (2020). The authors apply several multiple hypotheses procedures and still "find" many profitable strategies before applying some economic restriction. Probably a problematic aspect of their study is that they fail to incorporate the information gained from the data about the null distribution into the procedures. Thus, this particular data set seems perfect to evaluate Procedure 2, which heavily relies on ranking hypotheses by the data-driven $lfdr$ (i.e., the one that relies on the empirical null): 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.

Figure 4: Data representation under different assumptions

The figure compares histograms of the distribution of t-statistics for 2,396,456 trading strategies. The left panel is created by drawing a 90% of 2,396,456 from a $\mathcal{N}(0,1)$, the center panel presents the histogram of the actual data, and the rightmost panel is the density of the estimated empirical null. We use the analytical method of estimating the empirical null distribution parameters and null proportion described in Section 4 of Efron (2008). The data contains 2,396,456 trading strategies for the period between 1972 and 2015.



In Figure 4 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 widespread 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 the 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., a few genuinely non-zero alphas). How much of that might be in the data is a question that can only be answered by correcting for multiple hypotheses.

We implement Procedure 2 and report the number of strategies that are selected at different levels of γ and α in Table 7, where γ denotes the maximum allowable proportion of false discoveries (FDP), and α refers to the allowable tail probability. Similar to what we do for the

simulation exercise described in Section 5.3, we compare the results obtained from applying Procedure 2 to the one proposed by Guo and Romano (2007) (GR), the FDP-StepM procedure of Romano and Wolf (2007) and Romano et al. (2008) (RSW), and the FDR procedure of Sun and Cai (2007) (SC).

Table 7: Discoveries in a sample of 2,396,456 trading strategies

The table reports the number of trading strategies selected by Procedure 2, Guo and Romano (2007) (GR), the FDP-StepM procedure of Romano and Wolf (2007), and Romano et al. (2008) (RSW), and the FDR procedure of Sun and Cai (2007) (SC) when applied to the set of 2,396,456 stock trading strategies constructed from accounting and stock price information for the period between 1972 and 2015.

Panel A: Procedures based on empirical null						
	Procedure 2			GR		
γ/α	0.01	0.05	0.10	0.01	0.05	0.10
0.05	1	4	5	1	2	3
0.10	1	20	24	1	2	3
0.20	40	47	51	1	2	3
Panel B: Procedures based on theoretical null						
	Procedure 2			GR		
γ/α	0.01	0.05	0.10	0.01	0.05	0.10
0.05	253,837	254,894	255,454	204,011	205,113	205,654
0.10	416,217	417,137	417,627	351,377	352,334	352,907
0.20	698,052	698,881	699,322	626,247	627,418	628,007
Panel C: Alternative procedures						
	RSW			SC		
γ/α	0.01	0.05	0.10	0.01	0.05	0.10
0.05	5,528	32,812	65,001	290,932	411,855	549,435
0.10	21,867	90,722	15,1614	446,077	549,435	673,698
0.20	96,241	235,708	328,007	722,506	808,424	915,176

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 a $\gamma = 0.2$ 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.

Compared to the standard frequentist procedure of Guo and Romano (2007), which emerges from our simulation as the most powerful alternative solution in the frequentist's paradigm, Procedure 2 selects more strategies. 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 and $\gamma = 0.10$ 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 thousand strategies, while GR selects 352 thousand. This is understandable as Efron's method imposes a much less stringent condition on the null specification. By conforming to the data, it 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, which aims to control FDP at the same levels of γ and α; still, it is very conservative, and the SC which controls FDR at a level $\alpha + (1 - \alpha)\,\gamma$. The exercise reinforces the idea that the application of a multiple comparisons procedure to a vast number of tests can give questionable answers when all the information in the data is not taken into account. 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 paper, which relies on the *local false discovery rate* as one of its primary inputs.

7 Discussion

The proposed method is an (γ, α)-level FDX control method, which provides an instrumental framework for experiments run only once or a few times. Unlike FDR methods, which only offer a long-run average guarantee, FDX methods provide a high-probability control for individual experiments. Under an empirical Bayes framework, this method uses the Poisson binomial distribution to theoretically guarantee that the probability of false discoveries exceeding γ proportion is no more than α. There remain several open issues to address as the scope of this work progresses.

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 γ. Such exciting and highly relevant error control is left for future explorations.

An essential question for FDX and FDR methods based on the empirical Bayes framework is how to estimate the *lfdr* statistic in practice. This paper 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.

Also, several extensions may be worth investigating for this FDX method. For example, it may be possible to incorporate auxiliary information or develop a weighted FDX method for asymmetric hypotheses. Applied researchers may also need to design FDX controlling methods for discrete test statistics. These are interesting problems that we leave for future development. Another line of future work will be to construct FDX controlling procedures by

modeling dependencies within the tests using the hidden Markov model (HMM), following the work of Sun and Cai (2009) and more recently by Perrot-Dockès et al. (2021).

References

Basu, P. (2016). *Model selection principles and false discovery rate control.* Ph. D. thesis, University of Southern California.

Basu, P., T. T. Cai, K. Das, and W. Sun (2018). Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association 113*(523), 1172–1183. https://doi.org/10.1080/01621459.2017.1336443.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Blanchard, G., P. Neuvial, and E. Roquain (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics 48*(3), 1281–1303. https://doi.org/10.1214/19-aos1847.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82. https://doi.org/10.2307/2329556.

Chen, A. and T. Zimmermann (2021). Open-source cross-sectional asset pricing. *Critical Finance Review*, Forthcoming. https://doi.org/10.2139/ssrn.3604626.

Chen, S. X. and J. S. Liu (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica 7*, 875–892.

Chi, Z. and Z. Tan (2008). Positive false discovery proportions: Intrinsic bounds and adaptive control. *Statistica Sinica 18*(3), 837–860.

Chordia, T., A. Goyal, and A. Saretto (2020). Anomalies and false rejections. *The Review of Financial Studies 33*(5), 2134–2179. https://doi.org/10.1093/rfs/hhaa018.

Delattre, S. and E. Roquain (2015). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *The Annals of Statistics 43*(3), 1141–1177. https://doi.org/10.1214/14-aos1302.

Döhler, S. and E. Roquain (2020). Controlling the false discovery exceedance for heterogeneous tests. *Electronic Journal of Statistics 14*(2), 4244–4272. https://doi.org/10.1214/20-ejs1771.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association 99*(465), 96–104. https://doi.org/10.1198/016214504000000089.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science 23*(1), 1–22. https://doi.org/10.1214/07-sts236.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association 104*(487), 1015–1028. https://doi.org/10.1198/jasa.2009.tm08523.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*(456), 1151–1160. https://doi.org/10.1198/016214501753382129.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research 17*(4), 347–388. https://doi.org/10.1177/0962280206079046.

Farcomeni, A. (2009). Generalized augmentation to control the false discovery exceedance in multiple testing. *Scandinavian Journal of Statistics 36*(3), 501–517. https://doi.org/10.1111/j.1467-9469.2008.00633.x.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631. https://doi.org/10.1198/016214502760047131.

Fu, L. (2018). *Nonparametric empirical Bayes methods for large-scale inference under heteroscedasticity*. Ph. D. thesis, University of Southern California.

Genovese, C. and L. Wasserman (2004). A stochastic process approach to false discovery control. *The Annals of Statistics 32*(3), 1035–1061. https://doi.org/10.1214/009053604000000283.

Genovese, C. R. and L. Wasserman (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association 101*(476), 1408–1417. https://doi.org/10.1198/016214506000000339.

Goeman, J. J., J. Hemerik, and A. Solari (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics 49*(2), 1218–1238. https://doi.org/10.1214/20-aos1999.

Gordon, A. Y. and P. Salzman (2008). Optimality of the Holm procedure among general step-down multiple testing procedures. *Statistics & Probability Letters 78*(13), 1878–1884. https://doi.org/10.1016/j.spl.2008.01.055.

Gu, J. and R. Koenker (2020). Invidious comparisons: Ranking and selection as compound decisions. *arXiv preprint arXiv:2012.12550*.

Guo, W. and J. Romano (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology 6*(1), 1–35. https://doi.org/10.2202/1544-6115.1247.

Heller, R. and S. Rosset (2021). Optimal control of false discovery criteria in the two-group model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 83*(1), 133–155. https://doi.org/10.1111/rssb.12403.

Hemerik, J., A. Solari, and J. J. Goeman (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika 106*(3), 635–649. https://doi.org/10.1093/biomet/asz021.

Katsevich, E. and A. Ramdas (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics 48*(6), 3465–3487. https://doi.org/10.1214/19-aos1938.

Korn, E. L., J. F. Troendle, L. M. McShane, and R. Simon (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*(2), 379 – 398. https://doi.org/10.1016/s0378-3758(03)00211-8.

Lehmann, E. L. and J. P. Romano (2005). Generalizations of the familywise error rate. *The Annals of Statistics 33*(3), 1138–1154. https://doi.org/10.1214/009053605000000084.

Perone Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *Journal of the American Statistical Association 99*(468), 1002–1014. https://doi.org/10.1198/0162145000001655.

Perrot-Dockès, M., G. Blanchard, P. Neuvial, and E. Roquain (2021). Post hoc false discovery proportion inference under a hidden Markov model. *arXiv preprint arXiv:2105.00288*.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008). Formalized data snooping based on generalized error rates. *Econometric Theory 24*, 404–447. https://doi.org/10.1017/s0266466608080171.

Romano, J. P. and M. Wolf (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics 35*(4), 1378–1408. https://doi.org/10.1214/009053606000001622.

Roquain, E. and N. Verzelen (2021). False discovery rate control with unknown null distribution: Is it possible to mimic the oracle? *The Annals of Statistics*, Forthcoming.

Roquain, E. and F. Villers (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *The Annals of Statistics 39*(1), 584–612. https://doi.org/10.1214/10-aos847.

Shaked, M. and J. G. Shanthikumar (2007). Stochastic orders. *Springer Science & Business Media*.

Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association 102*(479), 901–912. https://doi.org/10.1198/016214507000000545.

Sun, W. and T. T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 393–424. https://doi.org/10.1111/j.1467-9868.2008.00694.x.

van der Laan, M. J., M. D. Birkner, and A. E. Hubbard (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), 1–32. https://doi.org/10.2202/1544-6115.1143.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (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), 1–27. https://doi.org/10.2202/1544-6115.1042.