Online false discovery rate control for LORD++ and SAFFRON under positive, local dependence

Aaron Fisher

Foundation Medicine Inc., Cambridge, Massachusetts, USA

Correspondence
Aaron Fisher, Foundation Medicine Inc., Cambridge, MA, USA.
Email: afishe27@alumni.jh.edu

Abstract
Online testing procedures assume that hypotheses are observed in sequence, and allow the significance thresholds for upcoming tests to depend on the test statistics observed so far. Some of the most popular online methods include alpha investing, LORD++, and SAFFRON. These three methods have been shown to provide online control of the “modified” false discovery rate (mFDR) under a condition known as CS. However, to our knowledge, LORD++ and SAFFRON have only been shown to control the traditional false discovery rate (FDR) under an independence condition on the test statistics. Our work bolsters these results by showing that SAFFRON and LORD++ additionally ensure online control of the FDR under a “local” form of nonnegative dependence. Further, FDR control is maintained under certain types of adaptive stopping rules, such as stopping after a certain number of rejections have been observed. Because alpha investing can be recovered as a special case of the SAFFRON framework, our results immediately apply to alpha investing as well. In the process of deriving these results, we also formally characterize how the conditional super-uniformity assumption implicitly limits the allowed p-value dependencies. This implicit limitation is important not only to our proposed FDR result, but also to many existing mFDR results.

KEYWORDS
batch testing, FDR control, p-values, sequential experimentation, sequential hypothesis testing

1 INTRODUCTION

A typical goal in multiple hypotheses testing is to limit the probability of any false discoveries (the familywise error rate, or FWER), or the false discovery rate (FDR; Benjamini & Hochberg, 1995), to be no less than a constant denoted by $\alpha$. Unfortunately, many traditional testing procedures require all hypotheses to be prespecified before a sequence of experiments begins, and require all data to be collected before formal conclusions can be made. This limits scientists’ ability to take early action in response to preliminary results. For example, if early clinical trial data suggest that a new drug is beneficial in a certain subpopulation, it would be desirable to be able to reject the null hypothesis in that subgroup before the remainder of the trial concludes (see, e.g., Zehetmayer et al., 2022). Similar situations arise when analysts must react quickly in response to data streams representing potential threats, such as when monitoring a series of transactions for credit card fraud (Zrnic et al., 2020). One simple strategy for early decision making is to “budget” or “spend” a predefined...
fraction of the allowed FWER on each planned test, so that the sum of thresholds used across all tests is equal to \( \alpha \). However, in scenarios where the total number of experiments that a lab will conduct is uncertain, budgeting the alpha level across all hypotheses may not be feasible.

**Online** testing procedures address these problems by allowing researchers to test hypotheses associated with a (possibly infinite) sequence of experiments. Early work in this field was pioneered by Foster and Stine (2008), who proposed an online method known as *alpha investing*. This method allows the rejection thresholds for future tests to depend on the statistics observed so far. Each rejected hypothesis replenishes the available alpha budget, or “alpha wealth,” allowing users to continue testing indefinitely.

Alpha investing formed the inspiration for a wave of new online methods. Aharoni and Rosset (2014) extend this idea to “generalized” alpha investing (GAI), and Ramdas et al. (2017) propose a version of GAI with improved power (GAI++). The latter class of methods includes a special case known as “LORD++,” which is based on the “LORD” (significance Levels based On Recent Discovery) method developed by Javanmard and Montanari (2018). Building on LORD++, Ramdas et al. (2018) develop a similar method that incorporates adaptive estimates of the proportion of true null hypotheses, referred to as “SAFFRON” (Serial estimate of the Alpha Fraction that is Futility Rationed On true Null hypotheses). Due to their relatively high power, LORD++ and SAFFRON are often described as being among the state-of-the-art methods available for online testing (Chen & Kasiviswanathan, 2020; Zhang et al., 2020). The concepts underlying these two methods have also influenced several related online approaches (Tian & Ramdas, 2019; Weinstein & Ramdas, 2020; Xu & Ramdas, 2020; Zrnic et al., 2020, 2021). Moreover, Ramdas et al. (2018) show that SAFFRON encompasses the original alpha investing method as a special case.

Existing theoretical studies of alpha investing, LORD++, and SAFFRON typically require a condition known as *conditional super-uniformity* (CS), which states that each null \( p \)-value is stochastically larger than a uniform random variable, conditional on the information used to specify its rejection threshold (Zrnic et al., 2021; see also Foster & Stine, 2008; Ramdas et al., 2017, 2018). Thus, each \( p \)-value in the sequence may be “locally dependent” with its neighbors, so long as its associated rejection threshold is specified sufficiently far in advance (see details in Section 2.2).

A key use case where local dependence holds is when estimating the effects of several treatments in several subpopulations. Here, the resulting \( p \)-value sequence follows local dependence due to the fact that tests conducted in one population are independent of tests conducted in another. Similar forms of local dependence hold, at least approximately, when the test statistics are associated with a temporal process with an autoregressive correlation structure.

Under the CS assumption, LORD++ and SAFFRON have been shown to control a “modified” version of the false discovery rate (mFDR) (Ramdas et al., 2017, 2018; Zrnic et al., 2021). To our knowledge however, existing results for online control of the traditional FDR (Benjamini & Hochberg, 1995) require an additional assumption of \( p \)-value independence (Ramdas et al., 2017, 2018; Zrnic et al., 2021).

Our primary contribution is to show that LORD++ and SAFFRON also ensure online control of the FDR under a positive, local dependency condition on the \( p \)-values. In particular, they control FDR whenever (1) the null \( p \)-values are CS given the information used to define their testing thresholds (as in Zrnic et al., 2021); (2) the null \( p \)-values follow the conventional assumption of positive regression dependence on a subset (PRDS; Benjamini & Yekutieli, 2001), and (3) users choose significance thresholds that are monotonically nonincreasing in the \( p \)-values that have been observed so far. Because alpha investing can be written as a special case, the same results immediately apply to alpha investing as well. As a secondary contribution, we show that FDR control is also possible under weakened forms of monotonicity, which allows us to account for certain forms of adaptive stopping times.

Finally, we introduce an important caveat regarding the CS assumption, which is relevant to many results in the literature beyond our own (e.g., Aharoni & Rosset, 2014; Fithian & Lei, 2022; Foster & Stine, 2008; Ramdas et al., 2017, 2018; Rebjock et al., 2021; Zhang et al., 2020; Zrnic et al., 2021). For well-powered \( p \)-value sequences, we show that CS can only be satisfied under local dependence. To our knowledge, this is the first formal characterization of CS as a dependence requirement.

The remainder of this paper is organized as follows. In Section 2, we introduce relevant notation, summarize LORD++ and SAFFRON, and present existing results for FDR control. In the process, we introduce the CS assumption and show how it implicitly restricts \( p \)-value dependencies. Section 3 contains our main result, and discusses variations on the required positive dependence and monotonicity conditions. Section 4 illustrates our main result in a series of simulations. All proofs are contained in the Supporting Information.
Following Foster and Stine (2008), we consider the setting where analysts observe a (possibly infinite) sequence of hypotheses \( \{H_1, H_2, \ldots\} \), along with corresponding p-values \( \{P_1, P_2, \ldots\} \). Let \( H_0 \subseteq \mathbb{N} \) be the subset of indices corresponding to the truly null hypotheses. At each stage \( t \) of testing, the researcher observes \( P_t \) and must decide to reject or not reject \( H_t \) before observing the next test statistic \( P_{t+1} \).

Let \( \alpha_t \) be the significance threshold used for testing \( H_t \), where \( H_t \) is rejected whenever \( P_t \leq \alpha_t \). In order to capture different types of adaptive decision making, Foster and Stine (2008) allow \( \alpha_t \) to be a function of the preceding p-values \( \{P_i\}_{i<t} \). To highlight that \( \alpha_t \) may only depend on some summary function of \( \{P_i\}_{i<t} \), we use \( I_t \) to denote the information used to determine the threshold \( \alpha_t \), where \( I_t \) must be a set of real-valued random variables. For example, if a decision rule defines \( \alpha_t \) only as a function of which hypotheses have been rejected so far, then \( I_t = \{1(P_i < \alpha_i)\}_{i<t} \). In this framing, an online testing method is essentially a set of rules for translating previous p-values into thresholds for the future tests. Each threshold \( \alpha_t \) is a random variable, but must be fully determined given \( I_t \).

Let \( R_t = \{i \leq t : P_i \leq \alpha_i\} \) be indices for the hypothesis that are rejected by the \( t \)th stage of testing. At each stage \( t \), the false discovery proportion (FDP) and the FDR are defined, respectively, as

\[
\text{FDP}(t) = \frac{|H_0 \cap R_t|}{1 \lor |R_t|} \quad \text{and} \quad \text{FDR}(t) = E \left[ \frac{|H_0 \cap R_t|}{1 \lor |R_t|} \right],
\]

where \((a \lor b)\) is shorthand for the maximum over \([a, b]\). Similarly, the “modified” FDR is defined as

\[
\text{mFDR}(t) = \frac{E [ |H_0 \cap R_t| ]}{E [ 1 \lor |R_t| ]}.
\]

### 2.1 LORD++ AND SAFFRON approaches

Ramdas et al. (2017) suggest choosing the thresholds \( \{\alpha_i\}_{i=1}^\infty \) in a way that ensures that an empirical estimate of the FDP never exceeds \( \alpha \). The authors estimate the FDP as

\[
\hat{\text{FDP}}_0(t) = \frac{\sum_{i \leq t} \alpha_i}{1 \lor |R_t|}
\]

and propose a specific algorithm (LORD++) for defining thresholds \( \{\alpha_i\}_{i=1}^\infty \) so that \( \hat{\text{FDP}}_0(t) \leq \alpha \) for all \( t \in \mathbb{N} \). A key attribute of LORD++ is that it is a monotonic algorithm, meaning that rejecting more p-values early on can only lead to higher testing thresholds later on (see Theorem 2 for a formal definition). Thus, observing a stronger signal in early tests can only improve power for later tests.

Intuitively, we may expect that constraining \( \hat{\text{FDP}}_0(t) \leq \alpha \) will result in a small FDP, since

\[
\hat{\text{FDP}}_0(t) \geq \frac{\sum_{i \leq t : i \in H_0} \alpha_i}{1 \lor |R_t|} \approx \sum_{i \leq t : i \in H_0} 1(P_i \leq \alpha_i) \frac{1}{1 \lor |R_t|} = \text{FDP}(t).
\]

Thus, \( \hat{\text{FDP}}_0(t) \) is an approximate overestimate of FDP(\( t \)), and so we can intuit that controlling \( \hat{\text{FDP}}_0(t) \) will result in conservative control for FDR(\( t \)). Indeed, Ramdas et al. (2017) show that, under certain conditions, LORD++ guarantees both \( \text{mFDR}(t) \leq \alpha \) and \( \text{FDR}(t) \leq \alpha \) for all \( t \in \mathbb{N} \). Importantly, these results apply not only to the specific LORD++ algorithm, but to any monotonic algorithm satisfying \( \hat{\text{FDP}}_0(t) \leq \alpha \) for all \( t \) (see details in Theorem 2). Since \( \hat{\text{FDP}}_0(t) \) is directly observable, the requirement that \( \hat{\text{FDP}}_0(t) \leq \alpha \) is straightforward to implement.
Building on this idea, Ramdas et al. (2018) suggest controlling an alternative estimate of FDP(t) that is expected to be less conservative. Leveraging strategies from Storey (2002) and Storey et al. (2004), Ramdas et al. propose the estimator

\[
\widetilde{FDP}_\lambda(t) = \sum_{i \leq t} \alpha_i \frac{1_{(P_i > \lambda_i)}}{1 - \lambda_i} \frac{1 - \lambda_i}{1 \vee |R_i|},
\]

(2)

where \(\{\lambda_i\}_{i=1}^{\infty}\) is a series of user-defined constants within the interval (0, 1). Like \(\alpha_i\), each \(\lambda_i\) is required to be a deterministic function of \(I_t\). The intuition of \(\widetilde{FDP}_\lambda(t)\) is that \(1_{(P_i > \lambda_i)}/(1 - \lambda_i)\) has expectation lower bounded by \(1\) when \(i \in \mathcal{H}_0\), but has a smaller expectation when \(i \notin \mathcal{H}_0\). Thus, the numerator in Equation (2) will ideally have an expectation close to \(\sum_{i \leq t : I_t} \alpha_i\), and so \(\widetilde{FDP}_\lambda(t)\) will ideally resemble right-hand side of Line (1). In simulations, Ramdas et al. generally found setting \(\lambda = 1/2\) to produce algorithms with relatively high power.

Ramdas et al. develop a monotonic algorithm known as SAFFRON that assigns the threshold parameters \(\{\alpha_i\}_{i=1}^{\infty}\) and tuning parameters \(\{\lambda_i\}_{i=1}^{\infty}\) such that \(\widetilde{FDP}_\lambda(t)\) is constrained to be no more than \(\alpha\) at all stages \(t \in \mathbb{N}\). The authors show that, under certain conditions, any monotonic algorithm satisfying \(\widetilde{FDP}_\lambda(t) \leq \alpha\) for all \(t \in \mathbb{N}\) (including SAFFRON) controls mFDR and FDR. We review the required conditions in the next two sections.

2.2 Conditional super-uniformity

Many online testing procedures are based on an assumption known as CS. Formally, we say that a p-value \(P_t\) satisfies CS if \(P(\mathbb{P}(P_t \leq u | I_t) \leq u\) for all \(u \in [0, 1]\), that is, \(P_t\) is a valid p-value even conditional on the information used to define its rejection threshold, \(I_t\). The Supporting Information discusses a slight variation of this definition. The CS assumption is nontrivial to verify: If joint distribution of underlying test statistics is completely unknown, there is no clear way to produce p-values that are CS.

Zrnic et al. (2021) propose a clever means of circumventing this problem when partial knowledge of this joint distribution is available. In particular, the authors consider cases where each p-value is dependent only with a subset of its neighbors, and refer to this subset as a “conflict set.” Reflecting these local relationships, we say that a p-value \(P_t\) follows local dependence if it is independent of the information used to define its thresholds (i.e., if \(P_t \perp I_t\)).

For example, suppose that test statistics are observed in batches, and are known to be independent across batches. Let \(b_t\) be the batch label for the \(t\)th hypothesis. Even if the within-batch dependencies are unknown, we can still proceed by choosing parameters for upcoming tests only based on the test statistics from previous batches. By constraining \(I_t\) to be a function of \(\{P_j : b_i < b_t\}\), we can effectively ignore the within-batch dependencies, as \(P(P_t \leq u | I_t) = P(P_t \leq u)\).

While local dependence implies CS (for marginally super-uniform p-values), it is not immediately obvious whether the reverse is true. A more precise understanding of how CS restricts dependencies does not appear to have been illustrated in the literature. To explore this link further, we introduce the following, novel remark.

Remark 1 (CS Necessary Conditions). For a given \(i \in \mathcal{H}_0\), if \(P_i\) is (strictly) uniformly distributed (i.e., \(P(P_i \leq u) = u\), then CS holds for \(P_i\) if and only if \(P_i \perp I_t\).

The implication of Remark 1 is that, in order to achieve CS, we must either plan testing thresholds for \(P_i\) only using variables that are uninformative of \(P_i\) (as in Zrnic et al., 2021), or we must use p-values that are not marginally uniform. Resorting to nonuniform p-values may cause our tests to be underpowered, although there are corrective methods that can mitigate this concern (see, e.g., Tian & Ramdas, 2019).

2.3 Existing FDR bounds for LORD++ and SAFFRON, under independence

Zrnic et al. (2021) show that LORD++ and SAFFRON each controls mFDR under the CS condition (see their Theorem 2; as well as Ramdas et al., 2017; and Ramdas et al., 2018).

Theorem 1 (mFDR under CS). Assume that the null p-values satisfy CS (i.e., \(P(P_t \leq u | I_t) \leq u\) for all \(t \in \mathcal{H}_0\) and \(u \in [0, 1]\)), and that \(\alpha_t\) and (when applicable) \(\lambda_i\) are deterministic functions of \(I_t\). Under these conditions, the following two results hold.


1. (LORD++ mFDR Control) If the parameters \( \{\alpha_i\}_{i \in \mathbb{N}} \) are selected so that \( \hat{FDP}_0(t) \leq \alpha \) for all \( t \in \mathbb{N} \) (e.g., LORD++), then \( mFDR(t) \leq \alpha \).

2. (SAFFRON FDR Control) If the parameters \( \{\alpha_i, \lambda_i\}_{i \in \mathbb{N}} \) are selected so that \( \hat{FDP}_\lambda(t) \leq \alpha \) for all \( t \in \mathbb{N} \) (e.g., SAFFRON), then \( mFDR(t) \leq \alpha \).

Additionally, Ramdas et al. (2017) and Ramdas et al. (2018) show that traditional FDR control is achieved under a combination of CS, a monotonicity condition, and a p-value independence condition.

**Theorem 2** (FDR under independence). In addition to the conditions of Theorem 1, we make the following assumptions.

1. (Independence) The null p-values are independent of each other and the nonnulls.
2. (Monotonicity) For each \( t \in \mathbb{N} \), the parameters \( \alpha_t \) and (when applicable) \( \lambda_t \) are deterministic, coordinatewise non-decreasing functions of the set of indicators \( I_t \subseteq (1(P_1 \leq \alpha_1), \ldots, 1(P_{t-1} \leq \alpha_{t-1}), 1(P_1 \leq \lambda_1), \ldots, 1(P_{t-1} \leq \lambda_{t-1})) \). Note that, for LORD++, this condition can be simplified by fixing each \( \lambda_t = 0 \).

Under the above conditions, the following two results hold.

1. (LORD++ FDR Control) If the parameters \( \{\alpha_i\}_{i \in \mathbb{N}} \) are selected so that \( \hat{FDP}_0(t) \leq \alpha \) for all \( t \in \mathbb{N} \) (e.g., LORD++), then \( FDR(t) \leq \alpha \).
2. (SAFFRON FDR Control) If the parameters \( \{\alpha_i, \lambda_i\}_{i \in \mathbb{N}} \) are selected so that \( \hat{FDP}_\lambda(t) \leq \alpha \) for all \( t \in \mathbb{N} \) (e.g., SAFFRON), then \( FDR(t) \leq \alpha \).

The monotonicity condition in Theorem 2 essentially states that lower p-values never lead us to require stricter thresholds in future tests—the more hypotheses we reject, the easier it will be to reject future hypotheses. This monotonicity condition can be ensured by design.

In the next section, we show that the additional assumptions made in Theorem 2 can be greatly relaxed. Most notably, the independence assumption can be weakened to a positive dependence assumption. That said, we still require CS.

### 3. FDR CONTROL UNDER POSITIVE DEPENDENCE

In Section 3.1, below, we introduce a monotonicity condition and a nonnegative dependence condition. We then present our main result, along with a sketch of the proof. In Section 3.2, we discuss how certain forms of adaptive stopping rules can be formalized under our assumptions.

#### 3.1. Main result

Our first required condition for online FDR control is analogous to the monotonicity conditions described in Theorem 2, and can similarly be ensured by design. We will require that any decrease to a p-value (i.e., making it “more significant”) cannot decrease the total number of rejections.

**Condition 1** (Relaxed Monotonicity). For any \( t \in \mathbb{N} \) and for any two p-value vectors \( \mathbf{p} = (p_1, \ldots, p_t) \) and \( \mathbf{p}' = (p'_1, \ldots, p'_t) \) that could possibly result from the first \( t \) tests, if \( p_i \leq p'_i \) for all \( i \leq t \), then \( \mathbf{p} \) must produce at least as many rejections as \( \mathbf{p}' \).

The most straightforward way to ensure Condition 1 is to require that each threshold \( \alpha_i \) be a monotonic, nonincreasing function of the preceding p-values \( (P_1, \ldots, P_{t-1}) \). However, we will see in the next section that weaker versions of monotonicity can also satisfy Condition 1.

While many testing procedures satisfy Condition 1, one notable exception is the adaptive discarding (ADDIS) method proposed by Tian and Ramdas (2019). This method is designed for the case where the null hypotheses are conservative, that is, stochastically larger than a uniform distribution. Roughly speaking, ADDIS screens and “discards” large p-values before testing. It can be most powerful when the null p-values are especially close to one, making them easier to distinguish from alternative p-values. For this reason, ADDIS does not satisfy Condition 1.
In order to relax the independence requirement in Theorem 2, we next introduce a version of the well-known PRDS condition developed by Benjamini and Yekutieli (2001). This condition will depend on the notion of increasing sets. A set of $k$-dimensional vectors $D \in [0,1]^k$ is called increasing if, for any vector $\mathbf{x} = (x_1, \ldots, x_t) \in D$ and any vector $\mathbf{y} = (y_1, \ldots, y_t)$ satisfying $x_i \leq y_i$ for all $i$, it must also be that $\mathbf{y} \in D$. For example, for any $r \leq t$, Condition 1 implies that the set of $p$-values that produce no more than $r$ rejections by stage $t$ is an increasing set.

**Condition 2** (Conditional PRDS). For any stage $t$, any null index $i \leq t$ satisfying $i \in \mathcal{H}_0$, and increasing set $D \subset [0,1]^t$, the probability $\mathbb{P}((P_1, \ldots, P_t) \in D | P_i = u, I_t)$ is nondecreasing in $u$.

Roughly speaking Condition 2 says that each null $p$-value is positively associated with the other $p$-values. Aswith local dependence, we may expect Condition 2 to hold when sequentially analyzing population subgroups, with several hypotheses tested per group. We may also expect Condition 2 to hold when studying temporal test statistics believed to follow an autoregressive structure.

The key take-away from Conditions 1 and 2 is that, together, they imply that low $p$-values in early stages of the sequence tend to be associated with a higher number of rejections by later stages of the sequence. This idea will play a central role in our derivation of FDR control.

**Theorem 3** (FDR under nonnegative, local dependence). Under Conditions 1 and 2, if the null $p$-values are conditionally super-uniform (i.e., $\mathbb{P}(P_t \leq u | I_t) \leq u$) for all $t \in \mathbb{N}$ and $u \in [0,1]$, then the following two results hold.

1. (LORD++ FDR Control) If the parameters $\{\alpha_i\}_{i \in \mathbb{N}}$ are selected so that $\hat{\text{FDP}}_0(t) \leq \alpha$ for all $t \in \mathbb{N}$ (e.g., LORD++), then $\text{FDR}(t) \leq \alpha$.

2. (SAFFRON FDR Control) If the parameters $\{\alpha_i, \lambda_i\}_{i \in \mathbb{N}}$ are selected so that $\hat{\text{FDP}}_{\lambda}(t) \leq \alpha$ for all $t \in \mathbb{N}$ (e.g., SAFFRON), then $\text{FDR}(t) \leq \alpha$.

Importantly, Theorem 3 still requires the CS condition. Thus, for well-powered $p$-value sequences, Theorem 3 effectively requires the $p$-values satisfy nonnegative, local dependence (see Section 2.2).

A sketch of the intuition for Theorem 3 is as follows. Consider the especially simple case where all parameters $\{\alpha_i, \lambda_i\}_{i \in \mathbb{N}}$ are fixed a priori (i.e., $I_i = \emptyset$ for all $i$). This setting greatly simplifies the notation needed, and still sheds light on how Theorem 3 can be proved when $\alpha_i$ and $\lambda_i$ are determined adaptively (see comments below). Under Conditions 1 and 2, smaller null $p$-values are generally associated with larger values for $(1 \lor |R_i|)$. Thus, for any $i \in \mathcal{H}_0$ and $t \geq i$, we would expect the rejection indicator $1(P_i \leq \alpha_i)$ to be negatively correlated with $1/(1 \lor |R_i|)$, that is,

\[
E\left[\frac{1}{1 \lor |R_i|} 1(P_i \leq \alpha_i)\right] \leq E\left[\frac{1}{1 \lor |R_i|}\right] E[1(P_i \leq \alpha_i)].
\]

Applying this, we have

\[
\text{FDR}(t) = \sum_{i \leq t; i \in \mathcal{H}_0} E\left[\frac{1(P_i \leq \alpha_i)}{1 \lor |R_i|}\right] \\
\leq \sum_{i \leq t; i \in \mathcal{H}_0} E\left[\frac{1}{1 \lor |R_i|}\right] E[1(P_i \leq \alpha_i)] \\
\leq \sum_{i \leq t; i \in \mathcal{H}_0} E\left[\frac{1}{1 \lor |R_i|}\right] \alpha_i \\
= E\left[\sum_{i \leq t; i \in \mathcal{H}_0} \frac{\alpha_i}{1 \lor |R_i|}\right],
\]

where the first inequality comes from Equation (3), and the second inequality comes from $\mathbb{P}(P_i \leq \alpha_i) = \mathbb{P}(P_i \leq \alpha_i | I_i) \leq \alpha_i$. Thus, if $\hat{\text{FDP}}_0(t) = \sum_{i \leq t} \frac{\alpha_i}{1 \lor |R_i|} \leq \alpha$, then monotonicity of expectations implies that $\text{FDR}(t) \leq \alpha$. 


To sketch the result for $\hat{\text{FDP}}_\lambda$, we build on Equation \((4)\) by multiplying each summation term by $E[1(P_i > \lambda_i)]/(1 - \lambda)$ $\geq 1$. We obtain

$$ \text{FDR}(t) \leq \sum_{i \leq t; i \in H_0} E \left[ \frac{\alpha_i}{1 \lor |R_i|} \right] \frac{E[1(P_i > \lambda_i)]}{1 - \lambda_i} \leq \sum_{i \leq t; i \in H_0} E \left[ \frac{\alpha_i[1(P_i > \lambda_i)]}{1 - \lambda_i} \right] \frac{1}{1 \lor |R_i|}, $$

where the second inequality comes from the fact that indicators of large $p$-values, $1(P_i > \lambda_i)$, are positively correlated with $(1 \lor |R_i|)^{-1}$. From here, if $\hat{\text{FDP}}_\lambda(t) \leq \alpha$, then monotonicity of expectations again implies that $\text{FDR}(t) \leq \alpha$. In the full details of the proof, we also iterate expectations over $I_i$ in order to account for adaptively defined parameters $\alpha_i$ and $\lambda_i$ (see the Supporting Information).

### 3.2 Incorporating certain forms of adaptive stopping rules

While Theorem 3 ensures FDR control at fixed times $t \in \mathbb{N}$, it does not uniformly control the FDR across all times $t$. That is, while Theorem 3 gives conditions under which $\sup_{t \in \mathbb{N}} E[\text{FDP}(t)] \leq \alpha$, it does not give conditions under which $E[\sup_{t \in \mathbb{N}} \text{FDP}(t)] \leq \alpha$. The practical relevance of this point is that some analysts may choose their final test stage adaptively, in the hope of rejecting a large proportion of the hypotheses tested. In other words, they may wish to stop testing early in the face of especially strong preliminary results. We show in this section that, for certain types of adaptive stopping rules, the conditions required for Theorem 3 still hold.

We will use $T$ to denote an adaptively determined stopping time, and will say that FDR is controlled under adaptive stopping times if

$$ E[\text{FDP}(T)] = E \left[ \frac{|H_0 \cap R_T|}{1 \lor |R_T|} \right] \leq \alpha. $$

We will generally assume that $1(T > t)$ is a deterministic function of $(P_1, ..., P_{t-1})$, and that $T$ is upper bounded (with probability 1) by a prespecified constant $t_{\max}$.

At first glance, it appears straightforward to account for these kinds of adaptive stopping times by simply setting $\alpha_t$ equal to zero for every stage $t > T$, and continuing testing until stage $t_{\max}$. Since $R_T = R_{t_{\max}}$, we can effectively control $E[\text{FDP}(T)]$ by controlling $\text{FDR}(t_{\max})$. Unfortunately, such a method of assigning thresholds is not monotonic in the observed $p$-values. Seeing a sufficient number of rejections may cause us to lower our thresholds for all future tests by setting them to zero, and so the monotonicity conditions in Theorem 2 are not satisfied.

However, even though these kinds of stopping rules are not compatible with monotonicity constraints in Theorem 2, there are several types of adaptive stopping rules that still maintain Condition 1. Such stopping rules will allow us to control $E[\text{FDP}(T)]$ by using Theorem 3 to control $\text{FDR}(t_{\max})$.

In particular, Condition 1 may still be satisfied if users combine a monotone threshold function with a monotone stopping rule. To formalize these types of rules, let $\{\beta_i\}_{i=1}^\infty$ be a sequence of functions used for defining thresholds, where $\beta_i : [0, 1]^{i-1} \rightarrow [0, 1]$ is a mapping from first $i-1$ $p$-values to the threshold $\alpha_i = \beta_i(P_1, ..., P_{i-1})$.

As a first example, consider the case where users assign each threshold $\alpha_i$ according to a coordinate-wise nonincreasing function $\rho_i^{\text{noninc}}$, but can also decide to stop testing if a certain critical threshold $\gamma_{\text{max-stage}}$ of rejections have been made, or if a certain number of tests $\gamma_{\text{max-stage}}$ have completed. This implies the following form for each function $\beta_i$:

$$ \beta_i(P_1, ..., P_{i-1}) = \rho_i^{\text{noninc}}(P_1, ..., P_{i-1}) \times 1(|R_{i-1}| < \gamma_{\text{max-R}}) \times 1(i \leq \gamma_{\text{max-stage}}). $$

Here, the overall function $\beta_i$ is not monotonic—smaller $p$-values will increase future thresholds at first, but eventually will cause them to drop to zero. Still, even though the thresholds themselves are not monotonic in the $p$-value sequence,
the total number of rejections is monotonic in the p-value sequence. Any decrease to a p-value can lead to a cease of testing, but cannot decrease the total number of rejections, and so Condition 1 still holds.

A similar result holds if we allow the maximum number of rejections, or the maximum number of stages, to be extended in the face of a strong signal in the preliminary tests. To formalize this, we can instead assume that each function $\beta_i$ has the following structure:

$$\beta_i(P_1, \ldots, P_{i-1}) = \beta^\text{noninc}_i(P_1, \ldots, P_{i-1})$$

$$\times 1(|R_{i-1}| < \beta^\text{max-R}_i(P_1, \ldots, P_{i-1}))$$

$$\times 1(i \leq \beta^\text{max-stage}_i(P_1, \ldots, P_{i-1})),$$

where $\beta^\text{max-R}_i : [0, 1]^{i-1} \rightarrow \mathbb{N}$ and $\beta^\text{max-stage}_i : [0, 1]^{i-1} \rightarrow \mathbb{N}$ are nonincreasing functions of $(P_1, \ldots, P_{i-1})$. Here, roughly speaking, larger p-values in the early stages must produce either stricter thresholds (via $\beta^\text{noninc}_i$), a lower number of maximum stages (via $\beta^\text{max-stage}_i$), or a lower number of maximum rejections (via $\beta^\text{max-R}_i$). Any of these changes can only reduce the number of total discoveries. We again see that Condition 1 holds: The threshold functions $\beta_i$ are not themselves monotonic, but the number of discoveries is still nonincreasing in the observed p-values.
FIGURE 2  Simulated power, under the same settings as Figure 1. Results for adaptive stopping times are omitted, as stopping early necessarily decreases power. Shaded ribbons again show ±2 Monte Carlo standard errors, although these are more difficult to visually distinguish than in Figure 1 because all standard errors are <0.01.

4  |  SIMULATIONS

Here, we illustrate our FDR control result using simulation. Based on the setup used by Ramdas et al. (2018), we simulate a vector of $t_{\text{max}} = 500$ normally distributed variables $(Z_1, \ldots, Z_{t_{\text{max}}}) \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_{t_{\text{max}}})$ is a vector of mean parameters and $\Sigma$ is a covariance matrix defined in detail below. For each statistic, our null hypothesis $H_i$ is that $\mathbb{E}(Z_i) = 0$, and our $p$-value $P_i = \Phi(-Z_i)$ is the (unadjusted) result of a one-sided test of $H_i$. In each simulated sample, we select a random subset of the parameters $(\mu_1, \ldots, \mu_{t_{\text{max}}})$ to be zero, and assign the remaining mean parameters to be 3. Let $\pi_1$ denote the proportion of mean parameters that are equal to 3, that is, the proportion of false nulls.

We define $\Sigma$ according to a block-covariance structure with block size denoted by $n_{\text{batch}}$, and within-block covariance $\rho$. As in Section 2.2, we define $b_i$ to be the block label for the $i$th test statistic. We define the each element of $\Sigma$ as follows:

$$
\Sigma_{ij} = \begin{cases} 
1 & \text{if } i = j \\
\rho & \text{if } i \neq j \text{ and } b_i = b_j \\
0 & \text{otherwise}.
\end{cases}
$$

We simulate all combinations of $n_{\text{batch}} \in \{1, 5, 10, 50\}$; $\rho \in \{0.3, 0.6\}$; and

$$
\pi_1 \in \{0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5\}.
$$

For each combination, we run 1000 iterations.
In each simulated sample, we apply the “batch” versions of LORD++ and SAFFRON in online FDR package, using the default settings and a desired FDR of $\alpha = 0.05$. This lets us require each threshold $\alpha_i$ to be chosen based only on the test statistics from previous batches (Robertson et al., 2019; see also Zrnic et al., 2021). That is, we set $I_i = \{P_i' : b_{i'} < b_i\}$, so that $P_i \perp I_i$. Here, we can see that CS holds from the fact that $P(P_i \leq u | I_i) = P(P_i \leq u) = u$ whenever $i \in H_0$.

For each of these settings, we also simulate adaptive stopping procedures in which analysts stop testing after a “rejection limit” of 10 or 50 discoveries have been made. That is, we set $\gamma_{\text{max-R}}$ in Section 3.2 to be 10 or 50, set $\gamma_{\text{max-stage}}$ to be $t_{\text{max}}$, and set $\beta_{\text{noninc}}$ to be the rejection thresholds from either LORD++ or SAFFRON. We implement these approaches again with the onlineFDR package, but with an added filtering step to accept all null hypotheses after $\gamma_{\text{max-R}}$ rejections.

4.1 Results

Figures 1 and 2 show the results of our analysis. We see that, in every scenario tested, LORD++ and SAFFRON control FDR($t_{\text{max}}$) at the appropriate rate. FDR is uniformly lower when adaptive stopping rules are used. We omit power results for adaptive stopping times, since early stopping necessarily decreases power.

5 CONCLUDING DISCUSSION

Where SAFFRON, LORD++, and alpha investing were previously only shown to control FDR under independence, we show that they additionally control FDR under positive, local dependence.

Although our work focuses on controlling FDR, this should not be taken as an implicit, blanket endorsement of FDR over mFDR. On the one hand, Javanmard and Montanari (2018) argue that FDR carries a more easily understood interpretation than mFDR. On the other, mFDR may be more easily applicable within a decision theory framework (Bickel, 2004), and may more naturally allow for decentralized control of the proportion of false positives across an entire scientific literature (Fernando et al., 2004; see also van den Oord, 2008; Zrnic et al., 2021). Further research into online control for both metrics, as well as control for FWER, remains vital (see also Tian & Ramdas, 2021; Weinstein & Ramdas, 2020).

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

This article uses only simulated data, which can be recreated from the Supporting Information.

ORCID

Aaron Fisher https://orcid.org/0000-0003-4089-585X

REFERENCES

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

Benjamini, Y., & Hochberg, Y. (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.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4), 1165–1188.

Bickel, D. R. (2004). Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? Statistical Applications in Genetics and Molecular Biology, 3(1), Article 8.

Chen, S., & Kasiviswanathan, S. (2020). Contextual online false discovery rate control. In S. Chiappa and R. Calandra (Eds.), Proceedings of the twenty third international conference on artificial intelligence and statistics (Vol. 108, pp. 952–961). PMLR.

Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., & Soller, M. (2004). Controlling the proportion of false positives in multiple dependent tests. Genetics, 166(1), 611–619.

Fithian, W., & Lei, L. (2022). Conditional calibration for false discovery rate control under dependence. Annals of Statistics, 50(6), 3091–3118.

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

Javanmard, A., & Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. Annals of Statistics, 46(2), 526–554.
SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Fisher, A. (2024). Online false discovery rate control for LORD++ and SAFFRON under positive, local dependence. *Biometrical Journal, 66*, 2300177. [https://doi.org/10.1002/bimj.202300177](https://doi.org/10.1002/bimj.202300177)