SEQUENTIAL SELECTION PROCEDURES AND FALSE DISCOVERY RATE CONTROL

By Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova and Robert Tibshirani

We consider a multiple hypothesis testing setting where the hypotheses are ordered and one is only permitted to reject an initial contiguous block, $H_1, \ldots, H_k$, of hypotheses. A rejection rule in this setting amounts to a procedure for choosing the stopping point k. This setting is inspired by the sequential nature of many model selection problems, where choosing a stopping point or a model is equivalent to rejecting all hypotheses up to that point and none thereafter. We propose two new testing procedures, and prove that they control the false discovery rate in the ordered testing setting. We also show how the methods can be applied to model selection using recent results on p-values in sequential model selection settings.

1. Introduction. Suppose that we have a sequence of null hypotheses, H_1, $H_2, \ldots H_m$, and that we want to to reject some hypotheses while controlling the False Discovery Rate (FDR, Benjamini and Hochberg, 1995). Moreover, suppose that these hypotheses must be rejected in an ordered fashion: a test procedure must reject hypotheses $H_1, \ldots, H_k$ for some $k \in \{0, 1, \ldots, m\}$. Classical methods for FDR control, such as the original Benjamini-Hochberg selection procedure, are ruled out by the requirement that the hypotheses be rejected in order.

In this paper we introduce new testing procedures that address this problem, and control the False Discovery Rate (FDR) in the ordered setting. Suppose that we have a sequence of p-values, p_1, ..., $p_m \in [0, 1]$ corresponding to the hypotheses H_j, such that p_j is uniformly distributed on $[0, 1]$ when H_j is true. Our proposed methods start by transforming the sequence of p-values p_1, ..., p_m into a monotone increasing sequence of statistics $0 \leq q_1 \leq \ldots \leq q_m \leq 1$. We then prove that we achieve ordered FDR control by applying the original Benjamini-Hochberg procedure on the monotone test statistics q_i.

Our setup is motivated by the problem of selecting stopping rules for model selection procedures such as the lasso, stepwise regression, and hierarchical clustering. Recent work (Lockhart et al., 2014; G'Sell, Taylor and Tibshirani, 2013; Taylor et al., 2013, 2014) develops p-values for the individual steps of these procedures. Under this formalism, taking an additional step along the model selection path is equivalent to rejecting the hypothesis that the existing model captures all the signal. A problem left open by this literature, however, is how these p-values can be used to select a model with inferential guarantees. Our results respond to this challenge: by applying our generic sequential FDR control rule to the step-specific p-values, we obtain a stopping rule with a guarantee on the fraction of useless steps taken by the model selection procedure.

Keywords and phrases: multiple hypothesis testing, stopping rule, false discovery rate, sequential testing

TABLE 1
Typical realization of p-values for LARS, as proposed by Taylor et al. (2014).

LARS step	1	2	3	4	5	6	7	8	9	10
Predictor	3	1	4	10	9	8	5	2	6	7
p-value	0.00	0.08	0.34	0.15	0.93	0.12	0.64	0.25	0.49	·

1.1. *A first example.*

To motivate our work further, consider a simple model selection problem. We have n observations from a linear model with p predictors:

$$y_i = \beta_0 + \sum_j x_{ij}\beta_j + Z_i \text{ with } Z_i \sim \mathcal{N}(0,1), \tag{1}$$

and apply the lasso (ℓ_1-penalized regression) to estimate the parameters. We apply the least angle regression procedure of Efron et al. (2004), which enters variables in a forward sequential manner. As explained in Section 4, the recent work of Taylor et al. (2014) provides a method for computing p-values for this sequence, each one testing whether the variables not yet in the model have a true coefficient of zero. Table 1 has a typical realization of these p-values; we generated data with

$$n = 50,\ p = 10,\ x_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\ \beta_1 = 2,\ \beta_3 = 4,\ \beta_2 = \beta_4 = \beta_5 \ldots \beta_{10} = 0.$$

These p-values are not exchangeable, and must be treated in the order in which the predictors were entered: 3, 1, 4 etc.

Figure 1 shows the result of applying one of our new procedures, *ForwardStop*, to 1000 realizations from model (1). This procedure, described in the next section, delivers for each target FDR α a stopping index k along with a guarantee that the model consisting of the first $k(\alpha)$ predictors entered has FDR at most α. The left panel shows the number of predictors selected by *ForwardStop* over the 1000 realizations, while the right panel shows that the achieved FDR is indeed less that the target level.

1.2. *Stopping Rules for Ordered FDR Control.*

In the ordered testing setting, a valid rejection rule is a function of p_1 , ..., p_m that returns a cutoff $\hat{k}$ such that hypotheses $H_1, \ldots, H_{\hat{k}}$ are rejected. The False Discovery Rate (FDR) is defined as $\mathbb{E}\left[V(\hat{k})/\max(1,\hat{k})\right]$, where $V(\hat{k})$ is the number of null hypotheses among the rejected hypotheses H_1, ... , $H_{\hat{k}}$.

We propose two rejection functions for this scenario, called *ForwardStop*:

$$\hat{k}_F = \max\left\{k \in \{1, \ldots, m\} : -\frac{1}{k}\sum_{i=1}^{k}\log(1-p_i) \le \alpha\right\}, \tag{2}$$

and *StrongStop*:

$$\hat{k}_S = \max\left\{k \in \{1, \ldots, m\} : \exp\left(\sum_{j=k}^{m}\frac{\log p_j}{j}\right) < \frac{\alpha k}{m}\right\}. \tag{3}$$

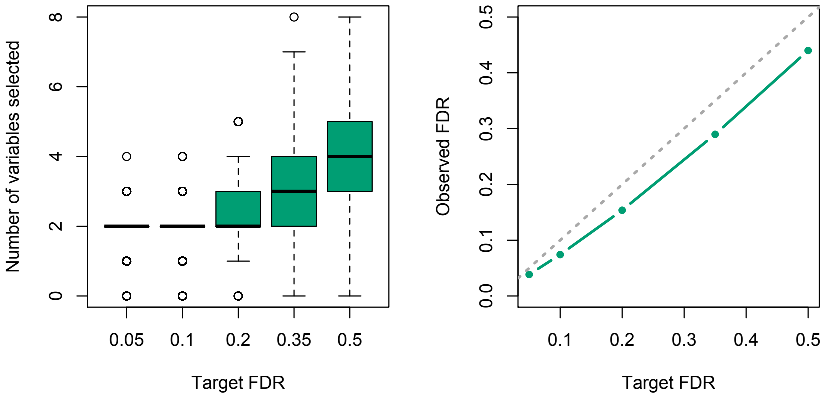

FIG 1. *Small simulated linear model example: lasso p-values for forward adaptive regression. The left panel shows the number of predictors selected by* ForwardStop *over* 1000 *realizations, while the right panel shows that the achieved FDR versus the target level, with the* 45^o *line drawn for reference.*

We adopt the convention that $\max(\emptyset) = 0$, so that $\hat{k} = 0$ whenever no rejections can be made. In Section 2 we show that both *ForwardStop* and *StrongStop* control FDR at level α.

ForwardStop first transforms the p-values, and then sets the rejection threshold at the largest k for which the first k transformed p-values have a small enough average. If the first p-values are very small, then *ForwardStop* will always reject the first hypotheses regardless of the last p-values. The rule is thus moderately robust to model misspecification, as it will not lose power even if the last p-values are a little bit too large.

Our second rule, *StrongStop* (3), comes with a stronger guarantee than *ForwardStop*. As we show in Section 2, provided that the non-null p-values precede the null ones, it not only controls the FDR, but also controls the Family-Wise Error Rate (FWER) at level α. Recall that the FWER is the probability that a decision rule makes even a single false discovery. If false discoveries have a particularly high cost, then *StrongStop* may be more attractive than *ForwardStop*. The main weakness of *StrongStop* is that the decision to reject at k depends on all the p-values after k. If the very last p-values are slightly larger than they should be under the uniform hypothesis, then the rule suffers a considerable loss of power.

1.3. *Related Work.* Although there is an extensive literature on FDR control (e.g., Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Efron et al., 2001; Storey, Taylor and Siegmund, 2004), no definitive procedure for ordered FDR control has been proposed so far. The best method we are aware of is an adaptation of the α-investing approach of Foster and Stine (2008). However, this procedure is not known to formally control the FDR (Foster and Stine prove that it controls the mFDR, defined as $\mathbb{E}V/(\mathbb{E}R+\eta)$ for some constant η); moreover, in our simulations, this approach has lower power than our proposed methods.

The problem of providing FDR control for the lasso has been studied, among others, by Bogdan et al. (2013), Benjamini and Gavrilov (2009), Lin, Foster and Ungar (2011), Meinshausen and Bühlmann (2010), Shah and Samworth (2012), and Wu, Boos and Stefanski (2007), using a wide variety of ideas involving resampling, pseudo-variables, and specifically tailored selection penalties. The goal of our paper is not to directly compete with these methods, but rather to provide "theoretical glue" that lets us transform the rapidly growing family of sequential p-values (Lockhart et al., 2014; G'Sell, Taylor and Tibshirani, 2013; Taylor et al., 2013, 2014) into model selection procedures with FDR guarantees.

1.4. *Outline of this paper.* We begin by presenting generic methods for FDR control in ordered settings. Section 2 develops our two main proposals for sequential testing, *ForwardStop* and *StrongStop*, along with their theoretical justification. We evaluate these rules on simulations in Section 3. In Section 4, review the recent literature constructing sequential p-values for model selection problems and discuss their relation to our procedures. In Section 5, we develop a more specialized version of *StrongStop*, called *TailStop*, which takes advantage of special properties of some of the proposed sequential tests. We provide demonstrations of all of our procedures on example model selection p-values in 6, and conclude with a discussion of some practical considerations in Section 7.

2. False Discovery Rate Control for Ordered Hypotheses.

In this section, we study a generic ordered layout where we a sequence of hypotheses that are associated with p-values $p_1, \ldots, p_m \in [0,1]$. A subset $M \subset \{0, \ldots, m\}$ of these p-values are null, with the property that

$$\{p_i : i \in M\} \overset{\text{iid}}{\sim} U([0,1]). \tag{4}$$

We can reject the k first hypotheses for some k of our choice. Our goal is to make k as large as possible, while controlling the number of false discoveries

$$V(k) = \left|\{i \in M : i \leq k\}\right|.$$

Specifically, we want to use a rule $\hat{k}$ with a bounded false discovery rate

$$FDR(\hat{k}) = \mathbb{E}\left[\frac{V(\hat{k})}{\max\left\{\hat{k},\, 1\right\}}\right].$$

We develop two procedures that provide such a guarantee.

Classical FDR literature focuses on rejecting a subset of hypotheses $R \in \{0, \ldots, m\}$ such that R contains few false discoveries. Benjamini and Hochberg (1995) showed that, in the context of (4), we can control the FDR as follows. Let $p_{(1)}, \ldots, p_{(m)}$ be the sorted list of p-values, and let

$$\hat{l}_\alpha = \max\left\{l : p_{(l)} < \frac{\alpha\, l}{m}\right\}.$$

Then, if we reject those hypotheses corresponding to $\hat{l}_\alpha$ smallest p-values, we control the FDR at level α. This method for selecting the rejection set R is known as the BH procedure. The key difference between the setup of Benjamini and Hochberg (1995) and our problem is that, in the former, the rejection set R can be arbitrary, whereas here we must always reject the first k hypotheses for some k. For example, even if the p-value corresponding to the third hypothesis is very small, we cannot reject the third hypothesis unless we also reject the first and second hypotheses.

2.1. *A BH-Type Procedure for Ordered Selection.* The main motivation behind our first procedure—*ForwardStop*—is the following thought experiment. Suppose that we could transform our p-values $p_1, \ldots, p_m$ into statistics $q_1 < \ldots < q_m$, such that the q_i behaved like a sorted list of p-values. Then, we could apply the BH procedure on the q_i, and get a rejection set R of the form $R = \{1, \ldots, k\}$.

Under the global null where $p_1, \ldots, p_m \overset{\text{iid}}{\sim} U([0,1])$, we can achieve such a transformation using the Rényi representation theorem (Rényi, 1953). Rényi showed that if $Y_1, \ldots, Y_m$ are independent standard exponential random variables, then

$$\left(\frac{Y_1}{m}, \frac{Y_1}{m} + \frac{Y_2}{m-1}, \ldots, \sum_{i=1}^{m} \frac{Y_i}{m-i+1}\right) \overset{d}{=} E_{1,m}, E_{2,m}, \ldots, E_{m,m},$$

where the $E_{i,m}$ are exponential order statistics, meaning that the $E_{i,m}$ have the same distribution as a sorted list of independent standard exponential random variables. Rényi representation provides us with a tool that lets us map a list of independent exponential random variables to a list of sorted order statistics, and vice-versa.

In our context, let

$$Y_i = -\log(1 - p_i), \tag{5}$$

$$Z_i = \sum_{j=1}^{i} Y_j / (m - j + 1), \text{ and} \tag{6}$$

$$q_i = 1 - e^{-Z_i}. \tag{7}$$

Under the global null, the Y_i are distributed as independent exponential random variables. Thus, by Rényi representation, the Z_i are distributed as exponential order statistics, and so the q_i are distributed like uniform order statistics.

This argument suggests that in an ordered selection setup, we should reject the first $\hat{k}_F^q$ hypotheses where

$$\hat{k}_F^q = \max\left\{k : q_k \leq \frac{\alpha\, k}{m}\right\}. \tag{8}$$

The Rényi representation combined with the BH procedure immediately implies that the rule $\hat{k}_F$ controls the FDR at level α under the global null. Once we leave the global null, Rényi representation no longer applies; however, as we show in the following results, our procedure still controls the FDR.

We begin by stating a result under a slightly restricted setup, where we assume that the s first p-values are non-null and the $m-s$ last p-values are null. We will later relax this constraint. The proof of the following result is closely inspired by the martingale argument of Storey, Taylor and Siegmund (2004).

LEMMA 1. *Suppose that we have p-values $p_1, \ldots, p_m \in (0,1)$, the last $m-s$ of which are null (i.e., independently drawn from $U([0,1])$. Define q_i as in (7). Then the rule $\hat{k}_F^q$ controls the FDR at level α, meaning that*

$$\mathbb{E}\left[\frac{\left(\hat{k}_F^q - s\right)_+}{\max\left\{\hat{k}_F^q, 1\right\}}\right] \leq \alpha. \tag{9}$$

Now the test statistics q_i constructed in Lemma 1 depend on m. We can simplify the rule by augmenting our list of p-values with additional null test statistics (taking $m \to \infty$), and using the fact that $\frac{1-e^{-x}}{x} \to 1$ as x gets small. This gives rise to one of our main proposals:

PROCEDURE 1 (ForwardStop). *Let $p_1, \ldots, p_m \in [0,1]$, and let $0 < \alpha < 1$. We reject hypotheses $1, \ldots, \hat{k}_F$, where*

$$\hat{k}_F = \max\left\{k \in \{1, \ldots, m\} : \frac{1}{k}\sum_{i=1}^{k} Y_i \leq \alpha\right\}, \tag{10}$$

$$\textit{and } Y_i = -\log(1-p_i).$$

We call this procedure *ForwardStop* because it scans the p-values in a forward manner: If $\frac{1}{k}\sum_{i=1}^{k} Y_i \leq \alpha$, then we know that we can reject the first k hypotheses regardless of the remaining p-values. This property is desirable if we trust the first p-values more than the last p-values. We prove the following result:

COROLLARY 1. *Under the conditions of Lemma 1, the* ForwardStop *procedure defined in (10) has FDR is controlled at level α.*

For simplicity, we have assumed so far that the non-null p-values are all concentrated at the beginning of the list. This hypothesis, however, is not required for FDR control.

THEOREM 1. *In the setup described in (4), the ForwardStop stopping rule $\hat{k}_F$ controls FDR at level α, meaning that*

$$\mathbb{E}\left[\frac{V(\hat{k}_F)}{\max\left\{\hat{k}_F, 1\right\}}\right] \leq \alpha.$$

We note that this last property does not hold for the more complicated rule from Lemma 1; this is a major advantage of the *ForwardStop* rule.

2.2. *Strong Control for Ordered Selection.* In the previous section, we created the ordered test statistics Z_i in (6) by summing transformed p-values starting from the first p-value. This choice was in some sense arbitrary. Under the global null, we could just as well obtain uniform order statistics q_i by summing from the back:

$$\widetilde{Y}_i = -\log(p_i), \tag{11}$$

$$\widetilde{Z}_i = \sum_{j=i}^{m} Y_j/j, \text{ and} \tag{12}$$

$$\tilde{q}_i = e^{-\widetilde{Z}_i}. \tag{13}$$

If we run the BH procedure on these backward test statistics, we obtain another method for control.

Procedure 2 (StrongStop). *Let p_1, ..., $p_m \in [0, 1]$, and let $0 < \alpha < 1$. We reject hypotheses* 1, ..., $\hat{k}$, *where*

$$\hat{k}_S = \max\left\{k \in \{1, \ldots, m\} : \tilde{q}_k < \frac{\alpha k}{m}\right\} \tag{14}$$

and $\tilde{q}_k$ is as defined in (13).

Unlike *ForwardStop*, this new procedure needs to look at the p-values corresponding to the last hypotheses before it can choose to make any rejections. This can be a liability if we do not trust the very last p-values much. Looking at the last p-values can however be useful if the model is correctly specified, as it enables us to strengthen our control guarantees: *StrongStop* not only controls the FDR, but also controls the FWER.

Theorem 2. *Suppose that we have p-values p_1, ..., $p_m \in (0, 1)$, the last $m - s$ of which are null (i.e., independently drawn from $U([0, 1])$. Then, the rule $\hat{k}_S$ from* (14) *controls the FWER at level α, meaning that*

$$\mathbb{P}\left[\hat{k}_S > s\right] \leq \alpha. \tag{15}$$

FWER control is stronger than FDR control, and so we immediately conclude from Theorem 2 that *StrongStop* also controls the FDR. Note that the guarantees from Theorem 2 only hold when the non-null p-values all precede the null ones.

3. Simulation Experiments: Simple Ordered Hypothesis Example. In this section, we demonstrate the performance of our methods in three simulation settings of varying difficulty. The simulation settings consist of ordered hypotheses where the separation of the null and non-null hypotheses is varied to determine the difficulty of the scenario.

We consider a sequence of $m = 100$ hypotheses of which $s = 20$ are non-null. The p-values corresponding to the non-null hypotheses are drawn from a $\text{Beta}(1, \beta)$

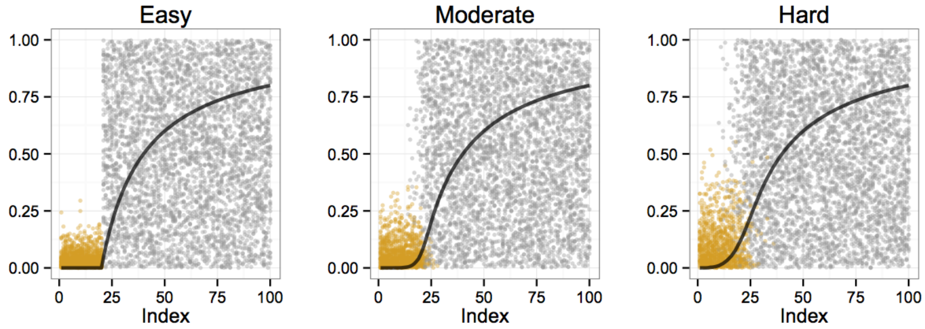

Fig 2. *Observed p-values for 50 realizations of the ordered hypothesis simulations described in Section 3. p-values corresponding to non-null hypotheses are shown in orange, while those corresponding to null hypotheses are shown in gray. The smooth black curve is the average proportion of null hypotheses up to the given index. Non-null p-values are drawn from a* Beta$(1, \beta)$ *distribution, with $\beta = 23, 14, 8$ for the easy, medium and hard settings, respectively.*

distribution, while those corresponding to true null hypotheses are $U([0, 1])$. At each simulation iteration, the indices of the true null hypotheses are selected by sampling without replacement from the set $\{1, 2, \ldots, m = 100\}$ with lower indices having smaller probabilities of being selected. We present results for three simulation cases, which we refer to as 'easy', 'medium', and 'hard.' In the easy setup, we have strong signal $\beta = 23$ and all the non-null hypotheses precede the null hypotheses, so we have perfect separation. In the medium difficulty setup, $\beta = 14$ and the null and non-null hypotheses are lightly inter-mixed. In in the hard difficulty setup, $\beta = 8$ and the two are much more inter-mixed.

For comparison, we also apply the following two rejection rules:

1. *Thresholding at α.* We reject all hypotheses up to the first time that a p-value exceeds α. This is guaranteed to control FWER and FDR at level α.
2. *α-investing.* We use the α-investing scheme of Foster and Stine (2008). While this procedure is not generally guaranteed to yield rejections that obey the ordering restriction, we can select parameters for which it does. In particular, defining an investing rule such that the wealth is equal to zero at the first failure to reject, we get

$$\hat{k}_{invest} = \min\left\{k : p_{k+1} > \frac{(k+1)\alpha}{1 + (k+1)\alpha}\right\}.$$

   This is guaranteed to control $\mathbb{E}V/(\mathbb{E}R + 1)$ at level α.

These are the best competitors we are aware of for our problem. Notice that, unlike *ForwardStop* and *StrongStop*, these rules stop at the first p-value that exceeds a given threshold. Thus, these methods cannot get past a few medium-sized p-values even if there are many very small p-values further along the sequence.

Figure 2 shows scatterplots of observed p-values for 50 realizations of the three setups. To help gauge the difficulty of the decision problem, a black curve is overlaid

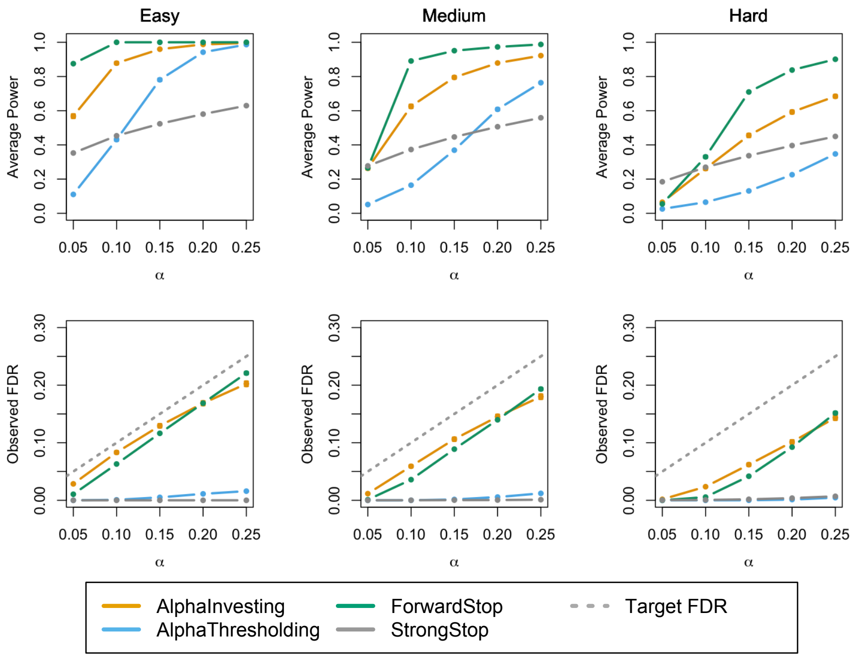

Fig 3. *Average power and observed FDR level for the ordered hypothesis example. All four stopping rules successfully control FDR across the three difficulty settings.* StrongStop *and* α-thresholding *are both very conservative in terms of FDR control. Even though* ForwardStop *and* α-investing *have similar observed FDR curves,* ForwardStop *emerges as the more powerful method.*

to show the average proportion of null hypotheses up to that point. This curve can be thought of as the FDR of a fixed stopping rule which always stops at exactly the given index.

Figure 3 summarizes the performance of *ForwardStop*, *StrongStop*, α-investing and α-thresholding averaged over 2000 simulations. The Figure shows plots of power and observed FDR for target FDR $\alpha \in [0.05, 0.25]$. The notion of power used here is that of average power, defined as the fraction of non-null hypotheses that are rejected (i.e., $(k - V)/s$).

Despite being fairly conservative in terms of FDR, the considered stopping rules perform quite well in terms of average power. *ForwardStop* is the most powerful method in the majority of examples. That being said, *StrongStop* appears to be more powerful than *ForwardStop* in the weak signal/low α settings. This may occur because, unlike the other methods, *StrongStop* scans p-values back-to-front and is therefore less sensitive to the occurrence of large p-values early in the alternative.

In our experiments, α-investing and *ForwardStop* have similar FDR curves, but *ForwardStop* tends to have far great power. Thus, although both rules are comparable in terms of reaching the nominal FDR, *ForwardStop* does better in terms of a precision-recall tradeoff.

4. Model Selection and Ordered Testing.

In the previous sections, we developed method for controlling FDR for ordered hypothesis testing, and showed that our methods out-perform existing alternatives in simulations. This work was motivated by recent developments in the model selection literature, where a string of papers following Lockhart et al. (2014) showed that several model selection problems could be usefully analyzed with ordered hypothesis tests using a "covariance test statistic." Here, we provide a brief review of some key results stemming from that literature.

The original paper of Lockhart et al. constructs asymptotic p-values for steps along the lasso path. In a similar vein, G'Sell, Taylor and Tibshirani (2013) develop asymptotic p-values for hierarchical clustering and the graphical lasso. Taylor et al. (2013, 2014) exit the asymptotic regime, and provide exact finite-sample p-values for a wider class of selection problems such as the group lasso and principal component analysis (PCA). Lee et al. (2013) use similar ideas to do inference for the lasso with a fixed regularization parameter.

The test statistics provided by these methods fall into two categories: Lockhart et al. (2014) and G'Sell, Taylor and Tibshirani (2013) produce harmonic p-values that get larger as we get deeper into the null, while Taylor et al. (2013, 2014) give exact p-values inside the null. In Section 5, we show how we can exploit harmonic behavior to gain more power in sequential testing.

Finally, we emphasize that our ordered FDR-controlling procedures are not limited to p-values from the papers described above. In particular, p-values obtained with permutation or bootstrap based approaches would also feed naturally into our stopping rules.

4.1. *The Covariance Test for the Orthogonal Lasso.*

Consider a linear regression model with response $y \in \mathbb{R}^n$ and predictor matrix $X \in \mathbb{R}^{n \times p}$:

$$y = X\beta + \varepsilon, \text{ with } \varepsilon \sim \mathcal{N}(0, \sigma^2). \tag{16}$$

Suppose, moreover, that σ is known, and that X is orthogonal. For any given $\lambda > 0$, the lasso estimator (Tibshirani, 1996) is

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{2} \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 .$$

This solution, as a function of λ, gives the *lasso path* $\hat{\beta}(\lambda)$.

Lockhart et al. (2014) construct test statistics corresponding to segments of the lasso path. In the case of orthogonal X, these test statistics have the particularly simple form

$$T_k = \lambda_k(\lambda_k - \lambda_{k+1}), \tag{17}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$ are the values of the regularization parameter where the sparsity of $\hat{\beta}$ changes.

Suppose that there are s true signal (non-null) variables. At each step along the lasso path, we consider testing the hypothesis $H_{0,j}$ that all signal predictors are contained in the current lasso model (the model with $j - 1$ predictors). In this case, Lockhart et al. show that if the s signal variables are entered into the model, then when X is orthogonal the subsequent test statistics T_{s+1}, ..., $T_{s+\ell}$ are asymptotically jointly distributed as independent exponential random variables:

$$(T_{s+1}, \ldots, T_{s+\ell}) \Rightarrow \left(\text{Exp}(1), \text{Exp}\left(\frac{1}{2}\right), \ldots, \text{Exp}\left(\frac{1}{\ell}\right)\right), \tag{18}$$

in the limit where $n, p \to \infty$; this result holds for any finite ℓ. Note that the test statistics T_i get smaller as we go deeper into the null. In Section 5, we show how to exploit this phenomenon to get more power. The test statistics of G'Sell, Taylor and Tibshirani (2013) exhibit similar harmonic behavior.

4.2. *Exact Testing for Least-Angle Regression.* Lockhart et al. (2014) also provide test statistics for non-orthogonal design matrices X. However, in the case of non-orthogonal designs, we obtained better results using the exact test statistics for the least-angle path developed by Taylor et al. (2014), called the *spacing test statistics.*

The first spacing test statistic T_1 has a simple form. Given a standardized design matrix X Taylor et al. (2014) show that, under the global null hypothesis $\beta = 0$,

$$T_1 = \frac{1 - \Phi\left(\frac{\lambda_1}{\sigma}\right)}{1 - \Phi\left(\frac{\lambda_2}{\sigma}\right)} \overset{H_{0,q}}{\sim} \text{Unif}(0, 1) \tag{19}$$

Thus T_1 provides a p-value for a test of the global null hypothesis. Remarkably, this result holds exactly for any for any design matrix X, and does not require n or p to be large. The spacing test is asymptotically equivalent to the covariance test for the first null variable.

Taylor et al. (2014) also derive similar test statistics for subsequent steps along the least-angle regression path, which can be used for testing whether a partial regression coefficient is zero. Assuming Gaussian noise, all the null p-values produced by this test are 1-dependent and uniformly distributed over [0, 1]. For the purpose of our demonstrations, we apply our generic FDR control procedures directly as though the p-values were independent.

5. False Discovery Rate Control for Harmonic Test Statistics.

Our *ForwardStop* and *StrongStop* procedures control the FDR for generic sequential selection problems where we have uniform p-values, and are appropriate for, e.g., the spacing test statistics for the least-angle path. However, in the orthogonal Lasso and the hierarchical clustering settings, we saw that l-th null test statistic had Exp$(1/\ell)$ distribution. To obtain p-values, one could transform these like Exp(1) variables, but the results would be wildly conservative. In this section, we design a stopping rule that is similar to *StrongStop*, but is able to take full advantage of the Exp$(1/\ell)$ behavior to dramatically increase power.

Suppose that we have a sequence of test statistics T_1, ..., $T_m \geq 0$ corresponding to m hypotheses. The first s test statistics correspond to signal variables; the subsequent ones are independently distributed as

$$(T_{s+1}, \ldots, T_m) \sim \left(\text{Exp}(1), \text{Exp}\left(\frac{1}{2}\right), \ldots, \text{Exp}\left(\frac{1}{m-s}\right)\right), \tag{20}$$

where $\mathrm{Exp}(\mu)$ denotes the exponential distribution with mean μ. As before, we wish to construct a stopping rule that controls the FDR.

We could try to apply either *ForwardStop* or *StrongStop*, using p-values based on (20). Of course, doing so would require knowledge of the number of signal variables s, and hence would not be practical. Fortuitously, however, an extension of this idea yields a variation of *StrongStop* that does not require knowledge of s and controls FDR.

Under (20), we have $j \cdot T_{s+j} \sim \mathrm{Exp}(1)$. Using this fact, suppose that we knew s and formed the *StrongStop* rule for the $m-s$ null test statistics. This would suggest a test based on

$$q_i^* = \exp\left[-\sum_{j=i}^{m} \frac{\max\{1,\, j-s\}}{j}\, T_j\right] \tag{21}$$

This is not a usable test, since it depends on knowledge of s. Now suppose we set $s = 0$, giving

$$q_i^* = \exp\left[-\sum_{j=i}^{m} T_j\right] \tag{22}$$

An application of the BH procedure to the q_i^* leads to the following rule.

PROCEDURE 3 (TailStop). *Let q_i^* be defined as in (22). We reject hypotheses* 1, ..., $\hat{k}_T$, *where*

$$\hat{k}_T = \max\left\{k : q_k^* < \frac{\alpha k}{m}\right\}. \tag{23}$$

Now the choice $s = 0$ is anti-conservative (in fact, it is the least conservative possibility for s), and so as expected we lose the strong control property of *StrongStop*. But surprisingly, in the idealized setting of (20), *TailStop* controls the FDR nearly exactly.

THEOREM 3. *Given (20), the rule from (23) controls FDR at level α. More precisely,*

$$\mathbb{E}\left[\frac{\left(\hat{k}_T - s\right)_+}{\max\left\{\hat{k}_T, 1\right\}}\right] = \alpha\,\frac{m-s}{m}.$$

The name *TailStop* emphasizes the fact that this procedure starts scanning the test statistics from the back of the list, rather than from the front. Scanning from the back allows us to adapt to the harmonic decay of the null p-values without knowing the number s of non-null predictors. An analogue to *ForwardStop* for this setup would be much more difficult to implement, as we would need to estimate s explicitly.

Remark: Current results yielding $\mathrm{Exp}(1/\ell)$ null distributions are asymptotic. In practice, the $\mathrm{Exp}(1/\ell)$ behavior has been observed to break down in these settings for large ℓ. A practical correction for this is to assume a more conservative null distribution for large values of ℓ by truncating the harmonic decay. Fixing $\tau \in \mathbb{N}$ ($\tau = 10$ works well) and letting s be the number of true steps, we assume $T_{s+\ell} \dot{\sim} \text{Exp}(1/\ell)$ for $\ell \leq \tau$, and $T_{s+\ell}$ stochastically smaller than $\text{Exp}(1/\tau)$ for $\ell > \tau$. The *TailStop* rule then becomes $q_i = \exp\left(-\sum_{j=i}^{n} \frac{\min(j,\tau)}{j} T_j\right)$. These q_i are strictly larger than in the original *TailStop* procedure, so the theoretical guarantees from Theorem 3 continue to hold.

6. Model Selection Experiments.

In this section, we demonstrate our stopping rules from Section 2 and 5 on sequential p-values for the lasso with orthogonal X and LARS with non-orthogonal X. As discussed in Section 4, these setups demonstrate the two most common types of distributional guarantees in the recent literature on sequential tests for model selection.

The goal of these experiments is to see what is the best way of transforming sequential p-values into a model selection rule with inferential guarantees; thus, we compare our proposed methods with the baselines discussed in Section 3.

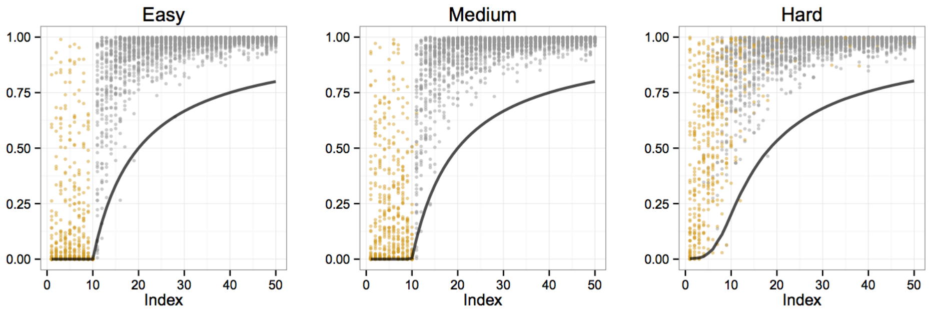

FIG 4. *Observed p-values for* 50 *realizations of the covariance test (Lockhart et al., 2014) with orthogonal X. p-values corresponding to non-null hypotheses are shown in orange, while those corresponding to null hypotheses are shown in gray. The smooth black curve is the average proportion of null hypotheses up to the given index. Note that these p-values behave very differently from those in the ordered hypothesis example presented in §3. The null p-values here exhibit* $\text{Exp}(1/\ell)$ *behaviour, as described in §4.1.*

6.1. *Simulations for the Covariance Test for the Lasso with Orthogonal X.*

In this section, we demonstrate these rules on the lasso with orthogonal X. Though we don't show it here for space concerns, very similar behavior appears when applying our rules to the hierarchical clustering setting.

We consider three scenarios which we once again refer to as easy, medium and hard. In all of the settings we have $n = 200$ observations on $p = 100$ variables of which 10 are non-null, and standard normal errors on the observations. The non-zero entries of the parameter vector β are taken to be equally spaced values from 2γ to $\sqrt{2 \log p}\gamma$, where γ is varied to set the difficulty of the problem. Figure 4 shows p-values from 50 realizations of each simulation setting; they exhibit harmonic behavior as described in Section 4.1.

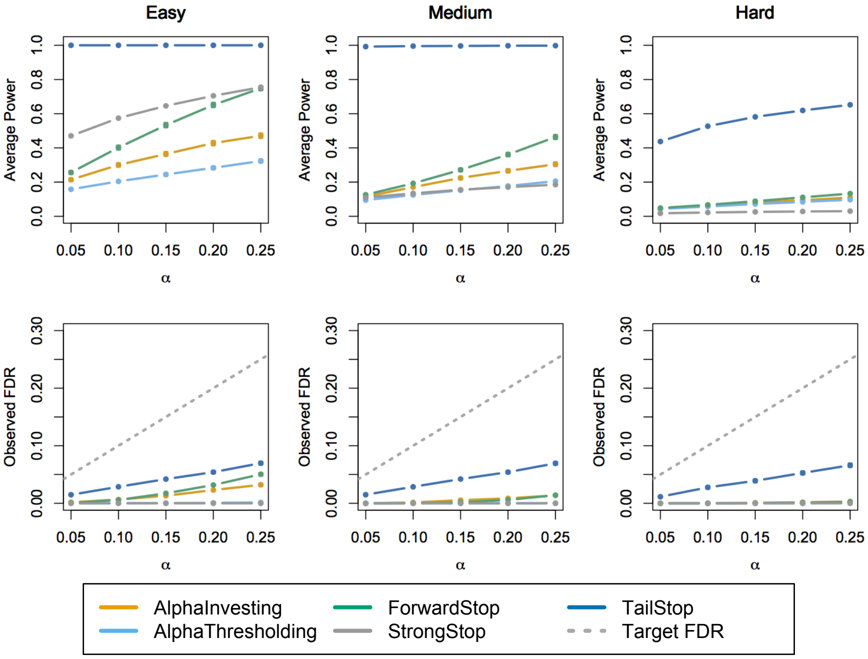

FIG 5. *Average power and observed FDR level for the orthogonal lasso using the covariance test of Lockhart et al. (2014). In the bottom panels, we see that all methods achieve FDR control. By taking advantage of the* $\mathrm{Exp}(1/\ell)$ *behaviour of the null p-values,* TailStop *far outperforms the other methods in power across all the difficulty settings.*

We compare the performance of *StrongStop*, *ForwardStop*, α-investing and α-thresholding. In this comparison, TailStop operates on the raw test statistics of Section 4.1. All other procedures operate on conservative p-values, $p_j = \exp(-T_j)$, obtained by bounding the null distributions by $\mathrm{Exp}(1)$.

Figure 5 shows plots of average power and observed FDR level across the three simulation settings. The FDR plots confirm that all of the procedures control the FDR. However, in the medium and hard settings *TailStop* is the only method that shows sensitivity to the choice of target α level. All other methods have an observed FDR level that's effectively 0, irrespective of the target α.

From the power plots we also see that *TailStop* has far higher power than the other procedures. The difference in power is particularly pronounced in the medium signal strength setting, where at low α *TailStop* achieves almost 10x higher power than any other method. The superior performance of *TailStop* is both desirable and expected, as it is the only rule that can take advantage of the rapid decay of the test statistics in the null.

6.2. *Simulations for the Spacing Test for Least-Angle Regression.*

As discussed in more detail in Section 4.2, Taylor et al. (2014) propose a sequence of spacing test p-values corresponding to each step in the Least Angle Regression (LARS) path.

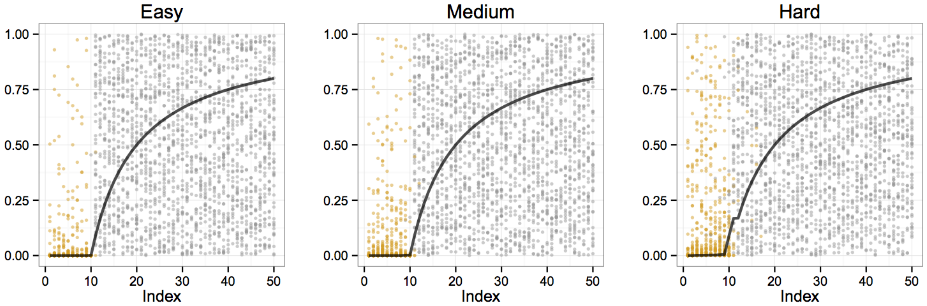

FIG 6. *Observed p-values for 50 realizations of the spacing test (Taylor et al., 2014) for LARS. p-values corresponding to non-null hypotheses are shown in orange, while those corresponding to null hypotheses are shown in gray. The smooth black curve is the average proportion of null hypotheses up to the given index. This example is similar to the Easy setting of the ordered hypothesis example of §3 in that the null and alternative are nearly perfectly separated. However, in the LARS setting the p-values under the alternative are highly variable and can be quite large, particularly in the Hard setting.*

These p-values are shown to be independent and uniformly distributed under the null hypothesis that the given variable has 0 regression coefficient in the model containing itself and all previously added variables. We compare the performance of *ForwardStop*, *StrongStop*, α-investing and α-thresholding in three settings of varying signal strength. *TailStop* is not included in this comparison because it should only be used when the null p-values exhibit $\text{Exp}(1/\ell)$ behaviour, whereas the spacing test p-values are uniform.

In all three settings we have $n = 200$ observations on $p = 100$ variables of which 10 are non-null, and standard normal errors on the observations. The design matrix X is taken to have iid Gaussian entries. The non-zero entries of the parameter vector β are taken to be equally spaced values from 2γ to $\sqrt{2 \log p}\gamma$, where γ is varied to set the difficulty of the problem.

Figure 6 shows p-values from 50 realizations of each simulation setting. Note that while all three settings have excellent separation—meaning that LARS selects all the signal variables before admitting noise variables—the p-values under the alternative can still be very large.

Figure 7 shows plots of average power and observed FDR level across the three simulation settings. All four successfully control FDR across the three simulation settings, and the control becomes increasingly conservative as the signal strength decreases. Interestingly, *StrongStop* attains both the highest average power and the lowest observed FDR level in all three settings. This is not unexpected, as *StrongStop* scans the p-values back-to-front and is therefore less hindered by stray occurrences of large p-values early in the LARS path.

7. Discussion: Practical considerations in model selection. When applying our procedures to model selection problems in practice, there are several practical concerns that arise. These can affect the power and FDR-controlling prop-

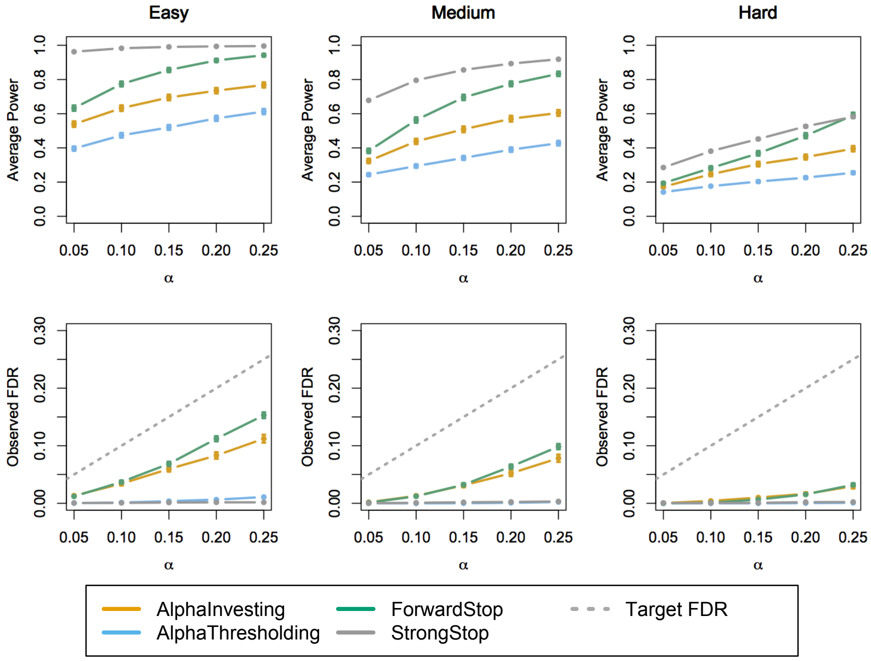

FIG 7. *Average power and observed FDR level for the spacing test p-values (Taylor et al., 2014). Even though there is nearly perfect separation between the null and alternative regions, the presence of large alternative p-values early in the path makes this a difficult problem.* StrongStop *attains both the highest average power and the lowest observed FDR across the simulation settings. Unlike the other methods,* StrongStop *scans p-values back-to-front, and is therefore able to perform well despite the occurrence of large p-values early in the path.*

erties of the procedures from the previous sections, as well as their interpretation. In this final section, we discuss some of these practical concerns, both in general and in the special case of the covariance testing literature reviewed above.

7.1. *Intermingling of signal and noise.* In finite samples with relatively weak signal, it is common that the signal selections and null selections can be intermingled in the selection path. In the lasso case, this happens when the signal variables are weak enough that the lasso path does not actually include the true solution at any point.

This means that there are two transitions that we could be trying to identify. One is the first time a noise selection is made. This is the interface between the region of pure signal and the intermingled region. The other is the last time a signal selection is made. This is the interface between the intermingled region and the region of only noise variables. When signal is strong, as in the conditions for Lockhart et al. (2014) and G'Sell, Taylor and Tibshirani (2013), these occur at the same point in the path. However, in practical settings they are often separated.

This means that one needs to be careful to specify which transition is intended when defining false discovery rates in these problems. This will depend on the specification of the null hypotheses for the particular test statistics being used. For the $\mathrm{Exp}(1/\ell)$ distributions we discuss in this paper, we are referring to selections that occur after the first noise variable enters as false selections (though the accuracy is impacted by the distributional distortion discussed later in this section). For some of the more exact tests and distributional guarantees that are being developed, the corresponding false discovery rate refers to null variables that enter after the last signal variable (rather than the first noise variable).

A more concrete impact of this intermingling is that it can theoretically distort the $\mathrm{Exp}(1/\ell)$ distributions of Lockhart et al. (2014) and G'Sell, Taylor and Tibshirani (2013). The existing theory for the lasso provides insight into the effects of this intermingling of noise and signal variables, since the noise variables enter independently of the weak signal variables in the orthogonal lasso.

Suppose that the ℓ and $\ell+1$ noise variables enter at λ_k and λ_{k+2}, but a weak signal variable enters at λ_{k+1}. According to the theory from Lockhart et al. (2014) for the noise variables, we actually expect $\lambda_k(\lambda_k - \lambda_{k+2})$ to be approximately $\mathrm{Exp}(1/\ell)$. The observed test statistics, $T_k = \lambda_k(\lambda_k - \lambda_{k+1})$ and $T_{k+1} = \lambda_{k+1}(\lambda_{k+1} - \lambda_{k+2})$, satisfy $T_k + T_{k+1} \le \lambda_k(\lambda_k - \lambda_{k+2}) \approx \mathrm{Exp}(1/\ell)$.

This is enough for *ForwardStop* and *StrongStop* to continue to control FDR (and FWER for *StrongStop*), since both T_k and T_{k+1} are still dominated by $\mathrm{Exp}(1)$. However, inserting these weak signal variables shifts the distributions in later steps to have larger means than expected (or equivalently shifts the locations of the noise variables). In the example above, T_{k+2} would be $\mathrm{Exp}(1/(\ell+1))$ instead of $\mathrm{Exp}(1/(\ell+2))$. Theoretically, this could lead the mean adjustments in *TailStop* to be too large, leading to anti-conservative behavior. In simulation, however, this anti-conservative behavior has been very difficult to produce.

7.2. *Correlated predictors in regression: Does FDR make sense as a criterion?.*

In regression settings, correlated predictors lead to further complications. While the $\mathrm{Exp}(1)$ bound of Lockhart et al. (2014) allows correlation, its conditions implicitly rely on those correlations being relatively weak. As a result, the distributional guarantee begins to break down in the presence of large correlations, as demonstrated in Table 2 of their paper. This can lead stopping rules based on these guarantees to be anti-conservative.

More recent approaches, like those of Taylor et al. (2013) and Taylor et al. (2014), provide guarantees that continue to hold in the presence of strong correlations. However, when the predictors are highly correlated, the appropriateness (and definition) of FDR as an error criterion comes into question. If a noise variable is highly correlated with a signal variable, should we consider it to be a false selection? This is a broad question that is beyond the scope of this paper, but is worth considering when discussing selection errors in problems with highly correlated X. This question is discussed in more detail in several papers (e.g. G'Sell, Hastie and Tibshirani, 2013; Lin, Foster and Ungar, 2011; Benjamini and Gavrilov, 2009; Wu, Boos and Stefanski, 2007).

8. Conclusions.

We introduced a new multiple hypotheses testing setting that has recently become important in some model selection problems. In this setting, the hypotheses are ordered, and all rejections are required to lie in an initial contiguous block. Because of this constraint, existing multiple testing approaches do not control criteria like the False Discovery Rate (FDR).

We proposed a pair of procedures, *ForwardStop* and *StrongStop*, for testing in this setting. We proved that these procedures control FDR at a specified level while respecting the required ordering of the rejections. Two procedures were proposed because they provide different advantages. *ForwardStop* is simple and robust to assumptions on the particular behavior of the null distribution. Meanwhile, when the null distribution is dependable, *StrongStop* controls not only FDR, but the Family-Wise Error Rate (FWER). We then applied our methods to model selection, and provided a modification of *StrongStop*, called *TailStop*, which takes advantage of the harmonic distributional guarantees that are available in some of those settings.

A variety of researchers are continuing to work on developing stepwise distributional guarantees for a wide range of model selection problems. As many of these procedures are sequential in nature, we believe that the stopping procedures from this paper provide a way to convert these stepwise guarantees into model selection rules with accompanying inferential guarantees. The simulations and comparisons in this paper suggest that our procedures perform favorably at this task.

A remaining challenge is to design stepwise p-values that yield particularly favorable performance when used in this fashion. For example, we have seen in Section 6 that the $\text{Exp}(1/k)$ null distributions available in some settings give a dramatic boost in power when used with appropriate stopping rules. However, null distributions of that form are only currently demonstrated in very specialized scenarios. Constructing stepwise statistics and distributions with such performance in more generalized scenarios is an important and open problem.

Acknowledgements. The authors are grateful to Jonathan Taylor, and to the Editor and Associate Editor for their helpful comments and suggestions. M.G., S.W. and A.C. are respectively supported by a NSF GRFP Fellowship, a BC and EJ Eaves SGF Fellowship, and a NSERC PGSD Fellowship; R.T. is supported by NSF grant DMS-9971405 and NIH grant N01-HV-28183.

References.

Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *The Annals of Applied Statistics* **3** 179–198.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

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

Bogdan, M., Berg, E. v. d., Su, W. and Candes, E. (2013). Statistical estimation and testing via the ordered L1 norm. *arXiv preprint arXiv:1310.1969*.

Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association* 1151-1160.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407–499. With discussion, and a rejoinder by the authors. MR2060166 (2005d:62116)

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

G'Sell, M. G., Hastie, T. and Tibshirani, R. (2013). False Variable Selection Rates in Regression. *arXiv preprint arXiv:1302.2303*.

G'Sell, M. G., Taylor, J. and Tibshirani, R. (2013). Adaptive testing for the graphical lasso. *arXiv preprint arXiv:1307.4765*.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2013). Exact inference after model selection via the Lasso. *arXiv preprint arXiv:1311.6238*.

Lin, D., Foster, D. P. and Ungar, L. H. (2011). VIF regression: A fast regression algorithm for large data. *Journal of the American Statistical Association* **106** 232–247.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics (with Discussion)*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72** 417–473.

Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Hungarica* **4** 191–231.

Shah, R. D. and Samworth, R. J. (2012). Variable selection with error control: Another look at Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.

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

Taylor, J., Loftus, J., Tibshirani, R. and Tibshirani, R. (2013). Tests in adaptive regression via the Kac-Rice formula. *arXiv preprint arXiv:1308.3020*.

Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for Least Angle Regression and the Lasso. *arXiv preprint arXiv:1401.3889*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.

Wu, Y., Boos, D. D. and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102** 235–243.

APPENDIX A: PROOFS

Lemma 1. We can map any rejection threshold t to a number of rejections k. For the purpose of this proof, we will frame the problem as how to choose a rejection threshold $\hat{t}$; any choice of $\hat{t} \in [0,1]$ immediately leads to a rule

$$\hat{k}_F = R(\hat{t}) = \left|\{i : q_i < \hat{t}\}\right|.$$

Similarly, the number of false discoveries is given by $V(\hat{t}) = \left|\{i > s : q_i < \hat{t}\}\right|$. We define the threshold selection rule

$$\hat{t}_\alpha = \max\left\{t \in [0,1] : t \leq \frac{\alpha\, R(t)}{m}\right\}.$$

Here, $R(\hat{t}_\alpha) = \hat{k}_F$ and so this rule is equivalent to the one defined in the hypothesis.

When coming in from 0, $R(t)$ is piecewise continuous with upwards jumps, so

$$\hat{t}_\alpha = \frac{\alpha\, R(\hat{t}_\alpha)}{m},$$

allowing us to simplify our expression of interest:

$$\frac{V(\hat{t}_\alpha)}{R(\hat{t}_\alpha)} = \frac{\alpha}{m}\,\frac{V(\hat{t}_\alpha)}{\hat{t}_\alpha}.$$

Thus, in order to prove our result, it suffices to show that

$$\mathbb{E}\left[\frac{V(\hat{t}_\alpha)}{\hat{t}_\alpha}\right] \leq m.$$

The remainder of this proof establishes the above inequality using Rényi representation and a martingale argument due to Storey, Taylor and Siegmund (2004).

Recall that, by assumption, $p_{s+1}, \dots, p_m \overset{\text{iid}}{\sim} U([0,1])$. Thus, we can use Rényi representation to show that

$$\begin{aligned}(Z_{s+1} - Z_s, \dots, Z_m - Z_s) &= \left(\frac{Y_{s+1}}{m-s}, \dots, \sum_{i=s+1}^{m} \frac{Y_i}{m-i+1}\right) \\ &\overset{d}{=} \left(E_{1,\,m-s}, \dots, E_{m-s,\,m-s}\right),\end{aligned}$$

where the $E_{i,\,m-s}$ are standard exponential order statistics, and so

$$\left(e^{-(Z_{s+1}-Z_s)}, \dots, e^{-(Z_m-Z_s)}\right)$$

are distributed as $m-s$ order statistics drawn from the uniform $U([0,1])$ distribution. Recalling that

$$1 - q_{s+i} = (1-q_s)\, e^{-(Z_{s+i}-Z_s)},$$

we see that $q_{s+1}, \dots, q_m$ are distributed as uniform order statistics on $[q_s, 1]$.

Because the last q_i are uniformly distributed,

$$M(t) = \frac{V(t)}{(t-q_s)}$$

is a martingale on $(q_s, 1]$ with time running backwards. Moreover, $\hat{t}_\alpha$ is a stopping time with respect to the relevant filtration. Thus, by the optional sampling theorem (plus some integrability arguments),

$$\mathbb{E}\left[M(\hat{t}_\alpha); \hat{t}_\alpha > q_s\right] \leq M(1) = \frac{m-s}{1-q_s}.$$

For all $t > q_s$,

$$\frac{V(t)}{t} = \frac{t-q_s}{t} M(t) \leq (1-q_s)\, M(t),$$

and so

$$\mathbb{E}\left[\frac{V(\hat{t}_\alpha)}{\hat{t}_\alpha}; \hat{t}_\alpha > q_s\right] \leq m - s.$$

Meanwhile,

$$\mathbb{E}\left[\frac{V(\hat{t}_\alpha)}{\hat{t}_\alpha} | \hat{t}_\alpha \leq q_s\right] = 0, \text{ and so, as claimed, } \mathbb{E}\left[\frac{V(\hat{t}_\alpha)}{\hat{t}_\alpha}\right] \leq m.$$

□ □

COROLLARY 1. We can extend our original list of p-values p_1, ..., p_m by appending additional terms

$$\tilde{p}_{m+1},\, \tilde{p}_{m+2},\, ...,\, \tilde{p}_{m^*} \overset{\text{iid}}{\sim} U([0,1])$$

to it. This extended list of p-values still satisfies the conditions of Lemma 1, and so we can apply procedure (8) to this extended list without losing the FDR control guarantee:

$$\hat{k}_F^{q,m^*} = \max\left\{k : \frac{m^* q_k^{m^*}}{k} \leq \alpha\right\}.$$

As we take $m^* \to \infty$, we recover the procedure described in the hypothesis:

$$\lim_{m^* \to \infty} \hat{k}_F^{q,m^*} = \hat{k}_F.$$

Thus, by dominated convergence, the rule $\hat{k}_F$ controls the FDR at level α. □ □

THEOREM 1. The proof of Lemma 1 used quantities

$$Z_i = \sum_{j=1}^{i} \frac{Y_j}{m-j+1} = \sum_{j=1}^{i} \frac{Y_j}{|\{l \in \{j,\, ...,\, m\}\}|}$$

to construct the sorted test statistics q_i. The key difference between the setup of Lemma 1 and our current setup is that we can no longer assume that if the i^{th} hypothesis is null, then all subsequent hypotheses will also be null.

In order to adapt our proof to this new possibility, we need to replace the Z_i with

$$Z_i^{ALT} = \sum_{j=1}^{i} \frac{Y_j}{\nu(j)}, \quad \text{where } \nu(j) = |\{l \in \{j,\, ...,\, m\} : l \in M\}|,$$

and M is the set of indices corresponding to null hypotheses. Defining

$$q_i^{ALT} = 1 - e^{-Z_i^{ALT}},$$

we can use Rényi representation to check that these test statistics have distribution

$$1 - q_i^{ALT} \overset{d}{=} r(i)\, \left(1 - U_{\nu(i),\,|M|}\right) := \exp\left[-\sum_{\{j \leq i : j \notin M\}} \frac{Y_j}{i}\right] \left(1 - U_{\nu(i),\,|M|}\right),$$

where the $U_{\nu(j),\,|M|}$ are order statistics of the uniform $U([0,1])$ distribution. Here $r(i)$ is deterministic in the sense that it only depends on the location and position of the non-null p-values.

If we base our rejection threshold $\hat{t}_\alpha^{ALT}$ on the q_i^{ALT}, then by an argument analogous to that in the proof of Lemma 1, we see that

$$\frac{V\left(\hat{t}_\alpha^{ALT}\right)}{\hat{t}_\alpha^{ALT}}$$

is a sub-martingale with time running backwards. The key step in showing this is to notice is that, now, the decay rate of $V(t)$ is accelerated by a factor $r^{-1}(i) \geq 1$. Thus, the rejection threshold $\hat{t}_\alpha^{ALT}$ controls FDR at level α in our new setup where null and non-null hypotheses are allowed to mix.

Now, of course, we cannot compute the rule $\hat{t}_\alpha^{ALT}$ because the Z_i^{ALT} depend on the unknown number $\nu(j)$ of null hypotheses remaining. However, we can apply the same trick as in the proof of Corollary 1, and append to our list an arbitrarily large number of p-values that are known to be null. In the limit where we append infinitely many null p-values to our list, we recover the *ForwardStop* rejection threshold. Thus, by dominated convergence, *ForwardStop* controls the FDR even when null and non-null hypotheses are interspersed. □ □

THEOREM 2. We begin by considering the global null case. In this case, the $\widetilde{Y}_i$ are all standard exponential, and so by Rényi representation the $\tilde{q}_i$ are distributed as the order statistics of a uniform $U([0,1])$ random variable. Thus, under the global null, the rule $\hat{k}_S$ is just Simes' procedure (Simes, 1986) on the $\tilde{q}_i$. Simes' procedure is known to provide exact α-level control under the global null, so (15) holds as an equality under the global null.

Now, consider the case where the global null does not hold. Suppose that we have $\hat{k}_S = k > s$. From the definition of $\tilde{q}_k$, we see that $\tilde{q}_k$ depends only on $p_k, \ldots, p_m$, and so the event $\tilde{q}_k \leq \alpha k/m$ is just as likely under the global null as under any alternative with less than k non-null p-values. Thus, conditional on s,

$$\sum_{k=s+1}^{m} \mathbb{P}\left[\hat{k}_S = k \middle| \text{alternative}\right] = \sum_{k=s+1}^{m} \mathbb{P}\left[\hat{k}_S = k \middle| \text{null}\right] \leq \alpha,$$

and so the discussed procedure in fact provides strong control. □ □

THEOREM 3. Let $Z_i^* = \sum_{j=i}^m T_i$. By Rényi representation,

$$\left(Z_{s+1}^*, \ldots, Z_m^*\right) \sim \left(E_{m-s,\,m-s}, \ldots, E_{1,\,m-s}\right),$$

where the $E_{i,\,j}$ are exponential order statistics. Thus

$$\left(q_{s+1}^*, \ldots, q_m^*\right)$$

are distributed as $m-s$ order statistics drawn from the uniform $U([0,1])$ distribution.

All the q_i^* corresponding to null variables are jointly distributed as independent uniform random variables. Thus, by the BH procedure, *TailStop* controls the FDR. The exact equality also follows directly from the result of Benjamini and Hochberg (1995). □ □

APPENDIX B: ADDITIONAL SIMULATIONS

In this section we revisit the ordered hypothesis example introduced in section 3 and present the results of a more extensive simulation study. We explore the following perturbations of the problem:

(a) Varying signal strength while holding the level of separation fixed. (Figures 8, 9, 10)
(b) Increasing the number of hypotheses while retaining the same proportion of non-null hypotheses (Figure 11)
(c) Varying the proportion of non-null hypotheses (Figures 12, 13, 14)

We remind the reader of the three simulation settings introduced in 3, which we termed Easy, Medium and Hard. These settings were defined as follows

Easy Perfect separation (all alternative precede all null), and strong signal (Beta$(1, 23)$)
Medium Good separation (mild intermixing of hypotheses), and moderate signal (Beta$(1, 14)$)
Hard Moderate separation (moderate intermixing hypotheses), and low signal (Beta$(1, 8)$)

All results are based on 2000 simulation iterations. Unless otherwise specified, the simulations are carried out with $m = 100$ total hypotheses of which $s = 20$ are non-null.

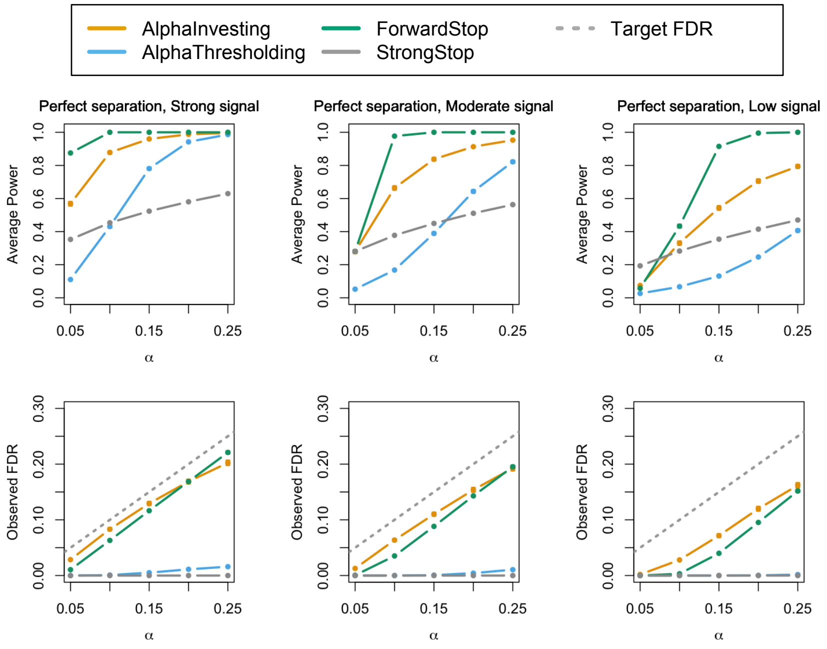

FIG 8. *Effect of signal strength on stopping rule performance: Perfect separation regime.* ForwardStop *remains the best performing method overall, except at the lowest α level in the moderate and low signal regimes. All of the methods become more conservative as the signal strength decreases.*

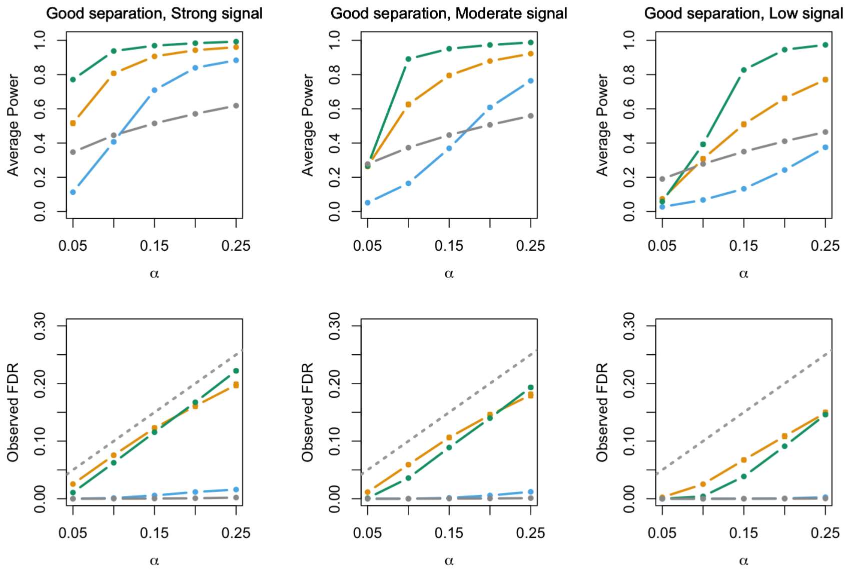

FIG 9. *Effect of signal strength on stopping rule performance: Good separation regime. The effect of signal strength is qualitatively the same as in the perfect separation regime.*

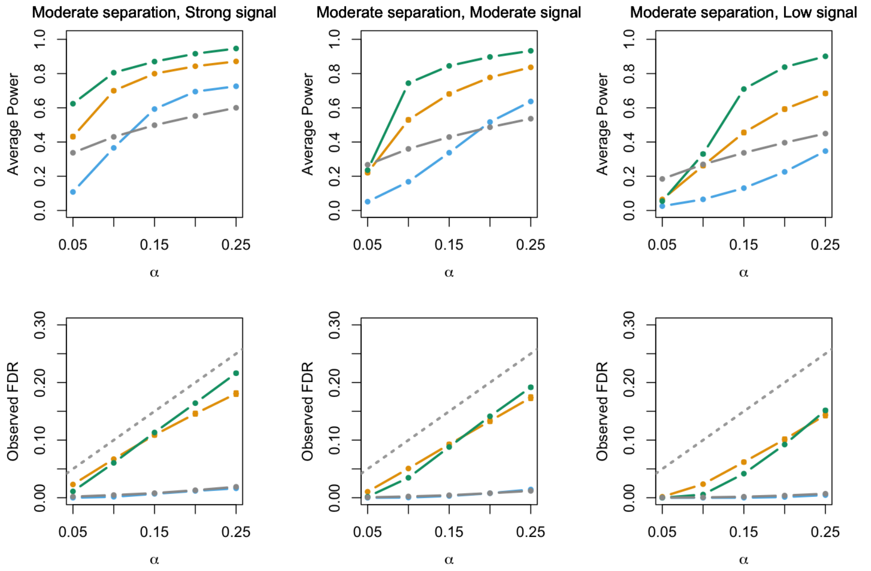

FIG 10. *Effect of signal strength on stopping rule performance: Moderate separation regime. The effect of signal strength is qualitatively the same as in the perfect separation and good separation regimes.*

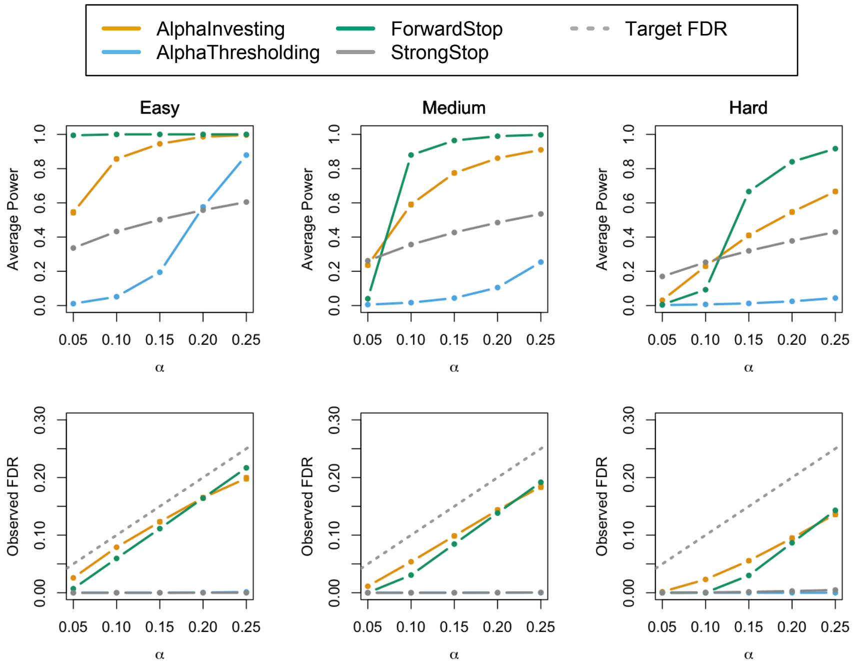

Fig 11. *Effect of increasing the total number of hypotheses. Instead of* 100 *hypotheses of which* 20 *are non-null, we consider* 1000 *hypotheses of which* 200 *are non-null. With the exception of α-thresholding, the performance of the methods remains largely unchanged. One small change is that* ForwardStop *loses power around $\alpha = 0.1$ in the Hard setting. The key difference is that the performance of α-thresholding considerably degrades. This is not surprising when we consider that α-thresholding is simply a geometric random variable. Thus as we increase the number of non-null hypotheses we expect the average power of α-thresholding to drop to* 0.

Department of Statistics
Stanford University
Sequoia Hall
Stanford, California 94305-4065
E-mail: maxg@stanford.edu
swager@stanford.edu
achould@stanford.edu
tibs@stanford.edu

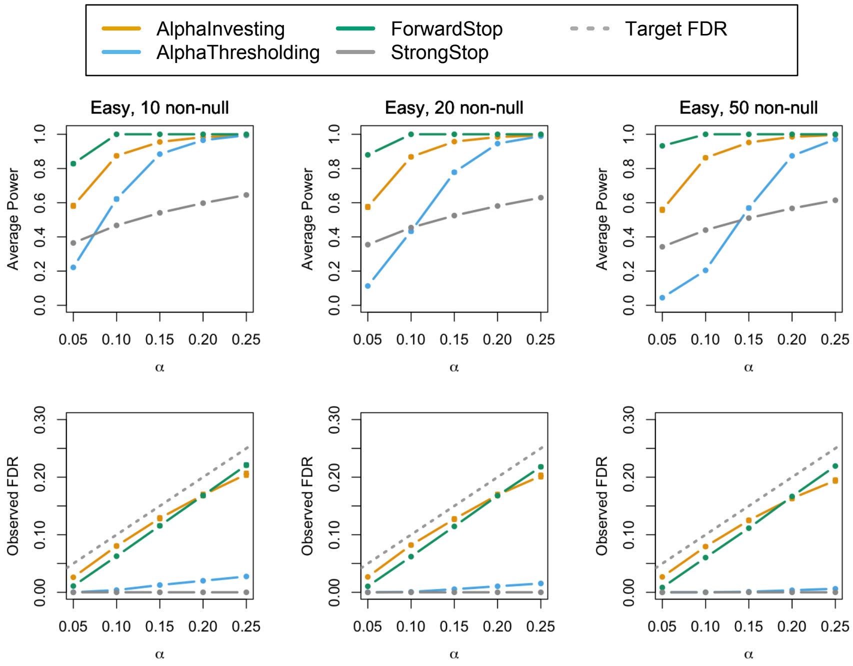

FIG 12. *Effect of varying the number of non-nulls out of $m = 100$ total hypotheses: Easy regime. With the exception of α-thresholding, the performance of the methods remains largely unchanged. The performance of α-thresholding degrades considerably as the number of non-null hypotheses increases. An explanation for this behaviour is presented in 11.*

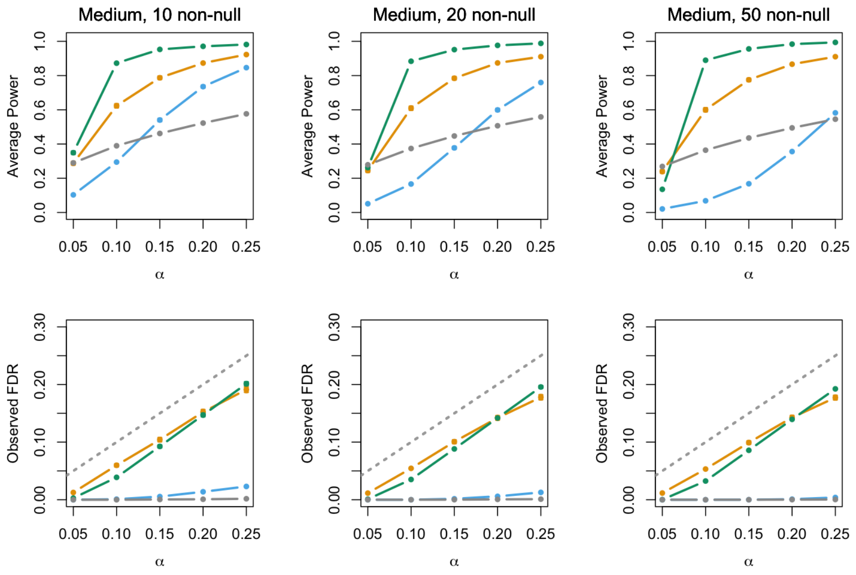

FIG 13. *Effect of varying the number of non-nulls out of $m = 100$ total hypotheses: Medium regime. The effect of varying the number of non-null hypotheses is qualitatively the same as in the Easy regime.*

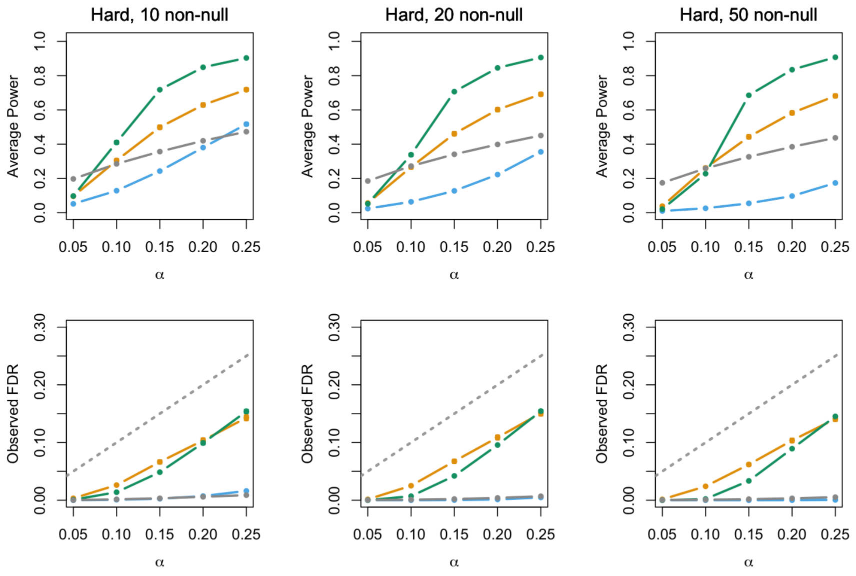

FIG 14. *Effect of varying the number of non-nulls out of $m = 100$ total hypotheses: Hard regime. The effect of signal strength is qualitatively the same as in the Easy and Medium regimes.*