Linear Models and Sequential Hypothesis Testing

Drew T. Nguyen

Abstract. This is an expository paper on FDR control of sequential hypotheses, with application to model selection in linear models with orthogonal design. We state the problem and outline, from start to finish, a complete method to perform model selection—including variations, recommendations, and some discussion of historical and current methods. Detailed proofs are provided for all main results and some of the important tools. In particular, the author proposes the use of a lesser-known test statistic, robust to nonorthogonality, and demonstrates its efficacy compared to a standard method.

Contents

Acknowledgements and Dedication 2
1. Introduction 2
2. Model Selection 3
3. The Lasso and p-values for Model Selection 5
4. Multiple Hypothesis Testing and the FDR 12
5. Continuous-time Martingales and the Optional Stopping Theorem 13
6. The Benjamini-Hochberg procedure 16
7. Sequential Hypothesis Testing 20
8. ForwardStop 22
9. Accumulation test procedures, SeqStep, and HingeExp 26
10. Simulation results 34
11. Discussion and Conclusion 35
12. References 38

Acknowledgements and Dedication

This is my undergraduate honors thesis in mathematics at UCSD, submitted in March 2018.

I would like to thank my parents, who would have liked me to be back home, but who supported my decision to go on working on projects at UCSD a little while longer; my brother, always a pillar of support; my thesis advisor Ian Abramson, whose floral lectures and dreamy attitude for statistics shaped my interest over the last several years; Jelena Bradic, whose class on statistical learning first got me into these problems; Ery Arias-Castro, whose occasional conversations helped me sharpen my understanding on my thesis topic in particular; Kenshi Yonezu and REOL, whose music got me through the winter I wrote it; and all the folks at NanoDesu Translations, for being such great friends, and missing me while I was gone. Sorry, guys, I have been a bit MIA. This is done now, though, so I'll see you again soon.

I dedicate this work to Madoka Kaname, who is an inspiration to us all.

1. Introduction

This work is motivated by the model selection problem in the usual linear regression setup with iid Gaussian error:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n \times p}, \boldsymbol{\beta}^* \in \mathbb{R}^p, \boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I})$.

Suppose we have a candidate estimate for the vector $\boldsymbol{\beta}^*$; call it $\hat{\boldsymbol{\beta}}$. Model selection (and its cousin variable selection) attempt to answer the problem of how to pick $\hat{\boldsymbol{\beta}}$ such that it has nonzero entries where $\boldsymbol{\beta}^*$ does, and zero entries where $\boldsymbol{\beta}^*$ does; a choice of nonzero $\hat{\beta}_i$'s is called a *model*, and if a $\hat{\beta}_i$ is nonzero we say it is included in the model.

One way to approach the problem is by considering an ordered sequence of models, constructed by some mechanism (to be specified): $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_N$ for some N, where each model $\mathcal{M}_k \subseteq \{1, ..., p\}$ indexes the included β_i's.

Given this sequence of models, how do we decide which $\mathcal{M}_k$ to settle on? Different approaches exist. One method is to associate each with a certain score function, and select the one that ranks highest (the score used is often the BIC score). Another method is to use cross-validation.

For the sake of this work, we will pursue a method based on sequential rejection of hypotheses, which, unlike the methods based on scores, give us probabilistic guarantees. This means that we will treat the $\mathcal{M}_k$ in order, using hypotheses for each that tests whether $\mathcal{M}_k$ is "right", per the model selection problem.

Specifically, we consider the hypotheses

$$H_{0k} : \text{supp}(\boldsymbol{\beta}^*) \subseteq \mathcal{M}_k, \quad H_{1k} : \text{supp}(\boldsymbol{\beta}^*) \not\subseteq \mathcal{M}_k \tag{2}$$

where $\text{supp}(\boldsymbol{\beta}^*) \subseteq \{1, ..., p\}$ denotes those true nonzero indices of the model. Though once again we note that even here there are alternatives; there are

different hypotheses one could use to quantify this idea of testing for the "right" $\mathcal{M}_k$, and they are not equivalent. We will discuss these at the end.

There are two parts to this method, and so the present work is modular, also in roughly two parts, plus some space for background material. One part is the problem of generating p-values corresponding to each of the hypotheses H_{0k}, this is section 2 and 3. The second part, which is all subsequent sections, is to combine these p-values in a sensible way to select a model while respecting multiple testing concerns.

First, in section 2 we begin by discussing difficulties involved in generating p-values, and then in section 3 describe an established method that works called the covariance test. Also in section 3 we pursue a suggestion by T. Tony Cai and Ming Yuan, and propose a second method which is more robust to inexact assumptions; Cai and Yuan did not provide a proof, but we provide it here. Second, we develop some background material in sections 4, 5, and 6, and then go into sequential hypothesis testing in sections 7, 8, and 9. That is, we consider p-values as generated from some distribution and forget about the calculation that produced them, and consider rejecting hypotheses in sequence using these p-values. The goal is to control an analogue of type-I error, called the FDR. Lastly, in section 10 we combine the p-values from generated from the first part using the techniques from the second part and compare performance between the covariance test and the proposed new statistic.

So as not to detract from the main points—namely p-value generation and FDR control—we will necessarily not prove everything. As far as facts not taught in yearly undergrad or graduate courses at UCSD, we'll use some basic properties of the lasso path, as well as some of the major theorems from extreme value theory and continuous time martingale theory. The relevant results require a lot of setup for their proofs, so much of that will be left to the references; in particular we will omit proofs for the needed standard results on lasso properties [1] [2], and extreme distributions [3], in part because those results can be simply stated, and we'll sometimes refer to them casually in prose. We will also assume the standard martingale results from discrete time, but then prove the main theorem we need for continuous martingales, because even though it takes some setup, the language of martingales is rather more technical and I find it illuminating to set them up in detail.

2. Model Selection

In the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, the p columns of $\boldsymbol{X}$ represent data corresponding to p *predictors*, while the rows correspond to the n samples. For instance, math test scores may depend on socioeconomic factors for n students. We may have their test score (an entry of $\boldsymbol{y}$) and collect n data rows with p entries (a row of $\boldsymbol{X}$), such as family income, age of parents, or whether they know a second language. In this case $p = 3$.

When fitting this model, we have to select which parameters to include and so which data to keep. Maybe we fear that are our parents' age aren't *really* relevant (though in this hypothetical situation, it is). A researcher may use either domain knowledge or, more often, statistical techniques to select correct parameters and thereby improve prediction and interpretability.

There are no shortage of statistical techniques for variable selection, but there are shortcomings with some of the more obvious (i.e. naive) procedures.

Here is a first try. Suppose we wish to test the nulls $H_{01}, ..., H_{0k}, ..., H_{0p}$ that $\beta_k = 0$ (from here I will simply write H_k for the kth null.) The Wald test statistic is

$$\frac{\hat{\beta}_k}{\widehat{\text{s.e.}}\left[\hat{\beta}_k\right]}$$

where $\hat{\beta}_k$ is obtained by OLS and the denominator is an estimate of the standard error (omitting here the exact formula). Under H_k, its distribution is $t_{n-(p+1)}$, so to perform variable selection, you can start with all the variables in the model, compute their p-values, and then leave out the variable with the largest p-value. Subsequently, you refit. This method is called "backwards selection".

There is no theoretical problem with this—but only as long as you don't ascribe any theoretical properties to it. Certainly what this will do is it will give us a sequence of variables $\beta_{j_1}, \beta_{j_2}, ...\beta_{j_p}$ to remove, but we cannot say we are doing valid hypothesis testing at each stage. After refitting, if we compute p-values again, then the new p-values we obtain have been affected by our *selection* of those variables and are thus invalid. In other words, the model we have selected and therefore the hypotheses we test have been influenced by our data. This is problematic, because we expect the hypotheses to be chosen beforehand.

Indeed, any procedure that reports p-values but does not account for the hypothesis having been selected based on the data is invalid. A procedure that does account for this selection is called an *adaptive procedure*.

As noted in the introduction, one does not *require* valid p-values or adaptive procedures to perform model selection. Backwards selection is still used, for instance, since it is a mechanism that gives us a sequence of models, and the researcher can select the set of variables with the best BIC score, or pick one based on cross-validation. Other methods in use are forward stepwise (like backwards elimination, but adds variables), forward stagewise (like forward stepwise, but more restrictive), and best subset regression.

These techniques remain controversial, however, because they do not give theoretical guarantees (except for best subset regression, which is computationally intractable for moderately sized datasets). But we can obtain some guarantees using adaptive procedures; in particular we can control how many type-I errors are made, as long as a correct null distribution can be derived.

In section 3 we introduce the lasso estimator, from which an adaptive procedure for model selection can be derived.

3. The Lasso and p-values for Model Selection

The *lasso estimator* is defined for the linear model (1) as the solution to the following optimization problem:

$$\hat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{3}$$

where one can think of $\lambda \geq 0$ as enforcing a penalty to the objective function based on the size of the coefficients of β. This has the property that as $\lambda \to \infty$, every $\beta_i \to 0$, while as $\lambda \to 0$, $\hat{\boldsymbol{\beta}}_\lambda \to \hat{\boldsymbol{\beta}}_{OLS}$, the ordinary least squares estimate.

For $A \subseteq \{1, ..., p\}$, we will also soon use the following definition:

$$\hat{\boldsymbol{\beta}}_A(\lambda) := \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}_A\boldsymbol{\beta}_A\|_2^2 + \lambda\|\boldsymbol{\beta}_A\|_1 \tag{4}$$

where a subscript $\boldsymbol{X}_A$ is the matrix $\boldsymbol{X}$ but only including those columns whose indices are in A, and likewise $\boldsymbol{\beta}_A$ includes only those entries from $\boldsymbol{\beta}$.

In practice, variables β_i both enter the model (go from zero to nonzero) and leave the model at finite values of λ. These values of λ are called *knots*, and we have $\infty > \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_N$, where λ_1 must be a knot where a variable enters the model and N is bounded above by 3^p. If the columns of $\boldsymbol{X}$ are in general position we have instead strict inequality, and the lasso solutions are unique [2]. If $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}$, we say that $\boldsymbol{X}$ satisfies orthogonal design, or simply that $\boldsymbol{X}$ is orthogonal. In this case variables never leave the model and $N = p$.

The lasso is used in high dimensional problems ($p > n$) because it is efficiently computable and has nice theoretical properties regarding consistency and support recovery[1], up to some technical conditions on $\boldsymbol{X}$. For us, the lasso is nice because it provides valid p-values to judge the significance of the models that it selects.

The null distribution will turn out to be exact for orthogonal $\boldsymbol{X}$, so we consider this scenario[2]; then we have that $N = p$. Consider a knot λ_k. Let $A_k \subseteq \{1, ..., p\}$ be the nonzero indices of $\boldsymbol{\beta}$ just before $\lambda = \lambda_k$ (this is called the *active set*) and suppose that at this knot the variable β_{i_k} enters the model. Let $\text{supp}(\boldsymbol{\beta}^*) \subseteq \{1, ..., p\}$ denote those true nonzero indices of the model.

With the hypotheses

$$H_{0k} : \text{supp}(\boldsymbol{\beta}^*) \subseteq A_k, \quad H_{1k} : \text{supp}(\boldsymbol{\beta}^*) \not\subseteq A_k \tag{5}$$

[1]Support recovery means that for some λ, the entries of $\hat{\boldsymbol{\beta}}_\lambda$ are zero where $\boldsymbol{\beta}^*$, and their signs match when they are not, so that this λ "recovers" the correct nonzero entries and signs.

[2]Certainly this is a restrictive condition; we discuss it at the end.

we define the test statistic

$$T_k = \frac{1}{\sigma^2}\left(\langle \boldsymbol{y}, \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_{k+1})\rangle - \langle \boldsymbol{y}, \boldsymbol{X}_{A_k}\hat{\boldsymbol{\beta}}_{A_k}(\lambda_{k+1})\rangle\right) \tag{6}$$

which is the difference in the (uncentered) covariance of the response $\boldsymbol{y}$ and the fitted values $\boldsymbol{X}\boldsymbol{\beta}$ with the inactive data left in and left out, normalized by the true variance. T_k is called the covariance test statistic.

Theorem 1 (Covariance Test)**.** *Under H_{0k} and with $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}$,*

$$T_k \xrightarrow{d} Exp(1) \tag{7}$$

as $p \to \infty$, with $n > p$, and their p-values are independent.

Before proceeding with the proof, we need a lemma. We only state it:

Lemma 1. *Under orthogonal design, T_k can be expressed in "knot form":*

$$T_k = \frac{1}{\sigma^2} \cdot \lambda_k(\lambda_k - \lambda_{k+1}) \tag{8}$$

The proof of Lemma 1 starts from a closed form for the lasso estimator $\hat{\boldsymbol{\beta}}$ given in [4] and proceeds through a (very) long chain of algebraic manipulations to obtain (8). The manipulations were done in great detail in the UCSD honors thesis [5] and we do not repeat them here.

We now use Lemma 1 and rewrite (8) into a form where the random variables involved are explicit. We will use a basic fact about the lasso, found in introductory sources such as [1]: Under orthogonal design, the lasso solution has the closed form

$$\hat{\beta}_j(\lambda) = S_\lambda(\hat{\beta}_j^{OLS}) \tag{9}$$

where $S_\lambda : \mathbb{R} \to \mathbb{R}$ is the soft-thresholding function

$$S_\lambda = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } -\lambda \le x \le \lambda \\ x + \lambda & \text{if } x < \lambda \end{cases}$$

We note also that under orthogonal design $\hat{\beta}_j^{OLS} = \boldsymbol{X}_j^T\boldsymbol{y}$. Let $U_j = \boldsymbol{X}_j^T\boldsymbol{y}$. Then the knots of the lasso are simply the values of λ where the coefficients become nonzero (cease to be thresholded):

$$\lambda_1 = \left|U_{(1)}\right|, \lambda_2 = \left|U_{(2)}\right|, ..., \lambda_p = \left|U_{(p)}\right| \tag{10}$$

where $\left|U_{(1)}\right| \ge \lambda_2 = \left|U_{(2)}\right| \ge ... \ge \lambda_p = \left|U_{(p)}\right|$ denote the order statistics in reverse order of $|U_1|, ..., |U_p|$. This is a slight abuse of notation—they are not the absolute values of the order statistics $U_{(1)}, ..., U_{(p)}$.

Under the null hypothesis, the entries of the OLS estimate are iid $\mathcal{N}(0, \sigma^2)$, so $|U_i|/\sigma^2 \overset{iid}{\sim} |\mathcal{N}(0,1)|$, and we write

$$T_k = \frac{1}{\sigma^2}\left|U_{(k)}\right| \cdot \left(\left|U_{(k)}\right| - \left|U_{(k+1)}\right|\right) \tag{11}$$

Proof of the covariance test. Because in the orthogonal case, variables never leave the model and $N = p$, we have $A_1 \subseteq A_2 \subseteq ... \subseteq A_p$. Then assuming the truth of the $H_{0k} : \text{supp}(\boldsymbol{\beta}^*) \subseteq A_k$ implies assuming the falsehood of every $H_{0k'} : \text{supp}(\boldsymbol{\beta}^*) \subseteq A_{k'}$ for $k' < k$. In other words, under H_{0k} all variables added to the model before step k are not null[3] and the remaining knots $\left|U_{(k)}\right|, ..., \left|U_{(p)}\right|$ are the order statistics of the remaining $|U_i|$, so $\left|U_{(k)}\right|$ may be thought of as the maximum of them and $\left|U_{(k-1)}\right|$ the second largest.

Hence we need only show the result for $T_1 = \left|U_{(1)}\right| \cdot \left(\left|U_{(1)}\right| - \left|U_{(2)}\right|\right)/\sigma^2$, which depends on a maximum and second largest order statistic, and the result follows with the exact same proof for the other T_k. For a similar reason, the resulting p-values are independent; the distribution we derive for T_k is conditional on the null being false for the $k' < k$ and true at k, while the $T_{k'}$ is conditional on the null being true at $k' < k$, which are disjoint events. So the p-values p_k are always independent of $p_1, ..., p_{k-1}$. Then any set of p-values $p_{k_1}, ..., p_{k_j}$ is mutually independent because

$$\begin{aligned} P(p_{k_1} \leq x_1, ..., p_{k_j} \leq x_j) &= P(p_{k_1} \leq x_1, ..., p_{k_{j-1}} \leq x_{j-1}) P(p_{k_j} \leq x_j) \\ &= ... = \prod_{i=1}^{j} P(p_{k_i} \leq x_i) \end{aligned}$$

Now we show the result for T_1. Each $|U_i|/\sigma^2$ has CDF

$$F(x) = (2\Phi(x) - 1)\mathbb{1}\{x > 0\}$$

We would like to know the distribution of its extreme order statistics as $p \to \infty$, and apply some basic extreme order statistics theory. Let H denote the left continuous inverse of $1/(1-F)$. Then Theorem 1.1.8, Remark 1.1.9, and Theorem 2.1.1. from de Haan and Ferreira (2006) [3] together immediately imply that if

$$\lim_{t \to \infty} \frac{(1 - F(t))F''(t)}{F'(t)^2} = -(\gamma + 1)$$

then for $a_p = H^{-1}(p)$ and $b_p = pF'(a_p)$, we have that $W_1 = b_p\left(\left|U_{(1)}\right| - a_p\right)$ and $W_2 = b_p\left(\left|U_{(2)}\right| - a_p\right)$ converge jointly in distribution:

$$(W_1, W_2) \xrightarrow{d} \left(\frac{E_1^{-\gamma} - 1}{\gamma}, \frac{(E_1 + E_2)^{-\gamma} - 1}{\gamma}\right)$$

where if $\gamma = 0$ the right hand side is interpreted as $(-\log(E_1), -\log(E_1 + E_2))$, which is its limit as $\gamma \to 0$, and the E's are iid Exp(1).

[3]Lockhart et al. assume some mild technical assumptions, but they don't condition on the first k' variables entering being those in the true model, and prove a stronger result in their Theorem 1 [6]. Since we'll always be rejecting hypotheses sequentially (see section 7), we won't need to do this.

Indeed $\gamma = 0$; we compute

$$\frac{(1-F(t))F''(t)}{F'(t)^2} = \frac{F''(t)}{F'(t)}\frac{1-F(t)}{F'(t)} = \frac{-2t\phi(t)}{2\phi(t)}\frac{2-2\Phi(t)}{2\phi(t)} = -t\cdot\frac{1-\Phi(t)}{\phi(t)}$$

so that

$$\lim_{t\to\infty}\frac{(1-F(t))F''(t)}{F'(t)^2} = \lim_{t\to\infty} -t\cdot\frac{1-\Phi(t)}{\phi(t)} = \lim_{t\to\infty} -t\cdot m(t)$$

where $m(t)$ is the Mills' ratio $(1-\Phi(t))/\phi(t)$ for the standard normal, for which it's well known that $m(t)\sim 1/t$ for large t. So the limit is -1, implying $\gamma = 0$ and

$$(W_1, W_2) \xrightarrow{d} (-\log(E_1), -\log(E_1+E_2))$$

Note that

$$\begin{aligned}\big|U_{(1)}\big|\cdot\big(\big|U_{(1)}\big| - \big|U_{(2)}\big|\big) &= (a_p + W_1/b_p)(W_1 - W_2)\frac{1}{b_p}\\ &= \frac{a_p}{b_p}(W_1-W_2) + \frac{W_1(W_1-W_2)}{b_p}\end{aligned}$$

Now, writing explicitly a_p and b_p, we have $a_p = F^{-1}(1-1/p) = \Phi^{-1}(1-1/2p)$ and $b_p = 2p\phi(a_p)$.

If we can show that $a_p/b_p \to 1$, then we are done, by the following argument. Because since $a_p\to\infty$ (since $F^{-1}(1)=\infty$) then $b_p\to\infty$ and would imply

$$\frac{a_p}{b_p}(W_1-W_2) + \frac{W_1(W_1-W_2)}{b_p} \to (W_1-W_2)$$

where (W_1-W_2) is asymptotically distributed as $\log(E_1+E_2)-\log(E_1) = -\log\Big(E_1/(E_1+E_2)\Big)$. The sum of n exponentials is Gamma$(n,1)$; a routine multivariate integral shows then that $E_1/(E_1+E_2)$ is uniform[4], so that (W_1-W_2) is asymptotically standard exponential (by inverse transform sampling), as we wanted to show.

To show that $a_p/b_p\to 1$, apply F to a_p to find $1-\Phi(a_p) = \frac{1}{2p}$, and recall $b_p = pF'(a_p)$. The Mills' ratio inequalities (well known, but can be found in [7]) say

$$\frac{x}{1+x^2}\cdot\phi(x) = \frac{1}{1+1/x^2}\cdot\frac{\phi(x)}{x} \le 1-\Phi(x) \le \frac{\phi(x)}{x}$$

Using $x = a_p$ and multiplying by $2p$ we find

$$\frac{1}{1+1/a_p^2}\cdot\frac{b_p}{a_p} \le 1 \le \frac{b_p}{a_p}$$

Taking $p\to\infty$ and and noting $a_p\to\infty$, as we observed earlier, we've shown $b_p/a_p\to 1$ by the squeeze theorem. □

[4]Or use the well known fact that for $X\sim$ Gamma$(\alpha_1,1)$ and $Y\sim$ Gamma$(\alpha_2,1)$, we have $\frac{X}{X+Y}$ is distributed as Beta(α_1,α_2), where $\alpha_1=\alpha_2=1$ is the special case of the standard uniform.

Before we move on we'll make a few remarks about the covariance test. Lockhart et al. in fact proved that this test is also applicable in the nonorthogonal setting, but then the true null distribution is stochastically smaller than Exp(1). What this means is you can control Type-I error if you wanted, but you've likely lost power.

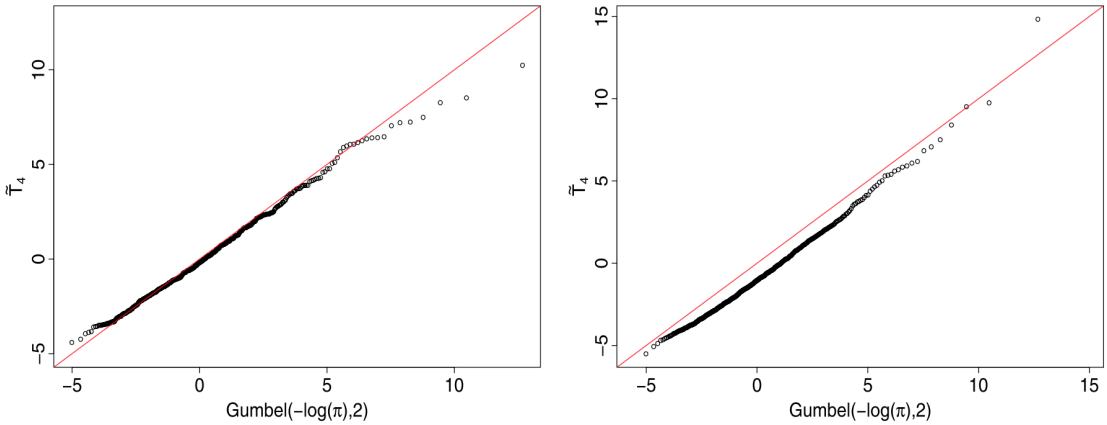
The covariance test is simple and easy to apply. I think it enjoys great popularity, being the one of the first tests of its kind to answer the question of adaptive p-values for model selection. And yet, as Lockhart et al. themselves note, it can be quite conservative in its rejection, such as in the nonorthogonal setting. Meanwhile, the more modern selective inference methods (I'll discuss some at the end) are complicated to state and to use, which makes them less appealing.

Now I would like to define what seems to me an underused and unknown test statistic, and prove its null distribution under the same null hypothesis as in (5). This statistic and associated hypothesis test, which I'll call the Lasso-G statistic (and test), was proposed by T. Tony Cai and Ming Yuan in a short discussion paper [8] in response to Lockhart et al. The Lasso-G statistic, which they denoted $\widetilde{T}_k$, was said to be asymptotically Gumbel$(-\log(\pi), 2)$. The CDF is

$$F(x) = \exp\left\{-e^{-(x+\log(\pi))/2}\right\}$$

The authors performed numerical experiments to show that it was robust to correlation between the features, and they reported that the that the covariance test statistic was not robust. We reproduce the plots from their paper in figure 1.

FIGURE 1. Experiments by T. Tony Cai and Ming Yuan for $\widetilde{T}_4$, computed from $\boldsymbol{X}$ generated by different ρ.



(a) $\rho = 0.2$ (b) $\rho = 0.8$

Figure 1 shows QQ-plots of $\widetilde{T}_4$ against the Gumbel. The authors choose $\boldsymbol{\beta} = (6, 6, 6, 0, 0, ...)^T$ for $n = 100$, $p = 50$ and generated $\boldsymbol{X}$ from a multivariate normal distribution where $\text{Cov}(X_i, X_j) = \rho^{|i-j|}$. Then they generated $\widetilde{T}_4$ for 500 independent datasets. The points match up well with Gumbel; we see that not only are they robust to ρ as big as 0.8, but that the approximation is quite good for moderately sized p.

Our treatment of the Lasso-G statistic in this section is theoretical; we provide a proof of its null distribution (which was not provided by Cai and Yuan). Our practical contribution is later, when we apply this statistic to the FDR control setting and show that it gets more power than the covariance test (see section 10).

Recall A_k is the active set just before the kth knot, and let $|A_k|$ be its cardinality. Suppose variables enter the lasso model in the order $\{j_1, j_2, ..., j_p\}$ and define $R_{j_k}(A_k) = \left(\text{RSS}_{A_k} - \text{RSS}_{A\cup\{j_k\}}\right)/\sigma^2$, which is the drop in residual sum of squares from OLS regression on only those variables in $\{A_k\}$ versus $\{A_k\}$ including the new variable entering the model. Then define the Lasso-G test statistic:

$$\widetilde{T}_k = R_{j_k}(A_k) - 2\log(|A_k|^c) + \log\log(|A_k|^c) \tag{12}$$

One can think of $-2\log(|A_k|^c) + \log\log(|A_k|^c)$ a correction factor to just the drop in RSS, which for the OLS problem is known to have a chi-square distribution. The factor corrects for the selection of $\{A_k\}$; without it we don't have an adaptive method.

Theorem 2 (Lasso-G test)**.** *Under H_{0k} and with $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}$,*

$$\widetilde{T}_k \xrightarrow{d} \textit{Gumbel}(-\log(\pi), 2) \tag{13}$$

as $p \to \infty$, with $n > p$, and their p-values are independent.

Proof. Let $\hat{\boldsymbol{\beta}}_A := \hat{\boldsymbol{\beta}}_A^{OLS}$ be the OLS solution regressed against just those variables in $A \subseteq \{1, ..., p\}$, and $\boldsymbol{X}_A$ be $\boldsymbol{X}$ just including those columns. Then

$$\text{RSS}_A = ||\boldsymbol{y} - \boldsymbol{X}_A\hat{\boldsymbol{\beta}}||_2^2$$

and expanding the sum and applying $\boldsymbol{X}^T\boldsymbol{X} = \mathbb{I}$ gives

$$\text{RSS}_A = \|\boldsymbol{y}\|_2^2 - 2\hat{\boldsymbol{\beta}}_A^T\boldsymbol{X}_A^T\boldsymbol{y} + ||\hat{\boldsymbol{\beta}}_A||_2^2$$

Again applying orthogonality, now using $\hat{\beta}_j^{OLS} = \boldsymbol{X}_j^T y$, we simplify

$$\text{RSS}_A = \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{y}^T\boldsymbol{y} + ||\hat{\boldsymbol{\beta}}_A||_2^2 = ||\hat{\boldsymbol{\beta}}_A||_2^2 - \|\boldsymbol{y}\|_2^2$$

so that the $\text{RSS}_A - \text{RSS}_{A\cup\{j\}}$ for some j not in A is

$$\text{RSS}_A - \text{RSS}_{A\cup\{j\}} = ||\hat{\boldsymbol{\beta}}_{A\cup\{j\}}||_2^2 - ||\hat{\boldsymbol{\beta}}_A||_2^2$$

which is maximized over j by picking the one with $|\boldsymbol{X}_j^T y|$ the largest. As for the lasso, in the case of orthogonal design, inspecting equations (10) tells

us that the jth variable to enter the model is also the one where $|\boldsymbol{X}_j^T y|$ is largest, out of all those that have not yet been added. So then

$$R_{j_k} = \max_{m \in A_k^c} R_m(A_k)$$

Classical regression theory says that each $R_m(A_k)$, under the null, has a χ_1^2 distribution. So R_{j_k} is a maximum of χ_1^2, whose distribution we can get through extreme value theory.

Before we do that let's note that the p-values are independent by the same reasoning as the proof of the covariance test. Also analogously to what we said in the covariance test, we only need to prove it for $\widetilde{T}_1$, because whenever we assume H_{0k} to be true, we can always think of R_{j_k} as a maximum order statistic of χ_1^2, and we may as well do the proof for R_{j_1}.

Let $V_1 \geq V_2 \geq ... \geq V_p$ be the order statistics of an absolute standard normal. Then every V_k is equal in distribution to the $U_{(k)}$ from the proof of the covariance test, and we can define $W_1 = b_p\,(V_1 - a_p)$ just as before and recall we proved it is distributed as $-\log(E)$ for E a standard exponential, when $a_p = \Phi^{-1}(1 - 1/2p)$ and $b_p = 2p\phi(a_p)$. The fact is that $-\log(E) \sim \text{Gumbel}(0, 1)$.

Now by Example 1.1.7 in de Haan and Ferreira, we could have chosen instead different constants a_p and b_p and still gotten this result. Following the logic in that example (with a minor modification since we have absolute standard normals, not standard normals)[5] we can redefine

$$a_p = \sqrt{2\log p} - \frac{\log\log p + \log \pi}{2\sqrt{2\log p}}$$

$$b_p = \sqrt{2\log p}$$

and have the same result of asymptotic convergence hold. Now

$$\frac{V_1 + a_p}{b_p} = \frac{V_1 - a_p}{b_p} + 2\frac{a_p}{b_p}$$

Clearly $\frac{a_p}{b_p} = 1 + o(1)$. But also $V_1 - a_p/b_p = W_1/b_p^2 \to 0$ almost surely, since $b_p \to \infty$ and W_1 goes to a nondegenerate distribution. So $V_1 + a_p/b_p \to 2$ almost surely, and then

$$V_1^2 - a_p^2 = W_1 \cdot \left(V_1 + a_p/b_p\right) \to 2 \cdot \text{Gumbel}(0, 1) = \text{Gumbel}(0, 2)$$

But after explictly squaring a_p^2 you can see $a_p^2 \approx 2\log(p) - \log\log(p) - \log(\pi)$ for large p. Rearranging, and noting that under the null R_{j_k} as χ_1^2 is distributed just like V_1^2 which is a squared normal, we have that $\widetilde{T}_k \to \text{Gumbel}(-\log(\pi), 2)$.

□

[5] If we'd had standard normals we would write $a_p = \sqrt{2\log p} - \frac{\log\log p + \log 4\pi}{\sqrt{2\log p}}$. According to a paper of Hall [9], this is, in a supremum error sense the best choice of constants for the standard normal (it converges fastest).

That's as far as we'll go for generation of p-values. Naturally this is not the end of that story; are there other statistics? Other tests? Other hypotheses? And are there methods that still have guarantees when $\boldsymbol{X}^T\boldsymbol{X} \neq \mathbb{I}$? Yes, of course, to all; since the covariance test was proposed, there came to exist something of a menagerie of methods, and they make up the field called *selective inference*. We mention some of these methods at the end. However, the covariance test stands out as the field's progenitor, and remains the simplest test. Meanwhile, the lasso-G test was simply another, simpler way to do the same thing, which may obtain better results under orthogonality and remain robust outside of it.

4. Multiple Hypothesis Testing and the FDR

We turn to multiple hypothesis testing—the problem of how to combine p-values to test multiple hypotheses, subject to false discovery concerns. The point of this section is to illustrate those concerns.

Consider now a genome-wide association study, which has a disease in mind and searches the entire human genome for genes that might be risk factors or causal factors for its development. There are tens of thousands of genes to search through, and each kth gene corresponds to another null hypothesis H_k—namely, does gene k contribute to the disease? As a different example, if one is searching for certain subsequences in a chromosome to understand the synthesis of a protein, the number of nulls towers into the millions [10].

Each hypotheses H_k has a p-value p_k, and we reject, say, if $p_k < t$, so t is some threshold. For N hypotheses, let $V(t) \leq N$ be a (unobserved) random variable equal to the number of false rejections.

Classically, to control type I error means to control for any mistake under the null hypothesis, so for a multiple testing analogue we may wish to choose t so that $P(V(t) \geq 1) < q$ for some desired q (this probability is called the family-wise error rate, or FWER). So we bound the probability that we falsely reject even once. But when we have millions of null hypotheses this is an unreasonably stringent criterion, because as we consider more and more nulls, it becomes more and more likely that a null p-value is less than α by chance. We could adjust t to make α very small, but then we will hardly ever reject, even when we should. So we have lost power.

Large datasets in technology and in biology have led researchers to the following attitude; making a few type I error type mistakes is not the end of the world. If we are interested the genes for which H_k is rejected, it is worth some false positives to have the actual interesting genes be revealed.

Let $R(t)$ be the total number of rejected hypotheses. There are some competitors for a looser criterion to control, but the standard one is the false discovery rate, or the FDR:

$$\mathrm{FDR}(t) = \mathbb{E}\left[\frac{V(t)}{\max(R(t), 1)}\right] \tag{14}$$

Type I error control becomes: Pick t so that $\text{FDR}(t) \leq q$ for some chosen threshold q. This is a loose criterion; controlling an expectation means that we may do badly on FDR control on any one experiment, but that on average we do not.

5. Continuous-time Martingales and the Optional Stopping Theorem

We have defined the FDR, but we have not actually given procedures for how to control it when testing N hypotheses.

Before we define these procedures and prove their FDR controlling properties, we need some results on martingales. This is essentially because the quantity $V(t)/t$, and other similar quantities (where recall $V(t)$ is our number of false rejections at t and t is our threshold) turn out to be martingales in t, or sub- or supermartingales. In any case, this allows us to apply the useful optional stopping theorem.

We assume basic results from martingales in discrete time—specifically we will assume the backwards martingale convergence theorem, which can be mostly proven by modifying the proof of the usual martingale convergence theorem[6].

Now we define martingales in continuous time, as well as some related concepts. Note that conditional expectation is taken over σ-algebras rather than events—one distinction is that a conditional expectation is a random variable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. We have the following definitions:

Definitions.

- A *filtration* is a collection $\{\mathcal{F}_t\}_{t\geq 0}$ of σ-algebras such that every $\mathcal{F}_t \subseteq \mathcal{F}$ and $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s < t$.
- A *stopping time* is a random variable $\tau : \Omega \to [0, \infty]$ such that for all t, $\{\omega \in \Omega : \tau(\omega) \leq t\} \in \mathcal{F}_t$.
- A *submartingale* is a stochastic process $\{X_t\}_{t\in[0,\infty)}$ on $(\Omega, \mathcal{F}, P)$ with:
  - $X_t \in \mathcal{F}_t$. We say that $\{X_t\}$ is *adapted* to $\{\mathcal{F}_t\}$.
  - $X_t \in \mathcal{L}_1$, i.e. $\mathbb{E}|X_t| < \infty$. We say that X_t is *integrable.*
  - $\mathbb{E}[X_t|\mathcal{F}_s] \geq X_s$ for $s \leq t$. It follows that $\mathbb{E}X_t \geq \mathbb{E}X_s$.

  A *supermartingale* is the same thing, but with the third bullet point replaced with $\mathbb{E}[X_t|\mathcal{F}_s] \leq X_s$ for $s \leq t$, and then $\mathbb{E}X_t \leq \mathbb{E}X_s$.
- A *martingale* is a submartingale that is also a supermartingale.
- A *backwards (sub, super) martingale* is the same as above, except the filtrations satisfy $\mathcal{F}_s \supseteq \mathcal{F}_t$ for $s < t$. Equivalently, keep the filtrations the usual way, but index t starting at 0 and going to $-\infty$. (We will

[6]The usual martingale convergence theorem is probably the more "major" theorem, but backwards martingales are important enough to be considered in detail in introductory texts like [11]

only have use for the discrete case, where t can be identified with the negative integers.)

- A stochastic process $\{X_t(\omega)\}_{t\geq 0}$ is *right continuous* if for all t, ω,

$$\lim_{s\downarrow t} X_s(\omega) = X_t(\omega)$$

For some stochastic process $\{X_t\}_{t\geq 0}$ on $(\Omega, \mathcal{F}, P)$, we say the *natural filtration* is defined by $\mathcal{F}_t = \sigma\left(\bigcup_{t'\leq t} \sigma(X_{t'})\right)$; the filtration generated by all random variables that occurred earlier than or at t.

The σ-algebras of a natural filtration refer to information present at time t, because $\sigma(X_t)$ contains all measurable events whose occurrence we can determine by observing X_t, and conditioning on $\sigma(X_t)$ essentially produces a random variable conditioned on all these events, weighted by their probabilities. It will be also be advantageous in proofs to append information aside from just the information contained in the process. Formally this means we generate $\mathcal{F}_t$ not just based on X_t, but based on an auxiliary random variable Y_t, too—or we can let $Y_t = y$ and condition on that event.

Our goal in this section is to prove the optional stopping theorem (for bounded stopping times) from continuous-time martingales for use in proofs in sections 6 and 8, which will allow us to simplify several calculations. Also, we use the discrete time analogue of this theorem in section 9, which is stated the same way.

Our proof is modified from one given in [12], where a less general theorem is proved.

Let $B \geq 0$; define S_B the set of all stopping times τ such that $\tau \leq B$ with probability 1.

Theorem 3 (Bounded Optional Stopping Theorem)**.** *Let $\{X_t\}_{t\geq 0}$ be a right-continuous nonnegative submartingale. Then $\mathcal{A} = \{X_\tau : \tau \in S_B, B \geq 0\}$ is uniformly integrable*[7]*, and*

$$\mathbb{E}[X_\sigma|\mathcal{F}_\tau] \geq X_\tau \text{ and } \mathbb{E}[X_\sigma] \geq \mathbb{E}[X_0] \tag{15}$$

for all $\tau, \sigma \in S_B$ with $\sigma \overset{a.s.}{\geq} \tau$.

Equations (15) show that the submartingale properties that $\mathbb{E}[X_t|\mathcal{F}_s] \geq X_s$ for $s \leq t$ and $\mathbb{E}X_t \geq \mathbb{E}X_s$ generalize to random stopping times given right continuity and a.s. boundedness of the stopping times. Without these kinds of conditions, this need not hold; consider the discrete martingale with respect to its natural filtration, where $X_0 = 0$ and $X_i = X_{i-1} + \xi_i$ for $i \geq 1$ and ξ_i are all iid, and equal to 1 or -1 with equal probability. Then $\{X_i\}$ is a simple random walk. However, if ν is the first time X_ν hits 1, then $\mathbb{E}X_\nu = 1 \neq \mathbb{E}X_0 = 0$, even though ν can be shown to be a stopping time

[7] In martingale theory, when this is satisfied they say that $\{X_t\}_{t\geq 0}$ is class DL. We won't actually use class DL for anything, but it's part of the statement of the theorem, and is important to be able to say a submartingale has a Doob-Meyer decomposition.

and almost surely finite. The problem here is that ν is not almost surely bounded by some B.

Proof. Let $\tau \in S_B^f \subseteq S_B$, where S_B^f is the set of stopping times such that τ takes only finitely many values, such as $0 \leq t_1 < t_2 < ... < t_n \leq B$.

First we show $\mathbb{E}[X_B|\mathcal{F}_\tau] \geq X_\tau$ for $\tau \in S_B^f$ and B our nonrandom bound, in order to establish uniform integrability for $\mathcal{A}^f = \{X_\tau : \tau \in S_B^f\}$. This holds because then $X_\tau = |X_\tau| \leq \mathbb{E}[X_B|\mathcal{F}_\tau]$. The conditional expectation of the integrable random variable X_B is integrable, so that the family $\mathcal{A}^f$ is dominated by an integrable function, which implies uniform integrability. After we show this, we move to general $\tau \in S_B$.

Let $A \in \mathcal{F}_t$. By the definition of conditional expectation, if $\mathbb{E}[X_\tau \mathbb{1}_A] \leq \mathbb{E}[X_B \mathbb{1}_A]$ for any such A then we do have $\mathbb{E}[X_B|\mathcal{F}_\tau] \geq X_\tau$, so we show this. Since $X_\tau = \sum_{k=1}^n X_\tau \mathbb{1}\{\tau = t_k\}$, we get the first equality, and then simplify:

$$\mathbb{E}[X_\tau \mathbb{1}_A] = \sum_{k=1}^n \mathbb{E}[X_\tau \mathbb{1}_A \mathbb{1}\{\tau = t_k\}] = \sum_{k=1}^n \mathbb{E}[X_{t_k} \mathbb{1}_A \mathbb{1}\{\tau = t_k\}]$$

Then, using that $\{X_t\}$ is a submartingale, and applying once again the definition of conditional expectation (because $\{\tau = t_k\}$ is an event in $\mathcal{F}_{t_k}$):

$$\leq \sum_{k=1}^n \mathbb{E}\Big[\mathbb{E}[X_B \mathbb{1}_A \mathbb{1}\{\tau = t_k\}|\mathcal{F}_{t_k}]\Big] = \sum_{k=1}^n \mathbb{E}[X_B \mathbb{1}_A \mathbb{1}\{\tau = t_k\}]$$

which is $\mathbb{E}[X_B \mathbb{1}_A]$, as needed.

Now let $\tau \in S_B$. We define a sequence $\{\tau_n\}$ of stopping times by $\tau_1 = B$ and for $n > 1$, $\tau_n = \max\{2^{-n}\lceil 2^n \tau \rceil, \tau_{n-1}, B\}$. These are strictly larger than τ and less than B and nonincreasing, and for each n only take finitely many values, because the ceiling function only takes finitely many values until $\tau_n = B$. So $\tau_n \in \mathcal{S}^f$, and we have $\tau_n \downarrow \tau$. By right continuity, this implies that $X_{\tau_n} \to X_\tau$ (for any fixed ω).

Fix $\epsilon > 0$; writing out the definition of uniform integrability for $\mathcal{A}^f$, we have $K \geq 0$ such that for every X_{τ_n}

$$\mathbb{E}\Big[|X_{\tau_n}| \cdot \mathbb{1}\{|X| \geq K\}\Big] \leq \epsilon$$

Applying DCT and taking $n \to \infty$ we see that appending X_τ to $\mathcal{A}^f$ doesn't break the uniform integrability. We can do this for all τ and see that the set $\mathcal{A} = \{X_\tau : \tau \in S_B\}$ is uniformly integrable.

Now we show the martingale properties (15) hold. Let's redefine $\tau_n = \max\{2^{-n}\lceil 2^n \tau \rceil, \tau_{n-1}, \sigma\}$, setting $\tau_1 = \sigma$. Since τ_n is nonincreasing, $\{X_{\tau_n}\}$ and its corresponding filtration define a discrete backwards submartingale. The statement of the backwards submartingale convergence theorem is that a discrete backwards submartingale has a limit both in a.s. (which we know is X_τ) and in $\mathcal{L}^1$, and that this is less than or equal to $\lim_n \mathbb{E}[V|\mathcal{F}_{\tau_n}]$, where

V is the first term of the backwards submartingale. X_σ is the first term, so

$$X_{\tau_n} \to X_\tau \leq \lim_n \mathbb{E}[X_\sigma | \mathcal{F}_{\tau_n}]$$

in both a.s. and $\mathcal{L}^1$.

To use these formulas, we take conditional expectations:

$$X_\tau = \mathbb{E}[X_\tau|\mathcal{F}_\tau] \leq \mathbb{E}\Big[\lim_n \mathbb{E}[X_\sigma|\mathcal{F}_{\tau_n}]\Big|\mathcal{F}_\tau\Big] = \lim_n \mathbb{E}\Big[\mathbb{E}[X_\sigma|\mathcal{F}_{\tau_n}]\Big|\mathcal{F}_\tau\Big]$$
$$= \lim_n \mathbb{E}[X_\sigma|\mathcal{F}_\tau] = \mathbb{E}[X_\sigma|\mathcal{F}_\tau]$$

where we can commute the limits because of $\mathcal{L}^1$ convergence and we use the law of total expectation at the end. This proves $\mathbb{E}[X_\sigma|\mathcal{F}_\tau] \geq X_\tau$.

To show $\mathbb{E}[X_0] \leq \mathbb{E}[X_\sigma]$, just set $\tau = 0$ so that $\mathbb{E}[X_\sigma|\mathcal{F}_0] \geq X_0$, and take expectations of both sides:

$$\mathbb{E}\Big[\mathbb{E}[X_\sigma|\mathcal{F}_0]\Big] = \mathbb{E}X_\sigma \geq \mathbb{E}X_0.$$

□

Some final remarks. This is not the most general form of the optional stopping theorem, either in discrete and continuous time; even if we revoke boundedness of the stopping times, there are other conditions we could impose so that equations (15) still hold. Lastly, we note that nonnegativity was not required for equations (15) to hold.[8]

6. The Benjamini-Hochberg procedure

With the needed results from martingale theory in place, we are all ready to set up to state and prove the Bejamini-Hochberg procedure for FDR control [13], abbreviated the BH procedure, or BH(q).

Benjamini and Hochberg's landmark 1995 paper, currently with around 45000 citations on Google Scholar, first defined the FDR and the BH method to control it. FDR has since become the mainstream criterion for multiple testing due to advances in data collection technology, and the BH procedure in particular has become a standard topic in applied statistics because of its simplicity and practicality. In this section we define the procedure and prove it controls FDR, using essentially the martingale proof due to Storey et al. [14], which greatly simplified the original authors'. The martingale idea is to exploit the optional stopping theorem, and this idea crops up again and again in the extensions and relatives of the BH procedure we give in sections 8 and 9.

The setting for BH is N unordered null hypotheses $H_1, H_2, ...H_N$, each of which is associated with independent p-values $p_1, p_2, ..., p_N$. Some unknown number $N_0 = \pi_0 N$ of these hypotheses are truly null hypotheses, and their

[8]Actually, if $\{X_t\}$ is a martingale, you don't even need nonnegativity to show it's of class DL (see previous footnote.) At the step in the proof where we show X_τ is dominated, we can instead show it's equal to a conditional expectation over a sub σ-algebra of $\mathcal{F}$, a set which is always uniformly integrable.

corresponding p-values are therefore distributed as $p_i^0 \overset{iid}{\sim} \text{Unif}[0,1]$, where the 0 superscript denotes a null p-value and $\pi_0 \in [0,1]$ is called the null proportion. The approach of BH is to specify an FDR controlling threshold t such that we reject all p-values p_i satisfying $p_i \leq t$.

Specifically let our total number of rejections be $R(t) = \#\{p_i \leq t\}$. Let $V(t) = \#\{p_i \leq t : p_i \text{ is null}\}$ be the number of false rejections. The threshold t is chosen to bound $\text{FDR}(t) = \mathbb{E}\Big[V(t)/\max(R(t),1)\Big]$, which is our analogue of type I error, by some $q \in [0,1]$.

For $p_1, ..., p_N$, define

$$\widehat{\text{FDR}}(t) = \frac{Nt}{\max(R(t),1)}$$

Then BH(q) prescribes

$$\hat{t}_q^{BH} := \sup\{0 \leq t \leq 1 : \widehat{\text{FDR}}(t) \leq q\} \tag{16}$$

In particular we have

$$\widehat{\text{FDR}}(\hat{t}_q^{BH}) \leq q. \tag{17}$$

Theorem 4 (BH(q))**.** *For the N hypotheses $H_1, ..., H_N$, with independent p-values $p_1, ..., p_N$, we reorder the p-values to form the order statistics $p_{(1)}, ..., p_{(N)}$, and order the N hypotheses the same way, labelling $H_{(1)}, ..., H(k)$.*

Fix $q \in [0,1]$, and reject $H_{(k)}$ for $p_{(k)} \leq t_q^{BH}$. This rule controls FDR at level q.

Before we give the proof let's make some remarks.

First, the usual formulation of BH(q) is more simply stated; it just says that for ordered p-values $p_{(1)} \leq ... \leq p_{(N)}$, we define

$$\hat{k}_q^{BH} = \max\left\{k : p_{(k)} \leq \frac{k}{N}q\right\} \tag{18}$$

and we reject all $H_{(k)}$ with $k \leq \hat{k}_q^{BH}$. The random integer $\hat{k}_q^{BH}$ is called a stopping rule.

The two formulations are equivalent because $R(p_{(k)}) = k$, so that $\widehat{\text{FDR}}(p_{(k)}) = Np_{(k)}/k$, and then

$$p_{(\hat{k}_q^{BH})} = \max\left\{p_{(k)} : p_{(k)} \leq \frac{k}{N}q\right\} = \max\left\{p_{(k)} : \widehat{\text{FDR}}(p_{(k)}) \leq q\right\}$$

Comparing the form of $p_{(\hat{k}_q^{BH})}$ to $\hat{t}_q^{BH}$, we conclude $p_{(k)} \leq p_{(\hat{k}^{BH})}$ (i.e. is rejected) if and only if $p_{(k)} \leq \hat{t}_q^{BH}$, since $p_{(\hat{k}^{BH})} \leq \hat{t}_q^{BH} < p_{(\hat{k}^{BH}+1)}$.

Secondly, why have I chosen the name $\widehat{\text{FDR}}$ (where the hat suggests an estimate)? Actually, it really is kind of a heuristic estimate of the true FDR as follows:

$$\mathbb{E}\left[\frac{V(t)}{\max(R(t),1)}\right] \approx \frac{\mathbb{E}V(t)}{\mathbb{E}\max(R(t),1)} \approx \frac{\mathbb{E}V(t)}{\max(R(t),1)}$$

Since the denominator is an observed random variable, we can just estimate its expectation by its observed value. But what is the numerator? The null p values are uniform, so they lie within the rejection region $[0, t]$ with probability t; if there were just one null p-value p_i, then $\mathbb{E}V(t) = P(p_i \leq t) = t$. Since there are $\pi_0 N$ of them, $\mathbb{E}V(t) = \pi_0 N t$.

Still π_0 is unknown. We may "estimate" $\pi_0 = 1$[9] , and find that

$$\frac{\pi_0 N t}{\max(R(t), 1)} \approx \frac{N t}{\max(R(t), 1)} = \widehat{\text{FDR}}(t)$$

So $\widehat{\text{FDR}}(t)$ is intuitively a kind of upward biased estimate of the true FDR, so that we would expect the true FDR to be "probably" less than $\widehat{\text{FDR}}$, and bounding the latter should bound the former. This is precisely what BH(q) does.

The idea of bounding an estimate for the FDR is also present in section 9; one can choose different kinds of estimates to derive different kinds of methods, and that is just what we do there.

Proof. We ultimately want to show that at the threshold $t = t_q^{BH}$, we have $\text{FDR}(t_q^{BH}) \leq q$, and we say FDR is controlled at level q.

Let $M(t) = V(t)/t$, and consider the stochastic process $\{M(t)\}_{t=1}^0$. The bounds mean we think of $M(1)$ as the starting point, with $t \to 0$. Define the σ-algebra $\mathcal{F}_t = \sigma(V(s), R(s) : 1 \geq s \geq t)$. Then $M(t)$ is adapted to the filtration $\{\mathcal{F}_t\}_{t=1}^0$, because if we know $V(t)$ we also know $M(t)$. (Alternately, $V(t)$ is $\mathcal{F}_t$-measurable, so $M(t)$ is as well.)

We show that $M(t)$ is a martingale in backwards time[10] (a calculation omitted in Storey et al [14].). $M(t)$ is integrable because $\mathbb{E}|M(t)| = \pi_0 N$. It remains to show that $\mathbb{E}[M(t)|\mathcal{F}_s] = M(s)$[11] for $1 \geq s \geq t$.

First, we compute $\mathbb{E}[M(t)|M(s)] := \mathbb{E}\big[M(t)|\sigma(M(s))\big]$:

$$\mathbb{E}[M(t)|M(s)] = \mathbb{E}[M(t)|V(s), s] = \frac{1}{t}\mathbb{E}[V(t)|V(s), s]$$

where we explicitly note our knowledge of s. Let p_i^0 denote a null p-value. Conditional on $V(s)$ and s, we have

$$V(t) = \#\{p_i^0 : p_i^0 \leq t \leq s\} = \sum_{i=1}^{V(s)} \xi_i \tag{19}$$

[9]There exist corrections [14] which properly estimate π_0, rather than just setting it to 1. Such methods yield greater power, but are not my focus.

[10]Not to be confused with backwards martingale; the first term of a martingale's filtration is the smallest σ-algebra of that filtration, and a backwards martingale's is the largest. $\mathcal{F}_1$ is the smallest here.

[11]This is why it makes sense to go backward in time. Conditioning on $\mathcal{F}_s$, which is to say, $V(s)$, for $1 \geq s \geq t$ excludes all the null p-values above s, which restricts the number of p-values we have left to classify by time t. On the other hand, knowing $V(t)$ doesn't give us useful information about $V(s)$.

where ξ_i is a random variable that's 1 if $p_i^0 \le t$ and 0 otherwise.

Unconditioned, $p_i^0 \overset{iid}{\sim} \text{Unif}[0,1]$. Then conditioning on s, a routine calculation gives that each $p_i^0 | \{p_i^0 \le s\} \overset{iid}{\sim} \text{Unif}[0,s]$. Additionally conditioning on $V(s)$, which is to say the state of null p-values besides the ith, gives no additional information because the p_i are independent.

Now just compute:[12]

$$\frac{1}{t}\mathbb{E}[V(t)|V(s),s] = \frac{1}{t}\sum_{i=1}^{V(s)} \mathbb{E}\xi_i = \frac{1}{t}\sum_{i=1}^{V(s)} P(p_i^0 \le t) = \frac{1}{t}V(s)P(p_i^0 \le t) \quad (20)$$

$$= \frac{1}{t}V(s)\frac{t}{s} = \frac{V(s)}{s} = M(s)$$

where t/s is the CDF of $\text{Unif}[0,s]$.

So we have $\mathbb{E}[M(t)|M(s)] = M(s)$. In addition, $\mathbb{E}[M(t)|M(s'), s \le s'] = M(s)$, because the knowledge of every $V(s')$ means we know which $p_i^0 > t$ as long as $s < p_i^0 \le s'$ (because we know the s' such that $V(s')$ has jumps), and therefore we know $V(t)$ does not count them. However, whether or not $V(s)$ counts the $p_i^0 \le s$, of which there are $V(s)$, is still random; this is the content of equation (19). Lastly, knowing $R(s)$ when we know $V(s)$ means we know which were the non-null p-values as well, but that's irrelevant to $M(t)$. So we have shown $\mathbb{E}[M(t)|\mathcal{F}_s] = M(s)$.

t_q^{BH} is clearly bounded, and it is a stopping time because if we know $R(s)$ for all $s \ge t$, then we know $\widehat{\text{FDR}}(s)$ (in particular we know whether or not it's less than q). This is enough to know if $t_q^{BH} \le t$.

However, $M(t)$ is not right continuous (taking "going right" to mean from 1 to 0). No matter; because the p_i^0 come from a continuous distribution, we can redefine $V(t) = \#\{p_i^0 : p_i^0 < t\}$ using strict inequality[13] and define $M(t)$ using this, and they will be equal almost surely, so that we have a right continuous version of $M(t)$.

Applying optional stopping for martingales,

$$\mathbb{E}M(t_q^{BH}) = \mathbb{E}M(1)$$

$$\frac{1}{t_q^{BH}}\mathbb{E}V(t_q^{BH}) = \mathbb{E}V(1) = N_0$$

where recall N_0 is the true number of nulls. From the definition of $\widehat{\text{FDR}}$, we also have

$$\max(R(t_q^{BH}), 1) = \frac{Nt_q^{BH}}{\widehat{\text{FDR}}(t_q^{BH})} \ge \frac{Nt_q^{BH}}{q}$$

[12] Saying that we condition on a random variable is shorthard for saying we condition on the σ-algebra it generates. However, one can show that it corresponds correctly with our intuition, so we can use it to do calculations without fear, as I'm doing here.

[13] If the original problem involved discrete distributions, the p-values may not have continuous distributions. Indeed, we are implicitly assuming they are all continuous.

where the inequality is implied by (17).

Putting these two together, and noting $N_0 \leq N$ we finally have

$$\text{FDR}(t_q^{BH}) = \mathbb{E}\left[\frac{V(t_q^{BH})}{\max(R(t_q^{BH}), 1)}\right] \leq \frac{q}{N}\mathbb{E}\left[\frac{V(t_q^{BH})}{t_q^{BH}}\right] = q\frac{N_0}{N} \leq q \tag{21}$$

□

7. Sequential Hypothesis Testing

So much for unordered hypothesis testing. We have now defined FDR and proved a procedure that controls it, called the BH procedure. However, we will actually have no use for BH directly. We include it because it is archetypal; its proof is the simplest example of an optional stopping based proof we have, not to mention it would be remiss to omit such a landmark result.

However, our real interest is in *sequential* hypothesis testing: we impose the condition that we can only reject hypotheses in a certain order, and the reordering of the p-values as done in BH are impossible in this case. Regardless, one can think of the forthcoming procedures as variations on the BH theme.

To motivate the problem, note that in section 3, the null hypothesis was $H_0 : \text{supp}(\boldsymbol{\beta}^*) \subseteq A_k$, where A_k was the active set just before the kth knot. If there are N knots (in fact, recall $N = p$ for orthogonal $\boldsymbol{X}$), then we may index the null hypotheses:

$$\begin{aligned} H_1 &: \text{supp}(\boldsymbol{\beta}^*) \subseteq A_1 \\ H_2 &: \text{supp}(\boldsymbol{\beta}^*) \subseteq A_2 \\ &\vdots \\ H_N &: \text{supp}(\boldsymbol{\beta}^*) \subseteq A_N \end{aligned} \tag{22}$$

and in the setting of the lasso with orthogonal $\boldsymbol{X}$, we have $A_1 \subseteq A_2 ... \subseteq A_N$ because variables only enter and never leave.

Rejecting a hypothesis means that we believe the model A_k selected at that stage is wrong, so we increase λ, reach another knot, and therefore add another variable and consider A_{k+1}. Since the falsehood of A_k implies the falsehood of $A_{k'}$ for all $k' < k$, the rejection of A_k implies rejection of all $A_{k'}$. So when we reject hypotheses, we can only reject sequentially; that is, we cannot reject A_1 and A_3 without rejecting A_2 as well. Subject to this condition, we would like to derive FDR controlling procedures.

Let's formalize the main idea. We have $H_1, ..., H_N$ hypotheses in a specific order and associated p-values, where rejection of hypotheses is not simultaneous but rather goes from the first to the last. Any rejection rule is constrained by the following requirement: if we reject, we may only reject a set of hypotheses of the form $\{H_1, ..., H_k\}$ for some $k \leq N$.

This restriction of rejecting hypotheses in order is called sequential hypothesis testing, because of the enforced sequence $\{H_k\}_{k=1}^N$. Each hypothesis is associated with a p-value, whose generation was the topic of section 3. We now consider them as generated from some distribution, and focus on the problem of combining them for testing in this more restricted setting, which is the natural one for the null hypotheses of equations (22).

Each FDR-controlling method we consider will define a stopping rule $\hat{k}$, which is a function of the p values, and tell us to reject all $H_1, ..., H_{\hat{k}}$. We had one for BH(q) as well, but now let's make it a formal definition.

Definition. A *stopping rule* for a sequential hypothesis testing problem $H_1, ..., H_N$ with associated p-values $p_1, ..., p_N$ is any function $\hat{k}$ of the N p-values which takes values in $\{0, 1, ..., N\}$, and rejects all hypotheses H_k with $k \leq \hat{k}$.

So in sequential testing, we only reject an initial block of hypotheses $H_1, ...H_{\hat{k}}$, up to the index returned by the stopping rule.

To illustrate the concept, we define probably the simplest stopping rule that works. It simply rejects until the first time a p-value exceeds q.

Proposition. *Suppose either that for some unknown $1 \leq k_0 \leq N$, every $H_{k_0}, ..., H_N$ is null, or that none of our hypotheses are null. Then define the stopping rule*

$$\hat{k} = \min\{k : p_k > q\} - 1$$

for some $q \in [0, 1]$. This controls FDR at level q. In fact, it controls $P(V \geq 1)$ at level q.

Proof. Recall V is false rejections and R is all rejections. The quantity $P(V(q) \geq 1)$ was mentioned in section 3 as an alternative to the FDR as a type-I-esque criterion. In fact $P(V(q) \geq 1) \geq \text{FDR}$, because $R(q) \geq V(q)$, so

$$\mathbb{1}\{V(q) \geq 1\} \geq \frac{V(q)}{\max(R(q), 1)}$$

and taking expectations on both sides shows that $P(V(q) \geq 1) \geq \text{FDR}(q)$. The given stopping rule controls $P(V(q) \geq 1)$ because

$$P(V(q) \geq 1) \leq P(p_{k_0} \leq q) = q$$

where k_0 is the first null hypothesis, and null p-values are uniform.

□

The quantity $P(V \geq 1)$ is known as the FWER, the family-wise error rate. Before Benjamini and Hochberg proposed the FDR as an alternative criterion, researchers used the FWER, and had a hard time because (as discussed in section 4) such a criterion is too stringent and sucks away power as the number of hypotheses increases. Since this simple stopping rule controls

FWER, it can't be any good. What's more, it requires all the non-nulls to precede the nulls, which may not occur in practice. How can we do better? Such methods are the focus of the next two sections 8, 9.

8. ForwardStop

It would be nice if we could use BH(q) in the sequential hypothesis setting of section 6. In fact, the BH procedure does perform sequential rejections—but only after the p-values are placed in ascending order, and the hypotheses reordered in the same fashion, producing $H_{(1)}, ..., H_{(N)}$. In the remarks after Theorem (4), we phrased BH(q) using a stopping rule $k \leq \hat{k}_q^{BH}$, which, as in sequential hypothesis testing, rejects an initial block $H_{(1)}, .., H_{(k)}$ with $k \leq \hat{k}_q^{BH}$.

If we want to reject an initial block in the *original* order of $H_1, ..., H_N$, we clearly can't reorder by ascending p-values and reject an initial block in the new order, as BH(q) says to do.

G'Sell et al. [15] were the first to recognize the sequential hypothesis testing setting, and to derive an FDR controlling procedure for it, called ForwardStop. Their idea was to, rather than reorder the p-values, first transform the p-values using the following theorem due to Alfred Renyi. It says that, through a linear transformation, one can transform iid exponential samples into exponential order statistics.

Theorem 5 (Renyi representation)**.** *Let* Exp(α) *denote the exponential distribution whose mean is* $1/\alpha$*. Let* $X_1, ..., X_n \overset{iid}{\sim}$ Exp(1)*. Then the random vector*

$$\left(\frac{X_1}{n}, \frac{X_1}{n} + \frac{X_2}{n-1}, ..., \sum_{i=1}^{n} \frac{X_i}{n-i+1} \right)$$

is jointly distributed like

$$(E_{1,n}, E_{2,n}, ..., E_{n,n})$$

where $E_{j,n}$ *denotes the* j*th order statistic of* n Exp(1) *variables.*

Proof. Note the memoryless property of exponentials, which says that if $Y \sim \text{Exp}(\alpha)$ and T is any nonnegative random variable (with a pdf p) such that T is independent of Y,

$$P(Y \geq y + T \mid Y \geq T) = \int_{\mathbb{R}} P(Y \geq y + t \mid Y \geq t, T = t) p(t) dt$$

$$\int_{\mathbb{R}} P(Y \geq y \mid T = t) p(t) dt = \int_{\mathbb{R}} P(Y \geq y) p(t) dt = P(Y \geq y)$$

Also note that $E_{1,n} \sim \text{Exp}(n)$, which you get by taking the exponential CDF to the nth power.

Let $Y_{(1)}, ..., Y_{(n)}$ be standard exponential order statistics. Consider the spacings $S_i = Y_{(i+1)} - Y_{(i)}$ for $1 \leq i \leq n-1$. $Y_{(i)}$ is characterized by having

above it $n-i$ order statistics $Y_{(i+1)}, ..., Y_{(n)}$, so $Y_{(i+1)}$ can be thought of as distributed as $E_{1,n-i}$ independently from $Y_{(i)}$, conditioned on $E_{1,n-i} \geq Y_{(i)}$.

Then

$$\begin{aligned} P(S_i > s) &= P(E_{1,n-i} - Y_{(i)} > s \mid E_{1,n-i} \geq Y_{(i)}) \\ &= P(E_{1,n-i} > s + Y_{(i)} \mid E_{1,n-i} \geq Y_{(i)}) \\ &= P(E_{1,n-i} > s) \end{aligned}$$

where we have used that $E_{1,n-i}$ is exponential. So $S_i \sim \text{Exp}\,(n-i)$.

Now $(n-i)S_i = (n-i)(Y_{(i+1)} - Y_{(i)}) := X_i' \sim \text{Exp}(1)$. Then $Y_{(i+1)} - Y_{(i)} = X_i'/(n-i)$. Using the algebraic identity

$$\sum_{i=0}^{j-1} Y_{(i+1)} - Y_{(i)} = Y_{(j)} \sim E_{j,n}$$

where we may define $Y_{(0)} = 0$ and $X_0' = nY_{(1)}$, substituting in $X_i'/(n-i)$ yields the result.

□

If we can convert exponential random variables to exponential order statistics, we can convert uniform p-values p_k to uniform order statistics p_k' by converting to exponential (through inverse transform sampling), applying Renyi representation, and then converting back (through the probability integral transform). The uniform order statistics act like ordered p-values, which allows us to apply BH(q). This idea translates more or less immediately to a proof if all the non-null hypotheses are assumed to precede the nulls, which may not be the case in practice. Regardless, the main idea is still present in the proof of FDR control for the ForwardStop procedure proposed by G'Sell et al., named so because it scans the p-values in a forward direction.

Theorem 6 (ForwardStop)**.** *Suppose we have N ordered hypotheses with associated independent p-values, $H_1, ..., H_N$ and $p_1, ..., p_N$ and a subset $\mathcal{H}_0 \subseteq \{1, ..., N\}$ which indexes the null hypotheses (so that $p_i \overset{iid}{\sim} \text{Unif}[0,1]$ for all $i \in \mathcal{H}_0$). Let $0 < q < 1$.*

Define the stopping rule

$$\hat{k}_q^F = \max \left\{ k \in \{1, ..., N\} : \frac{1}{k} \sum_{i=1}^{k} -\log(1-p_i) \leq q \right\} \tag{23}$$

Then this controls FDR at level q.

Proof. Define $Y_i = -\log(1-p_i)$, which is standard exponential random variables if p_i is null. Then define

$$Z_k = \sum_{i=1}^{l} \frac{Y_i}{\nu(i)}$$

where $\nu(i) = \#\{j \in \{i, ..., N\} : j \in \mathcal{H}_0\}$, the number of nulls from i to N.

If all the hypotheses had been truly null, then $\nu(i) = N - i + 1$, and Z_k would be an exponential order statistic by Renyi representation. That is not the necessarily the case under our assumptions, but we can still break the sum into two terms and apply Renyi representation to one of them.

Let $N_0 = \#\{\mathcal{H}_0\}$, the number of null hypotheses. Starting from $k = 1$ and going to $k = N$, take the p-value p_k and give it a label: $\mathcal{A}^{(i)}$ if it's the ith non-null from $\{p_1, ..., p_k\}$ you've seen, and $\mathcal{N}^{(i)}$ if it's the ith null. There are N_0 nulls and $N - N_0$ non-nulls; we can re-express Z_k as

$$Z_k = \sum_{i=1}^{k-(N_0-\nu(k+1))} \frac{Y_{\mathcal{A}^{(i)}}}{\nu(\mathcal{A}^{(i)})} + \sum_{i=1}^{N_0-\nu(k+1)} \frac{Y_{\mathcal{N}^{(i)}}}{N_0 - i + 1}$$

where $N_0 - \nu(k+1)$ is the number of nulls from 1 to k (defining $\nu(N+1) = 0$) and where we've simplified $\nu(\mathcal{N}^{(i)}) = N_0 - i + 1$ when p_i is null.

Define $Y_i' = Y_{\mathcal{N}^{(i)}}$, which is standard exponential and independent from the other Y's. By Renyi representation, the second summation is distributed as some $E_{N_0-\nu(k+1),N_0}$ variable.

Again consider the ideal situation with all the hypotheses truly null and Z_k a exponential order statistic. Then defining $p_k' = 1 - e^{-Z_k}$ would turn p_k' into a uniform order statistic (and $1 - p_k' = e^{-Z_k}$ are the uniform order statistics in the opposite order). We're not in the ideal situation, but regardless, we define p_k' as such anyway. Then we have $1 - p_k'$ equal to e^{-Z_k}, and

$$1 - p_k' \stackrel{d}{=} \exp\left\{ - \sum_{i=1}^{k-(N_0-\nu(k+1))} \frac{Y_{\mathcal{A}^{(i)}}}{\nu(\mathcal{A}^{(i)})} \right\} \cdot U_{\nu(k+1),N_0}$$

where $U_{\nu(k+1),N_0}$ is the $\nu(k+1)$th standard uniform order statistic out of N_0. Defining $r(k) = \exp\left\{ - \sum_{i=1}^{k-(N_0-\nu(k+1))} \frac{Y_{\mathcal{A}^{(i)}}}{\nu(\mathcal{A}^{(i)})} \right\}$ and noting that subtracting uniform order statistics from 1 gives uniform order statistics in the opposite order, we re-express

$$1 - p_k' \stackrel{d}{=} r(k)(1 - U_{N_0-\nu(k+1),N_0})$$

So the p_k' may not be uniform order statistics, but at least for the null p's, we can read off that they are closer to 1 than uniform order statistics. Why? Letting $p_k'^0 := p'_{\mathcal{N}^{(k)}}$, we have

$$1 - p_k'^0 \stackrel{d}{=} r(\mathcal{N}^{(k)})(1 - U_{k,N_0})$$

because $N_0 - \nu(k+1)$ is the number of nulls from 1 to k, and the number of nulls $N_0 - \nu(\mathcal{N}^{(k)} + 1)$ from 1 to $\mathcal{N}^{(k)}$ is k. Now since $0 \leq r(k) \leq 1$, we imagine shrinking the difference between 1 and U_{k,N_0}. In other words, the $p_k'^0$ are stochastically larger than uniform.

With this knowledge we attempt to apply BH(q) to the p'_k, as in equation (16), defining $R(t) = \#\{p'_i \leq t\}$ and $V(t) = \#\{p'_i \leq t : i \in S_N\}$. Then FDR and $\widehat{\text{FDR}}$ are also defined analogously as in section 6.

Now set

$$t_q^F := \sup\{0 \leq t \leq 1 : \widehat{\text{FDR}}(t) \leq q\}$$

using these definitions for R and V. We wish to reject all H_k with $p'_k \leq t_q^F$.

Analogously to the proof of BH(q), for the same filtration, we use a martingale idea, but this time we actually show that $M(t) = V(t)/t$ is a submartingale.

Conditional on $V(s)$, for $s \geq t$, express $V(t)$ as in equation (19):

$$V(t) = \#\{p_i'^0 : p'_i \leq t \leq s\} = \sum_{i=1}^{V(s)} \xi_i$$

with ξ_i a random variable that's 1 if $p_i'^0 \leq t$ and 0 otherwise. We know the $p_i'^0$ are stochastically larger than uniform order statistics, but $V(t)$ has the same value even if we permute the order of the $p_i'^0$ randomly, and if we do so, the shuffled $p_i'^0$ will each be stochastically larger than the standard uniform. This is because one can imagine they were they initially sampled iid from a uniform, made into order statistics, and then increased by some function of $r(\mathcal{N}^{(k)})$ that depends only on the nonnulls. Then shuffling them again will break any effect ordering has on their distribution, but the effect of r remains.

Suppose we have done this. Now compute $\mathbb{E}[M(t)|M(s)]$ as in (20):

$$\frac{1}{t}\mathbb{E}[V(t)|V(s), s] = \frac{1}{t}\sum_{i=1}^{V(s)} \mathbb{E}\xi_i = \frac{1}{t}\sum_{i=1}^{V(s)} P(p_i'^0 \leq t) = \frac{1}{t}V(s)P(p_i'^0 \leq t)$$

and then using that each $p_i'^0$ is stochastically larger than uniform, which means that their CDF is smaller,

$$\frac{1}{t}V(s)P(p_i'^0 \leq t) \leq \frac{1}{t}V(s)\frac{t}{s} = \frac{V(s)}{s} = M(s)$$

so that $\mathbb{E}[M(t)|M(s)] \leq M(s)$. Then, following exactly similar reasoning to the proof of BH(q) in section 6, conditioning on $\mathcal{F}_s$ we see that $M(t)$ is a submartingale. Because t_q^F is bounded and can be verified, as in section 6, to be a stopping time, (and because, assuming continuity of the transformed p-values p', we can redefine $V(t)$ with strict inequality to get right continuity, as in section 6), we can essentially rewrite equation (21) all over, so that rejecting all H_k such that $p'_k \leq t_q^F$ will control FDR.

However the theorem presented ForwardStop using an integer stopping rule, not a stopping time. Because we are applying BH(q) to the p' values, we can use equation (18) to see that rejecting based on $p'_k \leq t_q^F$ is equivalent

to rejecting based on all $k \leq \hat{k}_q$, where

$$\hat{k}_q = \max\left\{k : p'_k \leq \frac{k}{N}q\right\} = \max\left\{k : 1 - \exp\left[-\sum_{i=1}^{k}\frac{Y_i}{\nu(i)}\right] \leq \frac{k}{N}q\right\} \quad (24)$$

However, not only is $\hat{k}_q$ not equivalent to the rule proposed by the theorem, it is not even computable because it depends on knowledge of $\nu(k)$. To rectify this, we can append to the original untransformed list $p_1, ..., p_N$ additional null p-values (i.e. standard uniform) $p_{N+1}, ..., p_{N'}$. The procedure outlined so far still applies for any $N' \geq N$; that is, letting $t^F_{q,N'}$ be the resulting stopping time, $\text{FDR}(t^F_{q,N'}) \leq q$. Therefore, by DCT,

$$\begin{aligned}\lim_{N'\to\infty} \text{FDR}(t^F_{q,N'}) &= \lim_{N'\to\infty} \mathbb{E}\frac{V(t^F_{q,N'})}{\max(R(t^F_{q,N'}), 1)} \\ &= \mathbb{E}\left[\lim_{N'\to\infty} \frac{V(t^F_{q,N'})}{\max(R(t^F_{q,N'}), 1)}\right] \leq q\end{aligned}$$

so the rule obtained by taking $N' \to \infty$ controls FDR if indeed we obtain some rule. We can compute this rule through $\hat{k}_q$. Letting $\nu_{N'}(i)$ denote ν in this setting, we note it behaves like N' for large N'. Likewise $N + N' \approx N'$ for large N'. So taking limits in (24),

$$\begin{aligned}\lim_{N'\to\infty} \hat{k}_q &= \lim_{N'\to\infty} (N + N') \max\left\{k : p'_k \leq \frac{k}{N}q\right\} \\ &= \max\left\{k : N\left(1 - \exp\left[-\sum_{i=1}^{k}\frac{Y_i}{N}\right]\right) \leq kq\right\} \\ &= \max\left\{k : \sum_{i=1}^{k} Y_i \leq kq\right\} \\ &= \hat{k}^F_q\end{aligned}$$

a rule to which corresponds the random variable

$$\lim_{N'\to\infty} V(t^F_{q,N'})/\max(R(t^F_{q,N'}), 1)$$

so that we were justified in applying DCT, and which matches the rule specified in the theorem.

□

In this section we presented a method called ForwardStop which uses p-values to control FDR in a sequential hypothesis testing setting. At this point, the reader could take the statistics from section 3 on a model selection problem, turn them into p-values, and apply them with ForwardStop. But ForwardStop is not the only way to control FDR for sequential hypotheses, as we see in the next section.

9. Accumulation test procedures, SeqStep, and HingeExp

Consider the ForwardStop stopping rule $\hat{k}_q^F$. Specifically, forget about the Renyi and order statistic justification and consider the form ultimately arrived at in equation (23). Why should we believe, intuitively, that a function like $-\log(1-p_i)$ would work—as opposed to its cube, say? In this section, we show that ForwardStop is part of a class of FDR controlling methods called accumulation tests (though what we actually prove is that accumulation tests control something similar, to but slightly larger than, the FDR, and ForwardStop is special because it controls FDR exactly).

Let FDP $= V/\max(R,1)$, so that FDR $= \mathbb{E}(\text{FDP})$. The idea of ForwardStop is to reject as much as possible—hence the max—while controlling a certain quantity, namely $\frac{1}{k}\sum \log\left(\frac{1}{1-p_i}\right)$. It's reasonable to control this quantity because its expectation actually acts an upper bound of FDP (thereby kind of an overestimate of the FDR), so if it's bounded above, then so might FDP/FDR.

It's an upper bound of FDP because null p-values p_i have the pdf $f(x) = 1$ on $[0,1]$, so under the null

$$\mathbb{E}\log\left(\frac{1}{1-p_i}\right) = \int_0^1 \log\left(\frac{1}{1-p_i}\right) = 1$$

and we have

$$\frac{1}{k}\cdot\mathbb{E}\left[\sum_{i=1}^{k}\log\left(\frac{1}{1-p_i}\right)\right] \geq \frac{1}{k}\cdot\mathbb{E}\left[\sum_{\text{null } i\leq k}\log\left(\frac{1}{1-p_i}\right)\right] = \frac{V(k)}{k}$$

where k represents the value taken by our stopping rule, and stopping rules say to reject all H_i with $i \leq k$. So $R(k) = k$.

This intuition isn't unique to ForwardStop; indeed, any function $h : [0,1] \to [0,\infty)$ such that $p \sim \text{Unif}[0,1]$, and

$$\mathbb{E}[h(p)] = \int_0^1 h(p)dp = 1$$

will have a corresponding rule

$$\hat{k}_q^h = \max\left\{k : \frac{1}{k}\sum_{i=1}^{k} h(p_i) \leq q\right\}$$

where $(1/k)\sum_{i=1}^{k} h(p_i)$ is an overestimate of the FDP, and $\hat{k}_q^h = 0$ if it's otherwise undefined.

Any h satisfying these properties is called an accumulation function; ForwardStop just sets $h(p) = \log(\frac{1}{1-x})$. We formalize the definition, however simple:

Definition. Let $h : [0,1] \to [0,\infty)$ be such that $\int_0^1 h(p)dp = 1$. Then we call h an *accumulation function.*

This generalization was first noticed by Li and Barber (2015) [16], who stated and proved all the theorems in this section. Now before we state the main theorem of this section, we first define the modified FDR.

Definition. For $c \geq 0$, $\mathbb{E}\Big[V(k)/(c + R(k))\Big]$ is the *modified FDR* with parameter c, denoted mFDR_c. It is equal to the FDR when $c = 0$ and strictly smaller than the FDR otherwise.

Of course, if k is a stopping rule, we can just write $R(k) = k$.

Note that if $R(k)$ the number of rejections is moderately large with c quite small, this is not too far from the FDR, but otherwise, controlling mFDR is weaker than controlling FDR and has the potential to be much weaker. Now here's the main theorem. The setting is as in theorem 6. Suppose we have N ordered hypotheses with associated independent p-values, $H_1, ..., H_N$ and $p_1, ..., p_N$ and a subset $\mathcal{H}_0 \subseteq \{1, ..., N\}$ which indexes the null hypotheses (so that $p_i \overset{iid}{\sim} \text{Unif}[0, 1]$ for all $i \in \mathcal{H}_0$).

Theorem 7 (Accumulation tests)**.** *Let h be an accumulation function and let $q \in (0, 1)$ be a target FDR control level. Fix $C > 0$, and define the stopping rule*

$$\hat{k}_q^h = \max\left\{k \in \{1, ..., N\} : \frac{1}{k}\sum_{i=1}^{k} h(p_i) \leq q\right\}$$

Then

$$mFDR_{C/q}(\hat{k}_q^h) \leq \frac{q}{\int_{t=0}^{1} \min(h(t), C)dt}$$

This isn't as straightforward as ForwardStop. Notice that h does not depend on C; in fact, nothing depends on C except for the final control of the mFDR. In [16] C is simply taken to be 2. Also, the denominator of the right hand side is strictly less than 1, so we've lost exactly control at level q, and the quantity on the left side depends on q, as

$$\text{mFDR}_{C/q} = \mathbb{E}\left[V/\left(\frac{C}{q} + R\right)\right]$$

Taking q really, really small will tend to shrink the mFDR compared to the FDR on the left hand side. As for the right-hand side, we can get control at q if we pick $h(t)$ to be bounded and C to be its supremum, because then the denominator is 1. However, if the supremum is large, then the mFDR is much smaller than the actual FDR again[14]. So we see we must be reasonable when controlling mFDR, either picking h that is bounded by a small number or applying accumulation tests in situations where we will reject a moderate amount.

[14]Aside from this situation of bounded functions, where you can say something about exact control, there's no reason to "pick" different C; mFDR control holds true for all of them and the inequalities for different C seem equally useful.

I will defer the proof of this theorem to later, in favor of discussion now. It takes several lemmas and the proofs are not so terribly interesting, though I include them for completeness. The results, not the proofs, are the interesting thing here (one distinction between mathematics and statistics, I suppose).

Here are a few things to say about the accumulation test procedures.

- Is exact control of FDR important? In [16] Li and Barber also proposed a modified version of this method that gives exact control of the FDR, but I do not consider it here, because it saps away power. In section 10 we find that control of the modified FDR seems to not be so bad, and at least for that experiment we never exceeded the nominal level.
- The choice of h. We already see it might be advantageous to choose h to be bounded, for FDR control reasons. But what about power? As we march along k, we add up the values $h(p_i)$ until we spill over q, and we don't want a correct rejection to help accumulate evidence that we should stop rejecting. So $h(p_i)$ or $\mathbb{E}h(p_i)$ should be small on the non-nulls. Knowing the non-null distributions is impossible in principle, but today there are empiricial bayes methods that can do it [17], though they are not my focus.
- Where does ForwardStop fit in? We'll define two useful methods in this section, called SeqStep and HingeExp, both of which have more power than ForwardStop as seen in section 10. But they only control a modified FDR. To me so far it has seemed to be a worthwhile compromise because the power gains are large, but in a small signal setting, or when there are many nulls and few rejections we may want to use ForwardStop, or if we want to take q very small and reject stringently, because it has good power and controls FDR exactly.

We continue deferring the proof of theorem 7 to define the SeqStep [18] and HingeExp [16] stopping rules.

Definition (SeqStep)**.** In Theorem 7, take

$$h(p_i) = C \cdot \mathbb{1}\Big\{p_i > 1 - \frac{1}{C}\Big\}$$

So the definition of SeqStep depends on your choice of C; it is named for being a step function for sequential testing. The good thing about SeqStep is that in the class of bounded functions, it is in some sense the "best" choice, given a reasonable condition on the non-nulls.

Proposition. *Let h_S be the accumulation function corresponding to SeqStep where $C > 0$ is chosen. Consider any other accumulation function h bounded above by C. Suppose the non-null p-values p_i have a density $f_i : [0,1] \to [0,\infty)$, where f_i is nonincreasing.*

Then on those non-null p-values,

$$\mathbb{E}h(p_i) \geq \mathbb{E}h_S(p_i)$$

In the most recent bullet points we said that we should keep $\mathbb{E}h(p_i)$ small, and by this proposition SeqStep does just that.[15] We also note that the condition that f_i is nonincreasing is reasonable, because if a p-value is truly not null, it should have probability weight towards the low end of $[0, 1]$, not the high end.

Proof of Proposition.

$$\begin{aligned}
\mathbb{E}h(p_i) - \mathbb{E}h_S(p_i) &= \int_0^1 \big(h(t) - h_S(t)\big) \cdot f_i(t)\ dt \\
&= \int_0^{1-1/C} \big(h(t) - 0\big) \cdot f_i(t)\ dt + \int_{1-1/C}^1 \big(h(t) - C\big) \cdot f_i(t)\ dt \\
&= \int_0^{1-1/C} h(t) \cdot f_i(t)\ dt - \int_{1-1/C}^1 \big(C - h(t)\big) \cdot f_i(t)\ dt \\
\geq \int_0^{1-1/C} h(t) f_i(1-1/C)\, dt - \int_{1-1/C}^1 \big(C - h(t)\big) f_i(1-1/C)\, dt &\quad (f_i \text{ is non-increasing}) \\
= f_i(1-1/C) \cdot \left[\int_0^1 h(t)\ dt - \int_{1-1/C}^1 C\ dt\right] &= f_i(1-1/C) \cdot [1-1] = 0.
\end{aligned}$$

□

One remark. It is true that in the case of $\mathbb{E}h(p_i) - \mathbb{E}h_S(p_i) = 0$, then $h = h_S$ almost everywhere, but we leave the explanation to [16].

Now let's define HingeExp. Its name comes from the fact that it looks sort of like hinge loss from machine learning.

Definition (HingeExp)**.** In Theorem 7, take

$$h(p_i) = C \log\left(\frac{1}{C \cdot (1 - p_i)}\right) \mathbb{1}\Big\{p_i > 1 - \frac{1}{C}\Big\}$$

So HingeExp also depends on your choice of C. This function, which looks sort of like a fusion of ForwardStop and SeqStep, was proposed in [16] to combine the BH(q) heritage of ForwardStop with the optimality of SeqStep. There's currently not much to prove about it in particular, but it works really quite well.

Now we return to the proof of the main Theorem 7.

Unlike the proofs given in other sections, the proofs in this section 9 will follow extremely closely the original presentation in [18] and [16], which were very clear and left out very little (and which shared an author). In fact, I will be the one to leave things out; I will ask the interested reader to

[15] Of course this is not a proof that SeqStep actually has better power. In fact there is a proof in [18] which says that, given some somewhat technical conditions, that if $\mathbb{E}h_1(p_i) > \mathbb{E}h_2(p_i)$ on the non-nulls then asymptotically greater power is achieved. The proof is also somewhat technical (and very long) so we have left it out. Empirically we shall see the power gains.

refer to those references for the some of the omitted calculuations, which are presented with satisfying detail there.

We begin by proving a few lemmas.

Lemma 2. *Fix $c \in (0,1)$. Let $N^{(0)}$ be the largest element of $\mathcal{H}_0$. For $0 \leq k \leq \hat{N}$, put $V^+(k) = \#\{j \in \mathcal{H}_0 : 1 \leq j \leq k, p_j \leq c\}$ and $V^-(k) = \#\{j \in \mathcal{H}_0 : 1 \leq j \leq k, p_j > c\}$. Let $\mathcal{F}_k$ be the filtration generated by all the non-null p-values as well as $V^{\pm}(k')$ for all $k' \geq k$. Then the process*

$$M(k) = \frac{V^+(k)}{1 + V^-(k)}$$

is a supermartingale running in backwards k, adapted to $\mathcal{F}_k$, and for any k we have

$$\mathbb{E}\left[M(k)\right] \leq \frac{c}{1-c}$$

Proof. We assume $\mathcal{H}_0 = \{1, ..., \widehat{N}\}$ for some $\widehat{N}$. The proof for the general case is identical, but messier.

Because we know all the nonnulls whenever we condition on $\mathcal{F}_k$, we know whether or not k is null. If it is null then

$$M(k-1) = \frac{V^+(k) - I}{1 + V^-(k) - (1-I)} = \frac{V^+(k) - I}{\min(V^-(k) + I, 1)}$$

where $I = \mathbb{1}_{p_k \leq c}$. Otherwise if it is not null $M(k) = M(k-1)$. Conditioning on $\mathcal{F}_k$ tells us nothing more about I. What it does tell us however is for $k' \geq$ which of the null $p_{k'}$ was below c because then $V(k')$ goes up a notch when it does. Excluding those from consideration, at k we have

$$P(I = 1) = \frac{V^+(k)}{V^+(k) + V^-(k)}$$

where the denominator is just the total number of nulls left to consider by step k and the numerator is how many of them are known to be less than c, that is to say, how many we're allowed to pick from.

We'd like to calculate $\mathbb{E}\left[M(k-1)|\mathcal{F}_k\right]$. This is easily done by weighting the null possibilities by $P(I=1)$ and $P(I=0)$; we then find that $\mathbb{E}\left[M(k-1)|\mathcal{F}_k\right] \leq M(k)$, which establishes the supermarginale property. I omit the actual calculation and leave it for the reference [18].

To show the bound $\mathbb{E}M(k) = \mathbb{E}\left[\frac{V^+(k)}{1+V^-(k)}\right] \leq \frac{c}{1-c}$, we let $V(k) = V^+(k) + V^-(k)$ be the number of nulls from 1 to k (which is to say, the number of false rejections), and re-express

$$\mathbb{E}\left[\frac{V^+(k)}{1 + V^-(k)}\right] = \mathbb{E}\left[\frac{V^+(k)}{V(k) - V^+(k) + 1}\right]$$

where $V^+(k) \sim \text{Binomial}(V(k), c)$. Since the probability mass function is known, computing this expectation is a routine exercise; nevertheless it can be found in [18], though I omit it here.

The calculation proves the bound. □

Corrolary. *Let $B_1, ..., B_N \in \{0,1\}$ be independent, with $B_i \overset{iid}{\sim} Ber(\rho)$ for $i \in \mathcal{H}_0$. Let $\{\mathcal{F}_k\}_{k=N}^1$ be a filtration in reverse k such that*

- *$B_i \in \mathcal{F}_k$ for all $i \notin \mathcal{H}_0$, and for all $i > k$ with $i \in \mathcal{H}_0$*
- *$\sum_{i \leq k, i \in \mathcal{H}_0} B_i \in \mathcal{F}_k$*
- *The null B_i are exchangeable conditioned on $\mathcal{F}_k$*

Then

$$M(k) = \frac{1 + V(k)}{1 + \sum_{i \leq k, i \in \mathcal{H}_0} B_i}$$

is a supermartingale adapted to $\mathcal{F}_k$ and $\mathbb{E}M(k) \leq 1/\rho$ for all k.

Proof. Set $\rho = 1 - c$, where c is as in Lemma 2. Then identify $\sum_{i \leq k, i \in \mathcal{H}_0} B_i$ with $V^-(k)$ from Lemma 2. Let $M'(k)$ be the supermartingale from Lemma 2. Since a supermartingale plus a constant is still a supermartingale,

$$1 + M'(k) = 1 + \frac{V^+(k)}{1 + V^-(k)} = \frac{1 + V^+(k) + V^-(k)}{1 + V^-(k)} = \frac{1 + V(k)}{1 + V^-(k)} = M(t)$$

is still a supermartingale. Also,

$$\mathbb{E}M(k) = 1 + \mathbb{E}M'(k) \leq 1 + \frac{c}{1 - c} = \frac{1}{1 - c} = \frac{1}{\rho}.$$

□

Lemma 3. *Pick $C > 0$. Let $a_1, ..., a_N \geq 0$ and let h be an accumulation function. Define*

$$\hat{k} = \max \left\{ k \in \{1, ..., n\} : \sum_{i=1}^{k} h(p_i) \leq a_k \right\}$$

setting $\hat{k} = 0$ if it's otherwise undefined. Then

$$\mathbb{E}\left[\frac{1 + V(\hat{k})}{C + \sum_{i \leq \hat{k}, i \in \mathcal{H}_0} h(p_i)} \right] \leq \frac{1}{\int_0^1 \min(h(t), C) dt}$$

Proof. Define additional variables $U_i \overset{iid}{\sim} \text{Unif}[0,1]$ independently from the p_i's, and

$$B_i = \mathbb{1}\{U_i \leq h(p_i)/C\}$$

Then conditional on the $p_1, ..., p_N$, the B_i are independently

$$(B_i | p_1, ..., p_N) \overset{indep}{\sim} \text{Ber}\left(\min\left(\frac{h(p_i)}{C}, 1 \right) \right) = \text{Ber}\left(\frac{\min(h(p_i), C)}{C} \right) \tag{25}$$

Also, not conditioned on anything the null B_i are Bernoulli(ρ), where

$$\rho = \mathbb{E}\left[\frac{\max(h(p_i), C)}{C} \right] = \frac{1}{C} \int_0^1 \max(h(p_i), C) dt$$

We'd like to use the corollary. The filtration that works is where $\mathcal{F}_k$ is generated by knowing (p_i, U_i) for $i > k$ for $i \in \mathcal{H}_0$ and all the (p_i, U_i) when $i \notin \mathcal{H}_0$. One can verify that $\hat{k}$ is a stopping time with respect to this filtration. Then the corrolary, plus optional stopping[16], says

$$\mathbb{E}M(\hat{k}) = \mathbb{E}\left[\frac{1+V(\hat{k})}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i}\right] \le \frac{1}{\rho} = \frac{1}{\frac{1}{C}\int_0^1 \min(h(p_i), C)dt} \tag{26}$$

Next, we show

$$C\cdot\mathbb{E}\left[\frac{1+V(\hat{k})}{C+\sum_{i\le\hat{k},i\in\mathcal{H}_0} h(p_i)}\right] \le \mathbb{E}\left[\frac{1+V(\hat{k})}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i}\right]$$

which will prove the lemma when combined with (26).

$$\begin{aligned}
\mathbb{E}\left[\frac{1+V(\hat{k})}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{1+V(\hat{k})}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i}\Big| p_1,...,p_N\right]\right] \\
&= \mathbb{E}\left[(1+V(\hat{k}))\cdot\mathbb{E}\left[\frac{1}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i}\Big| p_1,...,p_N\right]\right] \\
&\ge \mathbb{E}\left[(1+V(\hat{k}))\cdot\frac{1}{\mathbb{E}\left[1+\sum_{i\le\hat{k},i\in\mathcal{H}_0} B_i | p_1,...,p_N\right]}\right]
\end{aligned}$$

where we've applied Jensen's inequality to $1/x$.

Then by (25)

$$= \mathbb{E}\left[\frac{1+V(\hat{k})}{1+\sum_{i\le\hat{k},i\in\mathcal{H}_0}\frac{\max(h(p_i),C)}{C}}\right] \ge C\cdot\mathbb{E}\left[\frac{1+V(\hat{k})}{C+\sum_{i\le\hat{k},i\in\mathcal{H}_0} h(p_i)}\right]$$

□

Finally we can prove the theorem.

Proof of Theorem 7.

$$\mathbb{E}[\mathrm{mFDP}_{C/q}(\hat{k}_h)] = \mathbb{E}\left[\frac{V(\hat{k}_q^h)}{C/q+\hat{k}_q^h}\right] = \mathbb{E}\left[\frac{V(\hat{k}_q^h)}{C+\sum_{i=1}^{\hat{k}_q^h} h(p_i)}\cdot\frac{C+\sum_{i=1}^{\hat{k}_q^h} h(p_i)}{C/q+\hat{k}_q^h}\right]$$

By definition of $\hat{k}_q^h$, and then simplifying,

$$\le \mathbb{E}\left[\frac{V(\hat{k}_q^h)}{C+\sum_{i=1}^{\hat{k}_q^h} h(p_i)}\cdot\frac{C+q\hat{k}_q^h}{C/q+\hat{k}_q^h}\right] = q\cdot\mathbb{E}\left[\frac{V(\hat{k}_q^h)}{C+\sum_{i=1}^{\hat{k}_q^h} h(p_i)}\right]$$

[16]We proved it for submartingales. It's also true for supermartingales because you can just multiply by negative signs to turn a supermartingale to a submartingale, apply the theorem, and then multiply by another minus.

(now we can think of the term as an estimate of mFDP.) Add 1 to the top and remove terms from the bottom to get

$$\leq q \cdot \mathbb{E}\left[\frac{1+V(\hat{k}_q^h)}{C+\sum_{i\leq\hat{k},i\in\mathcal{H}_0} h(p_i)}\right]$$

Now direct application of lemma 3 gives the result. We see that lemma 3 was indeed the crucial bound.

□

10. Simulation results

For the following results, we referenced R code from the website accompanying [16] as well as the R package implementing [19] and modified them to implement the Lasso-G test (the covariance test was already done), obtain p-values per section 3, and then pipe them into the methods in 8 and 9. We relied on the default plotting features.

For $n = 3000$ and $p = 1000$, we defined a vector $\boldsymbol{\beta}^* \in \mathbb{R}^{1000}$ which had the initial 200 entries nonzero. We looked at two settings; moderate signal (nonzero entries were 9) and strong signal (nonzero entries were 20). We generated orthogonal matrices $\boldsymbol{X}$ and data $\boldsymbol{y}$ based on the linear model (1), computed the lasso path, and computed both covariance test p-values and lasso-G p-values. Then each of ForwardStop, SeqStop with $C = 2$, and HingeExp with $C = 2$ were performed on these p-values in the order of the lasso path for varying FDR targets q. (Also pictured, in black, is SeqStep+, which is the modification of SeqStep in [16] that controls FDR rather than mFDR and which we mentioned back in section 9, but it has such low power that we did not bother to prove anything about it.)

This was performed 1000 times. The empirical power, which is the number of non-null hypotheses that were correctly rejected, was averaged over these 1000 times as q varied. Likewise the empirical FDR, which is the number of false rejections over rejections, was averaged 1000 times for varying q. The plots are shown in figures 2, 3, and 4.

The results are that FDR is successfully controlled in every situation (the observed FDR lies under the dotted line), even though SeqStep and HingeExp only control a modified FDR. HingeExp is the best performing method, and the covariance test did not obtain power greater than 0.7 on $q \in [0, 0.25]$ in the moderate signal regime, but it did on the strong signal regime.

However, the Lasso-G test, based on the Gumbel, maxed out on observed power almost immediately. The observed FDR was higher but still controlled. Evidently Lasso-G is not as conservative as the covariance test. In addition to the robustness reported by Cai and Yuan, it seems that Lasso-G is superior in this problem.

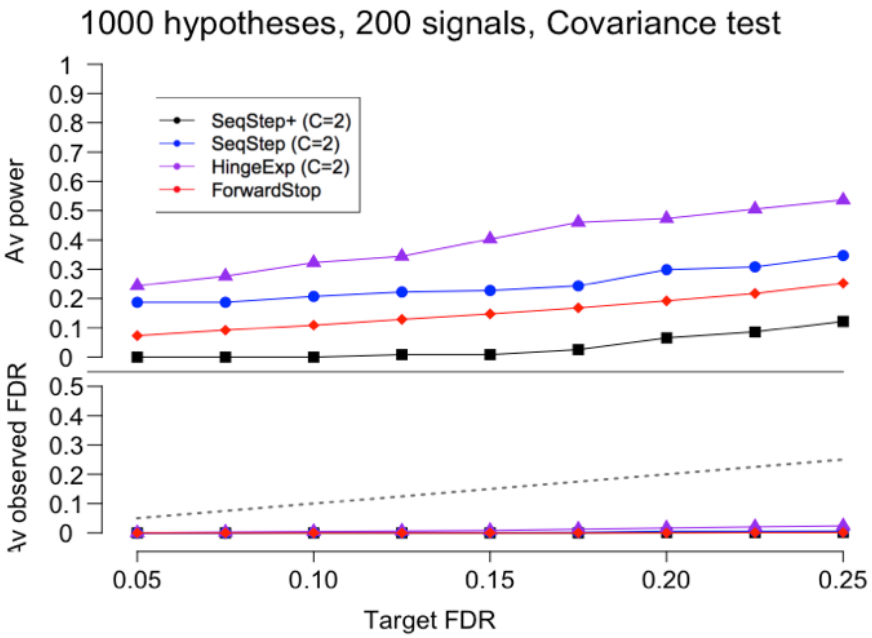


FIGURE 2. Covariance test with $\boldsymbol{\beta}^* = (9, 9, 9, ..., 0, 0, ...)$

11. Discussion and Conclusion

In this work we have presented two alternative test statistics and three alternative ways to combine them to control FDR. It appears that, for the orthogonal problem, the recommendation is to use Lasso-G with HingeExp. But we have analyzed a restricted set of methods under restricted assumptions and thereby only arrived at a partial answer. Good statistical practice is almost as much philosophy as it is mathematics. I would like to use this discussion to reflect on whether other methods might answer our question of model selection better, or if even we are asking the right question. In particular, the methods this work examines were all devised by 2016. How much has the field changed since then?

Robustness to nonindependence. Orthogonal design is a very restrictive condition. It simplifies many proofs, intuitively because we have stamped out the issue of multicollinearity and every OLS solution can be considered without affecting the others. Additionally, the assumption is necessary for independent p-values in the covariance test situation; the analogous result of Lockhart et. al for a general matrix $\boldsymbol{X}$ does not guarantee independence of p-values, which is an assumption in all our sequential procedures.

Indeed, getting p-values is not a real problem. The spacing test [20] is another way to get p-values and is even valid in finite samples for correlated

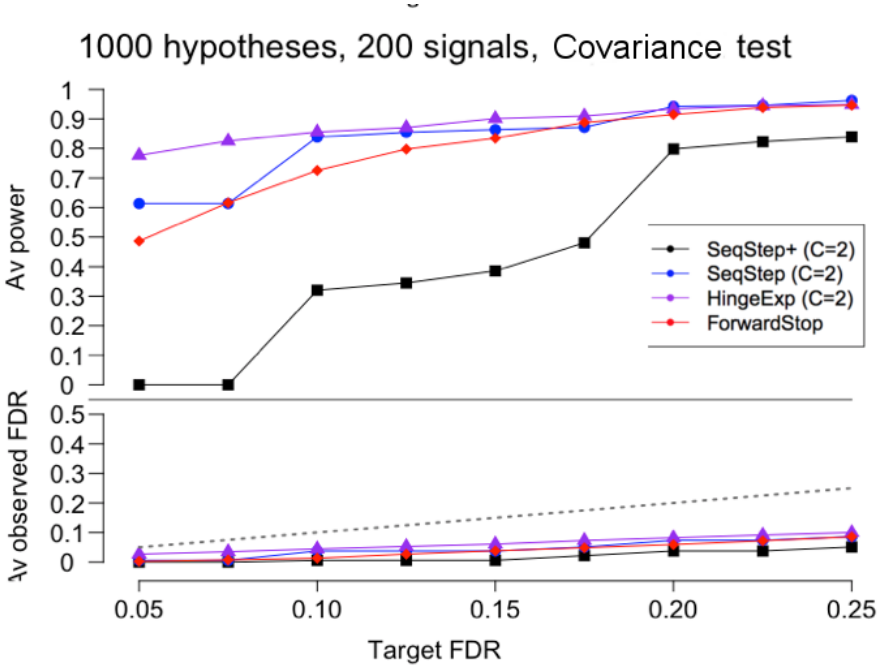


FIGURE 3. Covariance test with $\boldsymbol{\beta}^* = (20, 20, 20, ..., 0, 0, ...)$

matrices, and we have seen that the Lasso-G test is resistant to nonorthogonality. But there does not seem to be a way to get independent p-values. The Benjamini-Yekutieli procedure [21] for unordered hypotheses controls FDR under dependence, and BH(q) alone actually works under a kind of positive dependence [22], but no sequential stopping rules seem to have been devised that take advantage of these properties.

One thing that we do have is that under nonorthogonality, variables may enter and leave the model in the lasso, so that the null hypotheses are not nested, but all of our sequential hypothesis testing methods can handle some nulls interspersed along the non-nulls.

Which hypotheses? My discussion here parallels a passage in [23]. Our null hypotheses (5), which say the true support is contained in the current model, are conditional on the data as well as the whole sequence of hypotheses, and are typical in the literature. One approach is to condition on just the current null model, rather than the whole sequences of hypotheses; this is the approach in [23], and the result is power, with the drawback of computational difficulty.

Another hypothesis that is not conditional on the data is simply H_{0k} : $\beta_k = 0$. This is not a sequential testing problem anymore, and it has a different interpretation, being variable selection rather than model selection. If β_k is well correlated with other variables, it becomes difficult to interpret

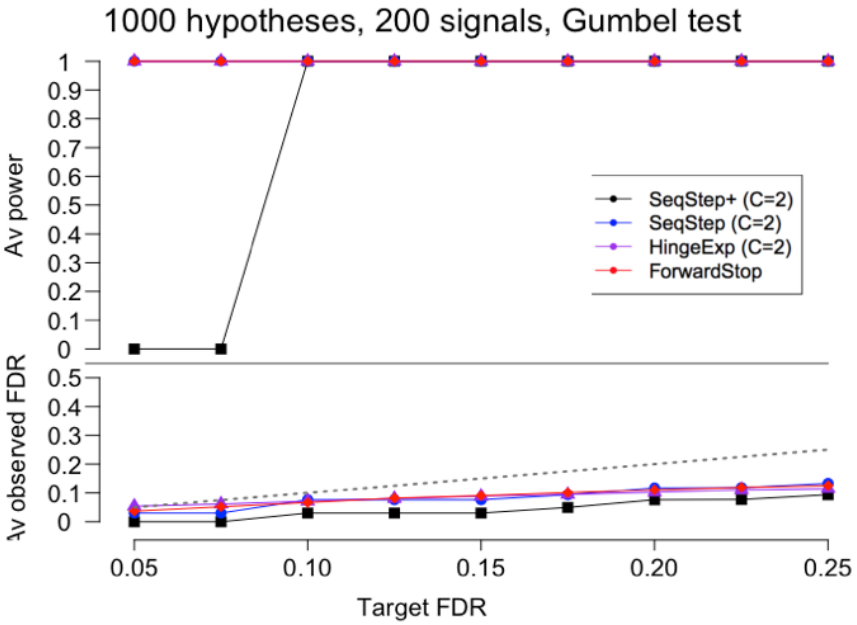


FIGURE 4. Lasso-G test with $\boldsymbol{\beta}^* = (9, 9, 9, ..., 0, 0, ...)$

a rejection in this case without knowing the nature of the correlation, and if p is large it may be impossible. If the model has some special scientific status and we know something about the other covariates then we may wish to test $H_{0k} : \beta_k = 0$; we may use debiased lasso methods [24] or desparsified lasso methods [25]. As for FDR control, in this setting there is exactly one method which achieves it, without even involving any null distributions or p-values. It is called the knockoff filter.

The knockoff. The knockoff [18] is one of the most important results in the FDR control literature and is many papers just by itself. It achieves FDR control of the hypotheses $H_{0k} : \beta_k = 0$ under completely general conditions on $\boldsymbol{X}$. The idea is to look at when variables enter the lasso path, denoted $Z_j = \sup\{\lambda : \hat{\beta}_j \neq 0\}$, and reject all H_j for which $Z_j \geq T$ where T is some data dependent threshold. How can T be calibrated to achieve FDR control? The knockoff method constructs fake data $\widetilde{\boldsymbol{X}}$ which has (almost) the same correlation structure as $\boldsymbol{X}$ is, but is (almost) independent from $\boldsymbol{X}$. Then the data $\widetilde{\boldsymbol{X}}$ are added to the model and the lasso path is computed. The knockoff variables are all null, so we can calibrate T based on whether β_j enters the model much earlier compared to its knockoff $\tilde{\beta}_j$, which would be good evidence that $\beta_j \neq 0$.

It turns out, and is shown in [18], that if you modify SeqStep (see section 9) slightly then the knockoff can be derived as a special case of it, and indeed the knockoff controls not FDR but some mFDR. Necessarily this

modification is no longer a sequential stopping rule, because the knockoff is not one. Personally, I wonder if some knockoff-esque method could be derived by considering the more powerful stopping rule, HingeExp.

So, what's the best method? For sequential stopping rules, the best we've got seems to be HingeExp, which assumes independence of the p-values (unless you really wanted to control FDR exactly, in which case you use ForwardStop). But the question of what p-values to use, and which hypotheses, is harder to decide, not least because of correlated design muddling the issue. This goes away if you wished to test $H_{0k} : \beta_k = 0$, where the knockoff is hands down the best method because it works under correlation. But do you want to? Ultimately I think it's probably a matter of philosophy, but my thoughts are not fully fleshed out. I suppose the point of this concluding discussion, aside from a gentle curtain close to my honors thesis, was to try to organize my thoughts in the hopes of learning an answer. Maybe I'll have a better one in a few years.

12. References

[1] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity*. CRC Press, 2015.

[2] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7:1456–1490, 2013.

[3] Laurens De Haan and Ana Ferreira. *Extreme Value Theory*. Springer, 2006.

[4] Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 04 2012.

[5] Wenjing Yin. Robust significance testing in sparse and high dimensional linear models, 2015. URL: https://math.ucsd.edu/programs/undergraduate/1415_honors_presentations/Wenjing_Yin_Honors_Thesis.pdf. Last visited on 2018/03/20.

[6] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of Statistics* 2014, Vol. 42, No. 2, 413-468, 2013, arXiv:1301.7161.

[7] Robert D. Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statist.*, 12(3):364–366, 09 1941.

[8] T. Tony Cai and Ming Yuan. Discussion: A significance test for the lasso. *Ann. Statist.*, 42(2):478–482, 04 2014.

[9] Peter Hall. On the rate of convergence of normal extremes. *Journal of Applied Probability*, 16(2):433–439, 1979.

[10] W. S. Noble. How does multiple testing correction work? *Nat. Biotechnol.*, 27(12):1135–1137, Dec 2009.

[11] Sidney I. Resnick. *A Probability Path*. Birkhauser, 2005.

[12] Gordan Zitkovic. Theory of Probability II Lecture Notes: Abstract Nonsense, 2015. URL: https://www.ma.utexas.edu/users/gordanz/notes/nonsense.pdf. Last visited on 2018/03/20.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[14] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency

of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2004.00439.x.

[15] Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12122.

[16] Ang Li and Rina Foygel Barber. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017, https://doi.org/10.1080/01621459.2016.1180989. Accompanying website: http://www.stat.uchicago.edu/~rina/accumulationtests.html. Last visited on 2018/03/20.

[17] Bradley Efron. *Large Scale Inference*. Cambridge University Press, 2013.

[18] Rina Foygel Barber and Emmanuel J. Candes. Controlling the false discovery rate via knockoffs. *Annals of Statistics 2015*, Vol. 43, No. 5, 2055-2085, 2014, arXiv:1404.5609.

[19] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics* 2004, Vol. 32, No. 2, 407-451, 2004, arXiv:math/0406456.

[20] Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016, https://doi.org/10.1080/01621459.2015.1108848.

[21] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196, 12 1999.

[22] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 08 2001.

[23] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection, 2015, arXiv:1512.02565.

[24] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. 2013, arXiv:1306.3171.

[25] Sara van de Geer, Peter Buhlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics 2014*, Vol. 42, No. 3, 1166-1202, 2013, arXiv:1303.0518.