Summary and discussion of: "Sequential selection procedures and false discovery rate control"

Statistics Journal Club, 36-825

Mattia Ciollaro and Calvin McCarter

Acknowledgements: Sincere thanks to Max Grazier G'Sell for taking the time to answer our questions about the paper and for the very useful conversations.

1 Summary

1.1 Preliminaries: Sequential Tests and FDR

Suppose we have a sequence of null hypotheses $H_1, \ldots, H_m$ and that we want to reject some of these hypotheses while controlling the expected number of 'mistakes' (in the sense of FDR, to be explained later). We also have a **constraint** on the way in which we can reject the hypotheses. Namely, the **hypotheses have to be rejected in an ordered fashion**: if the i-th hypothesis H_i is rejected, then all the preceding ones, $H_1, \ldots, H_{i-1}$, must be rejected as well (think of **step-wise regression**, but also **lasso** or **hierarchical clustering** for example).
We want to guarantee that the rule used to perform this ordered sequence of hypothesis tests **controls the proportion of false discoveries**, which is defined as

$$FDR = E\left(\frac{V}{R}\right)$$

with respect to the table below.

	null is true	alternative is true	
null is rejected	V	S	R
null is not rejected	U	T	$m - R$
	m_0	$m - m_0$	m

If there was no sequentiality in the hypothesis test, i.e. if this was a standard multiple testing problem, the well-known Benjiamini-Hochberg (BH) procedure[1] would provide a means to reject a subset of the m hypotheses and control the FDR at any level $\alpha \in (0, 1)$. Remember that the BH procedure amounts to the following three steps:

1. compute the p-values for the m hypotheses, $p_1, \ldots, p_m$
2. order the p-values in increasing order: $p_{(1)}, \ldots, p_{(m)}$

3. reject the hypotheses for which $p_{(\ell)} \leq \frac{\alpha \ell}{m}$.

In the ordered hypotheses testing setting, BH cannot be directly applied because of sequentiality. However, **if we had a way of transforming the sequence of p-values into a monotone increasing sequence of statistics $q_1, \ldots, q_m$, then we could apply the BH procedure on this sequence and thus obtain FDR control in the ordered setting as well**.

1.2 FDR Control for Ordered Hypothesis Tests

Among the m null hypotheses, assume that a subset $M \subset \{1, \ldots, m\}$ of them are true. Then

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

We can reject the first k hypotheses for a k of our choice. The goal is to find a procedure such that k is as large as possible and, at the same time, the FDR is controlled at a pre-specified level α.

1.2.1 Key Idea of the Paper

Assume temporarily the global null hypothesis according to which all of the m null hypotheses are in fact true, i.e.

$$\{p_i : i \in \{1, \ldots, m\}\} \overset{\text{iid}}{\sim} \mathrm{U}(0,1).$$

In this case, consider

$$\begin{aligned}
Y_i &= -\log(1 - p_i) && \text{uniform to exponential transform} \\
Z_i &= \sum_{j=1}^{i} \frac{Y_i}{m - j + 1} && \text{Rényi representation theorem} \\
q_i &= 1 - e^{-Z_i} && \text{back from (ordered) exponential to (ordered) uniform}
\end{aligned}$$

The Y_i's are iid exponential random variables and, thanks to a representation theorem by Rényi [1], the $Z's$ are distributed as exponential order statistics (i.e. as a sorted list of iid exponential random variables). The q_i's are therefore to be considered as 'ordered' p-values on which we can now apply the standard BH procedure. This suggests that k (the number of initial hypotheses to be rejected) should be determined according to the rule

$$\hat{k}_F^q = \max \left\{ k \in \{1, \ldots, m\} : q_k \leq \frac{\alpha k}{m} \right\}.$$

[1] In words, the representation theorem provides a way of mapping a collection of m iid exponential random variables to a collection of random variables whose joint distribution corresponds to the distribution of the first m exponential order statistics (and viceversa).

1.2.2 Some Surprises and Main Control Procedures

When the global null hypothesis is not true, the Rényi representation theorem does not hold. However, (**Surprise 1**) the above procedure still controls the FDR at any given level α! Instead of assuming the global null to be true, assume for the moment that the first s null hypotheses are false and the subsequent $m - s$ are true. Then we have

Lemma 1. *Suppose that the first s null hypotheses are false and the latter $m - s$ are true. The rule $\hat{k}_F^q$ controls the FDR at any pre-specified level α.*

Notice, however, that the q_i's constructed before depend on m. Given that the Rényi representation only works under the global null, it is reasonable to ask whether it is possible to derive a stopping criterion that does not rely on the corresponding transformation. Such criterion can be obtained by appending a list of null test statistics after the first m of interest (thus essentially letting $m \to \infty$). Then we have

$$q_k = 1 - e^{-Z_k} \approx \sum_{j=1}^{k} \frac{Y_j}{m - j + 1}$$

hence the rule

$$q_k = 1 - e^{-Z_k} \leq \frac{\alpha k}{m}$$

is approximated by the rule

$$\sum_{j=1}^{k} \frac{Y_j}{m - j + 1} \leq \frac{\alpha k}{m}$$

which in the limit becomes

$$\frac{1}{k} \sum_{j=1}^{k} Y_j \leq \alpha.$$

This suggests the **ForwardStop** rule

$$\hat{k}_F = \max \left\{ k \in \{1, \ldots, m\} : \frac{1}{k} \sum_{j=1}^{k} Y_j \leq \alpha \right\}.$$

The authors prove the following Corollary (**Surprise 2**).

Corollary 1. *Under the conditions of Lemma 1 (the s false null hypotheses are concentrated at the beginning of the list and precede the true $m - s$ null hypotheses), the rule $\hat{k}_F$ controls the FDR at any pre-specified level α.*

Interestingly enough, **the FDR is still controlled by the asymptotic procedure when the assumption that the s false null hypotheses are all concentrated before the true null hypotheses is removed (Surprise 3)**.

Theorem 1. *The stopping rule ForwardStop corresponding to $\hat{k}_F$ controls the FDR at any pre-specified level α under the setup described at* (1).

Looking at the p-values from left to right is somewhat arbitrary. We might also scan them from right to left. In particular, under the global null hypothesis that all of the m null hypotheses are true, we could define

$$\tilde{Y}_i = -\log(p_i) \qquad \text{uniform to exponential transform}$$
$$\tilde{Z}_i = \sum_{j=i}^{m} \frac{\tilde{Y}_j}{j} \qquad \text{Rényi representation theorem}$$
$$\tilde{q}_i = e^{-\tilde{Z}_i} \qquad \text{back from (ordered) exponential to (ordered) uniform}$$

This suggest the second procedure, **StrongStop**, which correspond to the following stopping rule:

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

We have the following Theorem, which gives **stronger theoretical guarantees to StrongStop** (compared to ForwardStop) **under the assumption that all the false null hypotheses are concentrated at the beginning of the list**.

Theorem 2. *The stopping rule StrongStop corresponding to $\hat{k}_S$ controls the family-wise error rate (FWER) at any pre-specified level α if the first s null hypotheses are false and the last $m - s$ are true. This means that the probability of making even just one false discovery is controlled:*

$$P\left(\hat{k}_S > s\right) \leq \alpha.$$

1.2.3 Competitor methods

The authors compare the proposed stopping rules with two competitors:

- thresholding at α: reject the hyoptheses up to the first time that a p-value exceeds α. This procedure controls both the FDR and the FWER at level α
- a tailored version of α-investing (Foster and Stine, 2008): $\hat{k}_{\text{invest}} = \min\left\{k : p_{k+1} > \frac{(k+1)\alpha}{1+(k+1)\alpha}\right\}$. This procedure controls $\frac{EV}{ER+1}$ (the so-called 'marginal FDR') at level α.

These procedures tend to be less powerful then ForwardStop, as they stop at the first 'medium-sized' p-value even though other small p-values follow in the list. This intuition is confirmed by some of the authors' simulations in the paper.

The StrongStop procedure is shown to be the most powerful for the spacing test. Also, StrongStop has the lowest observed FDR in the simulations performed by the authors. The intuition behind this fact is the following: while the distribution of the p-values (or of the test statistics) under the null hypotheses is known, in the sequential selection setting the distribution under the alternative hypothesis is much less well-behaved. In particular, early 'spikes' of moderate p-values in the first hypotheses can cause ForwardStop to stop too early (making the procedure not very powerful). On the other hand, StrongStop scans the p-values from right to left (i.e. from the true null hypotheses domain towards the false null hypotheses domain). StrongStop achieves greater power because it spots the 'discontinuity' arising in the transition from the distribution of the p-values under the null (uniform) towards the (possibly spiky) distribution of the p-values under the alternative.

1.3 Harmonic p-values

1.3.1 Motivating Example

Consider the orthogonal lasso model where $X \in \mathbb{R}^{n\times p}$ is the matrix containing the p orthogonal predictors and $X^TX = I$. Suppose that there are s signal (non-null) variables. At each step along the lasso path, we consider testing the hypothesis $H_{0,j}$ that all of the s signal predictors are contained in the current lasso model with $j-1$ predictors. For this purpose, Lockhart et al.[4] introduced the statistics

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

where $\lambda_1 \geq \lambda_2 \geq \ldots$ are the values of the regularization parameter in the lasso problem

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

at which the sparsity of the minimizer $\hat{\beta}$ changes. If the null $H_{0,s+1}$ is true, Lockhart et al. showed that

$$T_{s+1}, T_{s+2}, \ldots, T_{s+\ell} \xrightarrow{\mathrm{d}} \mathrm{Exp}(1), \mathrm{Exp}(1/2), \ldots, \mathrm{Exp}(1/\ell)$$

in the limit as $n, p \to \infty$ for finite ℓ. **However, in order to obtain p-values from the T's, one needs to know s.** Otherwise, the harmonic exponential distributions can be bounded by Exp(1), but that would yield conservative p-values and consequently reduced power.

1.3.2 FDR Control for Harmonic Test Statistics

The goal is to modify the StrongStop rule in order to take advantage of the exponential behavior of the test statistics under the null hypothesis.

Assume that we have a sequence of test statistics $T_1, \ldots, T_m$ and that only the first s statistics correspond to signal variables. Assume further that the subsequent statistics are independently distributed as

$$T_{s+1}, T_{s+2}, \ldots, T_m \sim \mathrm{Exp}(1), \mathrm{Exp}(1/2), \ldots, \mathrm{Exp}(1/(m-s)).$$

Suppose momentarily that we know s. We have

$$jT_{s+j} \sim \mathrm{Exp}(1)$$

(the Gamma family has a closure property w.r.t. multiplication by a scalar). The StrongStop rule would suggest a test based on

$$q_i^* = \exp\left(-\sum_{j=i}^{m} \frac{\max\{1, j-s\}}{j} T_j\right).$$

However, s is unknown. If we set anti-conservatively $s = 0$ (corresponding to the hypothesis that there are no signal variables, in which case any included variable would be a false discovery) we get

$$q_i^* = \exp\left(-\sum_{j=i}^{m} T_j\right).$$

Now, applying the BH procedure yields the **TailStop** rule

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

The choice $s = 0$ is strongly anti-conservative, so we would expect to lose the strong control property of the StrongStop procedure. Unexpectedly, TailStop still controls the FDR nearly exactly as shown in the following Theorem.

Theorem 3. *We have*

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

Caveat: current results with harmonic exponential test statistics under the null are only asymptotic. As far as we know, the harmonic exponential behavior of the test statistics is characteristic to very specialized scenarios.

1.4 Simulation: ForwardStop, StrongStop and Competitors

The aim is to compare the performance of ForwardStop, StrongStop, and the two competitors methods of Section 1.2.3, i.e. thresholding at α and *alpha*-investing. We aim at comparing the power of the four methods and their ability to effectively control the FDR. We consider two settings:

- ordered hypotheses: the first s hypotheses are false and the remaining $m - s$ are true
- unordered hypotheses: the true and false null hypotheses are intermixed.

In our simulations, we set $m = 100$, $s = 20$, and we consider five levels of required nominal FDR, $\alpha = 0.05, 0.10, 0.15, 0.20, 0.25$. In order to compute the observed FDR and the average power for each method (defined as the fraction of correctly rejected hypotheses), we average the results over 500 runs.

1.4.1 Ordered Setting

For the first s false null hypotheses, we generate the corresponding p-values by drawing from a Beta$(1, \beta)$ distribution. The parameter β is allowed to take four values, $\beta = 4, 8, 14, 23$. Low values of β correspond to low-signal settings and therefore to harder settings. The p-values for the true null hypotheses are drawn from a standard uniform distribution. Figure 1 depicts the four beta distributions used to generate the p-values for the false null hypotheses. A subset of the simulated p-values for the four different difficulty levels is depicted in Figure 2. The results of the simulations for the ordered setting are summarized in Figure 3.

By looking at Figure 3, we notice that all the methods meet the required FDR under all the four difficulty settings. Thresholding at α and StrongStop appear much more conservative in the FDR control than the other two methods. The α-investing procedure, together with ForwardStop, appears to be the most powerful. We notice that as the difficulty of the scenario increases, ForwardStop becomes dominated by α-investing in terms of power (remember, however, that while ForwardStop is guaranteed to control the FDR, α-investing is only guaranteed to control $EV/(ER + 1)$).

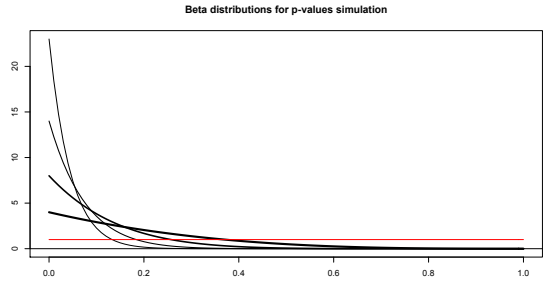


Figure 1: Beta distributions used for the simulation of the p-values. The red line is the uniform distribution used to draw the p-values for the true null hypotheses. Increasingly thick black beta densities correspond to increasingly harder simulation settings.

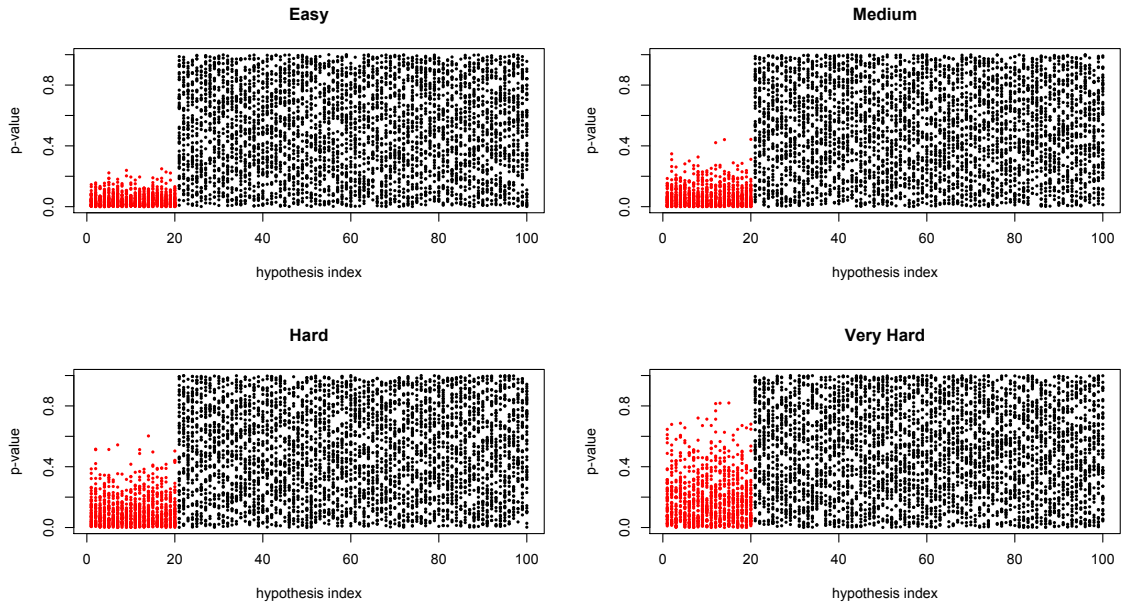


Figure 2: Realizations of the simulated p-values in the ordered hypotheses setting under the four level of difficulty considered. Red dots correspond to p-values for the false null hypotheses, black dots correspond to p-values for the true null hypotheses.

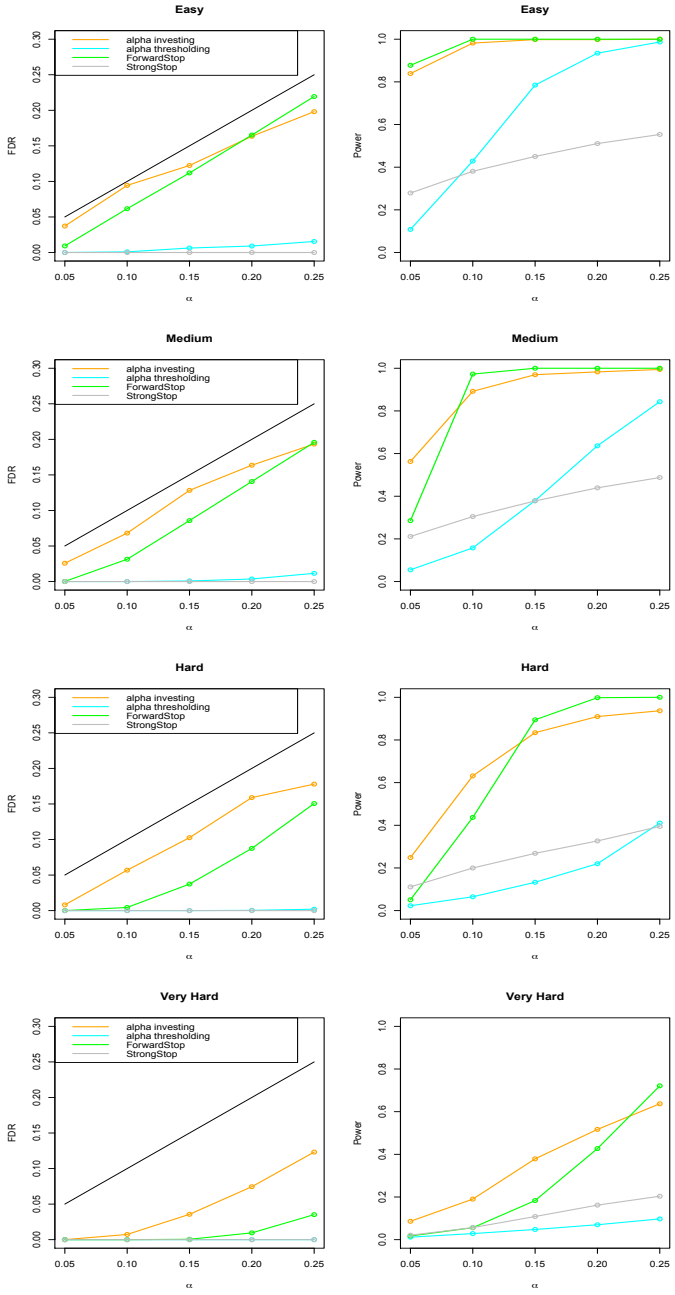


Figure 3: FDR and power analysis for the four methods in the ordered setting.

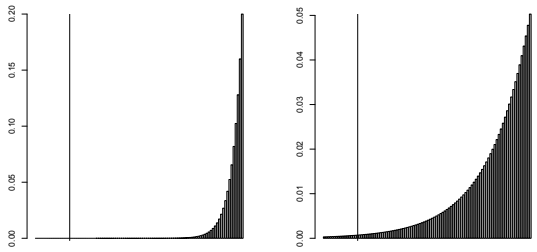


Figure 4: Probability mass function used to draw the indices for the true null hypotheses in the weak intermixing setting (left) and in the strong intermixing setting (right). The vertical line separates the first $s = 20$ and the latter $m - s = 80$ indices.

1.4.2 Unordered Setting

In this setting, the true and the false null hypotheses are intermixed. In particular, for each simulation we draw without replacement $m - s$ indices for the true null hypotheses from the following probability mass function

$$p(x) = \frac{(1-r)r^{x-1}}{1-r^m} \mathbb{1}_{\{1,\dots,m\}}(x)$$

for two choices of r, namely $r = 0.8$ (slight intermixing) and $r = 0.95$ (strong intermixing). Figure 4 contains the barplots for the two considered distributions. A subset of the simulated p-values under the mixing induced by the above probability mass function is depicted in Figure 5. The results of the simulations for the unordered setting are depicted in Figures 6a and 6b. Qualitatively, the results under weak intermixing resemble the results obtained in the ordered setting. Although StrongStop still appears to conservatively control the FDR, remember that in the unordered setting there are no theoretical guarantees regarding FDR control using the StrongStop procedure. The results in the strong mixing scenario are different: StrongStop and α-investing are no longer controlling the FDR (in fact, they are not expected to); ForwardStop and thresholding at α control the FDR as expected, although ForwardStop clearly has higher power.

1.5 Simulation: Model Selection for the Graphical Lasso

We compare the sequential selection procedures using significance tests for the graphical lasso.[3] In this setting, each significance test is performed when an additional edge merges two connected components in the graph structure. We compare the ForwardStop, StrongStop, and TailStop procedures where the true graph structure over 100 variables includes 53 connected components. Of these connected components, 50 are singletons, while 3 components contain the 50 remaining variables.

We ran all three procedures at fixed $\alpha = 0.10$ on datasets consisting of 10,000 and 20,000 samples, depicted in Figure 7a and Figure 7b, respectively. We observed that the p-values on the left side (where the alternative hypothesis holds) converged slowly to 0, with large occasional large p-values for correct merges. Thus, ForwardStop performs poorly

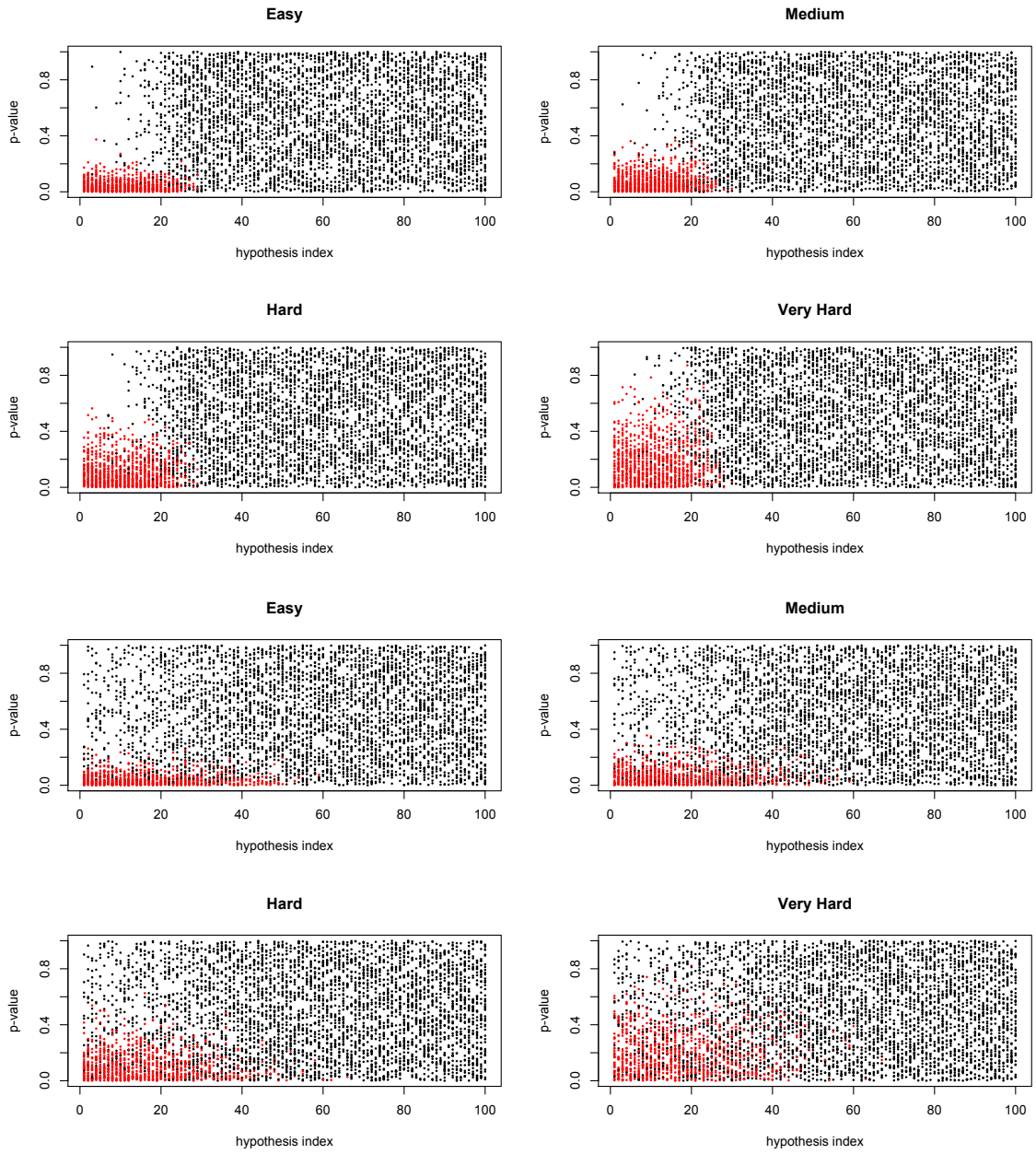
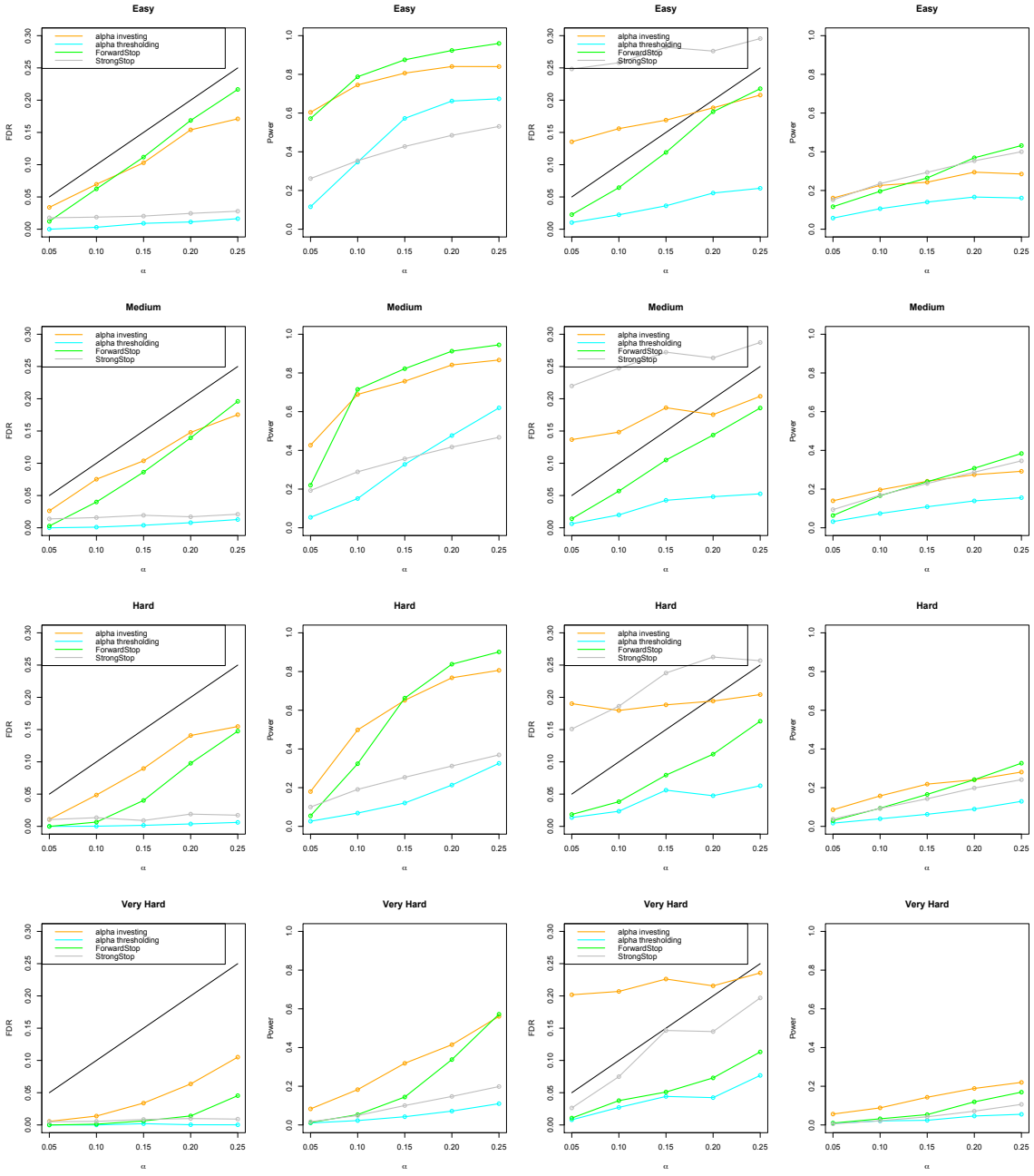
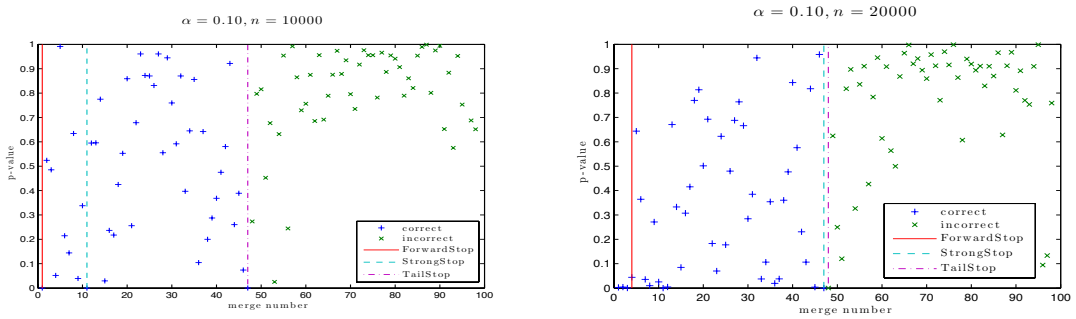


Figure 5: Realizations of the simulated p-values in the intermixed hypotheses setting under the four level of difficulty considered. Red dots correspond to p-values for the false null hypotheses, black dots correspond to p-values for the true null hypotheses. The top four panels correspond to the low mixing setting ($r = 0.8$) whereas the bottom four panels correspond to the strong mixing setting ($r = 0.95$).



(a) Slight intermixing (b) Strong intermixing

Figure 6: FDR and power analysis for the four methods in the unordered setting.



(a) Model selection with 10,000 samples.

(b) Model selection with 20,000 samples.

Figure 7: p-values and selected models for ForwardStop, StrongStop, and TailStop at $\alpha = 0.1$.

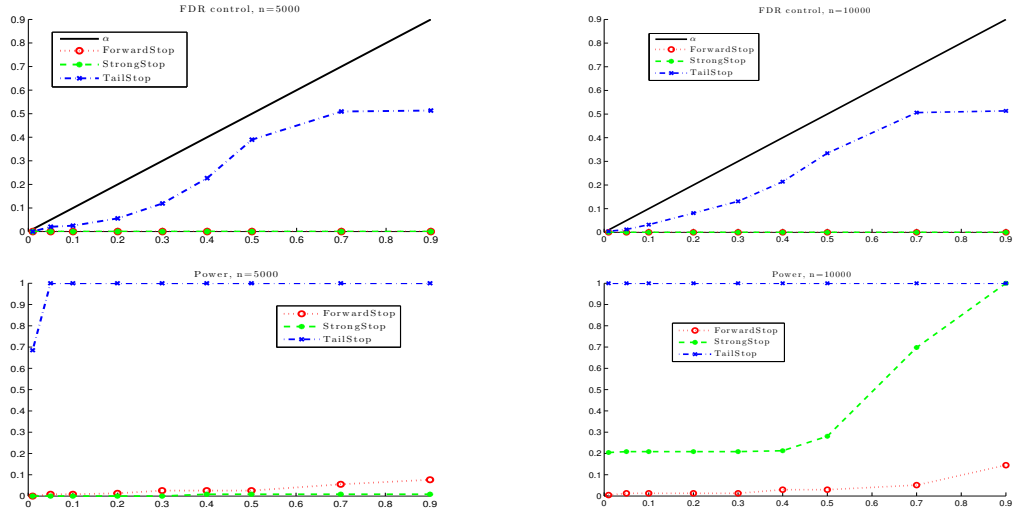


Figure 8: FDR and power for 5,000 samples and 10,000 samples, respectively.

in this setting, while StrongStop and TailStop work better by moving from the right hand side. As shown in the paper describing the significance test for the graphical lasso, the p-values under the null hypotheses exhibit harmonic behavior. Since it can exploit harmonic p-values, TailStop has greater power than StrongStop at smaller sample sizes.

We also compared the three procedures in terms of FDR and power for a range of α values, depicted in Figure 8. As expected, all three procedures control FDR, with ForwardStop and StrongStop being extremely conservative. TailStop exhibits much greater power than both ForwardStop and StrongStop at both sample sizes, although StrongStop seems to have high power when both the sample size and α are very large.

2 Discussion

A number of questions and comments concerned the use of the proposed methods in the lasso setting. The importance of orthogonal design was emphasized, as it leads to a monotone regularization path, which leads to the asyptotic harmonic behavior exploited by TailStop. Section 7.2 of the paper has a brief discussion about FDR control for non-orthogonal design and provides references to other methods that have theoretical guarantees when the orthogonality assumption is removed. In the case of orthogonal design, the theory guarantees *asymptotic* independence and *asymptotic* exponential harmonic behavior of the statistics under the null (note that independence of the test p-values is assumed throughout in the paper; also, the harmonic exponential distribution of the test statistics is assumed to be exact in the paper, rather than approximate). An interesting question concerns the robustness of the proposed methods in the setting when the test statistics (and the corresponding p-values) are not independent (as it is typically the case for non-orthogonal design across the steps of the LARS procedure).

The interpretation of the procedure in the lasso setting was also discussed. Max Grazier G'Sell suggested potential applications where sequential selection for the lasso would be useful. For instance, in biomedical applications, where measuring each covariate may be costly, a doctor would want to be confident that each test in a panel is actually relevant to predicting the response.

It was also pointed out that StrongStop was much more conservative than ForwardStop in the simulation experiments. This makes sense given that StrongStop also controls the FWER in the ordered setting.

References

[1] 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) (1995): 289-300.

[2] Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. *False Discovery Rate Control for Sequential Selection Procedures, with Application to the Lasso.* arXiv preprint arXiv:1309.5352 (2013).

[3] Max Grazier G'Sell, Jonathan Taylor, and Robert Tibshirani. *Adaptive testing for the graphical lasso.* arXiv preprint arXiv:1307.4765 (2013).

[4] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. *A Significance Test for the Lasso.* The Annals of Statistics 42.2 (2014): 413-468.