STAT 27850 PROJECT

WENYU CHEN, MICHAEL WANG

Contents

1. Abstract 1
2. Introduction 1
3. Preliminaries: LBORD 2
4. Formalization of Clustering 2
5. Cluster-Adaptive LBORD 3
6. Adaptive Forgiving LBORD 6
7. Simulation Study 7
8. Comparison of the Two Methods 11
9. Discussion 12
10. Conclusion 12
11. Acknowledgement 12
References 12

1. Abstract

This paper develops two adaptions of the LBORD method to increase power when there are clusters of signals. The first method, the Cluster Adaptive LBORD, yields higher power when there are consecutive sequences of signals (clusters). We have given a proof that this method controls the FDR. The second method, the Adaptive Forgiving LBORD, is a more liberal approach and further increases power. Although the Adaptive Forgiving method is not strictly below the controlled FDR α, we have shown that its FDR is bounded and acceptable in practice. To put the two methods into test, we conducted simulation studies and showed that the two methods indeed have higher power than the original LBORD. While the Cluster Adaptive method strictly controls the FDR, the Adaptive Forgiving method also has reasonable FDR in simulations.

2. Introduction

In an online testing setup, signals may have patterns. When a sequence of consecutive hypotheses all comes from the signal (as opposed to the null hypothesis), we call it a cluster of true signals. For instance, when streaming consumer preference data for a certain product, an online shopping site may start receiving a consecutive set of low p-values indicating a significant preference for the product. It is likely that such a cluster of significant findings represent some specific relations between these consumers, which makes their p values more likely to be significant.

Date: March 16 2016.

Assume that we do not have knowledge of the hypothesis stream before the test. However, if we can detect the existence of clusters and make use of the fact that a sequence of p-value is clustered, we can get more power and less false discoveries.

Therefore, when dealing with a cluster of rejections, we should increase the current level of rejection to allow more rejections. And we should still control the overall False Discovery Rate(FDR). In general, this study considers two factors:

(1) Improvement of power
(2) Control of FDR

The FDR control would be presented in a proof, and the improvement of power would be demonstrated by simulations.

3. Preliminaries: LBORD

We use the LBORD (also called LORD in the original paper) method for online testing as the basic method to modify[1].

LBORD uses a decreasing convergence sequence $\{\beta\}_i$ such that $\sum_{i=1}^{n} \beta_i = \alpha$. The rejection level is $\alpha_i = \beta_{i-\tau_i}$ where τ_i is the index of the last rejection before i. Apply on each p_i, the p-value of the i-th hypothesis.

As shown in [1], this method has $FDR(t) \leq \alpha$ where $FDR(t)$ is defined to be the FDR up tp the t-th rejection.

4. Formalization of Clustering

We formalize the definition of clusters. For a sequence of hypothesis, there are: 1. the underlying signal/null pattern (the truth about the hypothesis) and 2. the sequence of rejection/acceptance decisions produced by some algorithm.

Remark 4.1. For the latter part of this report, "cluster" means "cluster of length larger than r" for some r that we specify.

Definition 4.2 (True cluster, rejection cluster, and 0/1, T/F representation)**.** A *True cluster* is a sequence of consecutive signals. A *rejection cluster* is a sequence of p-values being rejected.

Consider a rejection cluster. Use 0 to denote true null, 1 to denote true signal. Use F to denote rejection, and T to denote acceptance of p-value.

Remark 4.3. For a true cluster, the corresponding sequence of rejection/acceptance might not form a rejection cluster, and the length might differ. Also for a rejection cluster, the corresponding pattern might not have a true cluster. See examples 4.5

Definition 4.4 (Head, body, and tail of a rejection cluster)**.** Define the *Head* of a rejection cluster to be the first r of this cluster. Define the *Body* of a rejection cluster to be the sequence from $r+1$-th rejection (if it is a signal) to the last *consecutive true signal* in the rejection cluster. If $r+1$ is null, then there would be no body. Define the *Tail* of a rejection cluster to be the rest of the rejection cluster.

Notice that by definition, the body of the rejection cluster contains only true signals. This definition splits the rejection cluster into three parts: the head, where the recognition of cluster happens, the body, where all discoveries are true signals, and the tail. Head is all rejected by LBORD rejection levels β sequence; body lowers FDR and increases power; tail might be troublesome if we are too liberal.

Examples 4.5. Consider a true cluster and its corresponding rejection sequence. Let $r = 4$

```
111111110000
FFFFFFFFFFTT
```

Observe a true cluster of length 8 and a rejection sequence of length 10.

$$\underbrace{FFFF}_{Head}\underbrace{FFFF}_{Body}\underbrace{FF}_{Tail}$$

There are 2 false discoveries in the the rejection cluster. All in the tail.

Consider the same signal pattern with another rejection pattern.

```
111111110000
FFFTFFFTFFTT
```

There is no rejection cluster (larger than r). In this case the algorithm that produces the rejections fails to detect a cluster.

Consider another signal pattern and its corresponding rejection pattern.

```
110111010000
FFFFFFFFFTTT
```

There is no true cluster. But the algorithm detects a rejection cluster of length 9. Thus the rejection cluster has pattern

$$\underbrace{FFFF}_{Head}\underbrace{FF}_{Body}\underbrace{FFF}_{Tail}$$

There is 1 false discovery in the head, and 2 in tail.

These examples gives the most important three conditions we will see in the proof of the first method, the Cluster-Adaptive LBORD.

5. Cluster-Adaptive LBORD

5.1. **Definition.**

Definition 5.1 (Cluster-Adaptive LBORD)**.** Let τ_i be the last rejection before hypothesis i. Let c_{i-1} be the number of consecutive rejections up tp $i-1$. We apply a constant function as adaptive rejection boundary.

$$\alpha_i = \begin{cases} \beta_{i-\tau_i} & \text{for } c_{i-1} < r \\ m\alpha & \text{for } c_{i-1} \geq r \end{cases} \tag{5.2}$$

where $\{\beta\}$ is a converging decreasing infinite sequence summing to α, and r is a positive integer and $m\alpha < 1$.

Then update

$$c_i = \begin{cases} c_{i-1} + 1 & \text{if } p_i \text{ is rejected} \\ 0 & \text{if fails to reject.} \end{cases} \tag{5.3}$$

Intuitively, the parameter r represents *the amount of evidence needed to admit a sequence as a cluster*, and parameter m represents *how much more liberal would clustered hypothesis receive.*

5.2. **FDR control.**

Theorem 5.4. *The Cluster-Adaptive LBORD method controls FDR.* [1]

Proof. Define a rejection cluster to be "observed true cluster" if its *head* contains no true null (i.e., all hypothesis in head are true discoveries). Let k be the total number of true clusters, let $\tilde{k}$ be the number of "observed true clusters", let q be the total number of other rejection cluster (called "fake clusters"). The total number of clusters in rejection sequence is $\tilde{k} + q$

There are four conditions

Four conditions		
	Not true cluster	True cluster
Not rejection cluster	Case 1	Case 2
Rejection cluster	Case 3: Fake cluster	Case 4: Observed true cluster

Case 1

This is the case of non-cluster pattern handled by non-cluster rejection boundaries. In this case, the algorithm tests hypothesis with the β_i sequence. We can view each rejection cluster as a single rejection. This generalization is valid because β_i it reset to β_1 at the beginning of the cluster, all the way till the end of the cluster. Then this method works exactly the same as LBORD method. So we have

$$FDR_1 \leq \alpha \tag{5.5}$$

trivially true.

Case 2

This is the case of true clusters being ignored. This case is exactly the same as Case 1 from the rejection sequence point of view. Though power is not as high as when the cluster is detected, the FDR is not affected. We have FDR control

$$FDR_2 \leq \alpha \tag{5.6}$$

Case 3 & 4

Denote the length of head, body, tail of each rejection cluster as h, b, t. Recall that $h = r$ by definition. We know $r + b + t \geq r$. Thus the total rejection in Case 4 is

$$D_4 = \sum_{j=1}^{\tilde{k}} (r_j + b_j + t_j) \geq r\tilde{k} \tag{5.7}$$

Similarly, for Case 3

$$D_3 = \sum_{j=1}^{q} (r_j + b_j + t_j) \geq rq \tag{5.8}$$

In these two cases, tests are using the β_i sequence in the head of the cluster, and using adaptive rejection bound $m\alpha$ in the body and tail. Again, notice that body does not have any false positive.

[1] Here we use the same definition of FDR as in the LBORD paper, i.e., define it as FDR(t), the FDR up to t-th rejection.

In case 4, all p-values in head are signals. Thus in Case 4, false positives can only occur in the tail, we have

$$FD_4 \leq \sum_{j=1}^{\tilde{k}} t_j \tag{5.9}$$

$$E(FD_4|\tilde{k}) \leq \tilde{k}E(t_j) = \tilde{k}\sum_{j=1}^{\infty} j(m\alpha)^j = \frac{m\alpha}{(1-m\alpha)^2}\tilde{k} \tag{5.10}$$

The first inequality is because not all rejections in tail are false positive (see example 3).

In Case 3, what happens in the tail is the same. However, there are also false positives in the head. Consider the head of a "false cluster". Given that the r-many p-values in head are all rejected, the probability of each being false negative is bounded by the condition of independence. Let π_0 be the probability of each hypothesis to be null.[2] Then $\pi_0\beta_1$ is the probability of i being null and getting rejected; $(1-\pi_0)\tau$ is the probability of i being a signal and also getting rejected in the head, where τ is the power of each single test, we know $\tau \leq 1$.

$$FD_3 \leq \sum_{j=1}^{q} t_j + \sum_{j=1}^{q} \text{false positive in head of } j \tag{5.11}$$

$$E(FD_3|q) \leq \frac{m\alpha}{(1-m\alpha)^2}q + q\sum_{j=1}^{r}\binom{r}{j} j(\pi_0\beta_1)^j((1-\pi_0)\tau)^{r-j} \tag{5.12}$$

Notice that the second part is increasing in τ, thus let $\tau = 1$ and then we can calculate the second part (which converges) bound FD_3 with

$$E(FD_3|q) \leq \frac{m\alpha}{(1-m\alpha)^2}q + q\left(\beta_1\pi_0 r((\beta_1-1)\pi_0+1)^{r-1}\right) \tag{5.13}$$

Denote $T = \dfrac{m\alpha}{(1-m\alpha)^2}$. Now we write the bound for FDR of the two cases

$$FDR_{3\&4} \leq E\left[\frac{FD_3 + FD_4}{D_3 + D_4}\right] \tag{5.14}$$

$$\leq E\left[\frac{T(\tilde{k}+q) + q\left(\beta_1\pi_0 r((\beta_1-1)\pi_0+1)^{r-1}\right)}{r(\tilde{k}+q)} \mid \tilde{k}, q\right] \tag{5.15}$$

$$\leq \frac{T}{r} + \beta_1\pi_0((\beta_1-1)\pi_0+1)^{r-1} \tag{5.16}$$

The second term is decreasing in r for $r \geq 1$. Thus it is bounded by $\beta_1\pi_0$. Since $\pi_0 \leq 1$,

$$FDR_{3\&4} \leq \frac{T}{r} + \beta_1 \tag{5.17}$$

Now we require that

$$\frac{T}{r} + \beta_1 \leq \alpha \tag{5.18}$$

$$\Leftrightarrow r \geq \frac{1}{\alpha-\beta_1}\frac{m\alpha}{(1-m\alpha)^2} \tag{5.19}$$

[2]We assume the non-clustered nulls and signals are each i.i.d. distributed

The term $\frac{1}{\alpha-\beta_1}$ is fixed by our choice of β sequence. If we pick r and m such that $r \geq \frac{1}{\alpha-\beta_1}\frac{m\alpha}{(1-m\alpha)^2}$, then put the result into (5.17), we get

$$FDR_{3\&4} \leq \alpha \tag{5.20}$$

We showed that in all 4 cases, FDR is controlled by the adaptive method if (5.19) is met. Thus we proved FDR is controlled in total by linearity of expected value. □

6. Adaptive Forgiving LBORD

Definition 6.1 (Adaptive Forgiving LBORD)**.** Let τ_i and c_{i-1} be as defined above. For each hypothesis i, use the rejection boundary

$$\alpha_i = \beta_{i-\tau_i} \tag{6.2}$$

If the p-value is rejected, go to the next hypothesis, and update

$$c_i = c_{i-1} + 1. \tag{6.3}$$

If p-value is not rejected. Test it again using

$$\zeta_i = \frac{1}{2}\alpha c_{i-1} \tag{6.4}$$

And update

$$c_i = c_{i-1} - \frac{1}{\alpha}. \tag{6.5}$$

otherwise, set $c_i = 0$

Remark 6.6. The sequence of rejection produced by this algorithm divides into a set that is rejected by the first boundary (by β sequence) and a set rejected by the second boundary (by ζ sequence).

Remark 6.7. We do not know if this method controls FDR. But we can show it is not very far off.

The next step is similar to what we did for the last method. Partition the rejection sequence into two cases.

Case 1: The part of rejection sequence that is not in rejection clusters of length ≥ 2

Notice all of the elements in this set are rejected or accepted by the first comparison (i.e., with the β sequence). Again, we view each rejection cluster as one rejection, because they do not affect the distance from the last rejection for p-values after the cluster. In this way, the method is functioning the same as LBORD does. Thus the FDR is controlled.

Case 2: The collection of rejection clusters of length ≥ 2

For a cluster (fake cluster or observed true cluster) of length k, we can write its length into

$$k = L + \lfloor \alpha L \rfloor \tag{6.8}$$

where L is the number of elements in this cluster that is rejected by the first comparison (β sequence), and $\lfloor \alpha L \rfloor$ is the times the second comparison (ζ sequence) is used. Within each cluster, the FDR is controlled by

$$FDR \leq E\left[\frac{\#\text{false discoveries by 1st comparison} + \#\text{2nd comparison}}{L}\right] \tag{6.9}$$

Bound the number of false positive in the first term by the same argument we did in the last proof. We have the first term less than $\beta_1 L$, because all the hypothesis (except the first) is rejected at level β_1.

If $\alpha L > \frac{2}{\alpha}$, then there may be at most $L - \frac{1}{2\alpha}$ second comparisons that is forcefully rejected by $\zeta > 1$. In this case

$$FDR \leq E\left[\frac{\beta_1 L + \sum_{i=1}^{1/2\alpha} \frac{i}{2}\alpha + (\lfloor \alpha L \rfloor - \lfloor \frac{2}{\alpha} \rfloor)}{L} | L\right] \tag{6.10}$$

$$\leq E\left[\frac{\beta_1 L + 1 + \frac{1}{16\alpha} + \lfloor \alpha L \rfloor - \lfloor \frac{2}{\alpha} \rfloor}{L} | L\right] \tag{6.11}$$

$$\lessapprox \frac{\alpha L + (\beta_1 L - \frac{7}{16\alpha})}{L} \tag{6.12}$$

$$\leq \frac{\alpha L + \beta_1 L}{L} \tag{6.13}$$

$$\leq 2\alpha \tag{6.14}$$

For $\alpha L \leq \frac{2}{\alpha}$,.

$$FDR \leq E\left[\frac{\beta_1 L + \sum_{i=1}^{\alpha L} \frac{i}{2}\alpha}{L} | L\right] \tag{6.15}$$

$$\approx \frac{\beta_1 L + \frac{\alpha^2}{4} L + \frac{\alpha^3}{4} L^2}{L} \tag{6.16}$$

$$= \frac{\beta_1 L + \frac{3}{4}\alpha^2 L}{L} \tag{6.17}$$

$$< \alpha \tag{6.18}$$

All the approximation above works when α is small. Thus even though we do not know if the method bounds FDR, we know it is not totally out of control. Also, in the condition of $\alpha L > \frac{2}{\alpha}$, the cluster is big enough that we would expect the actual FDR to be much smaller than the bound (now the bound assumes all second comparison are false discoveries), because it would be more likely that the cluster consists of all signals. When the rejection cluster is not large, this method controls FDR.

Later, in the simulation study, we would show that this method works better than the 2α bound.

7. Simulation Study

In our simulations we investigate the effects of different parameters on the power and FDR under the following 5 data patterns:

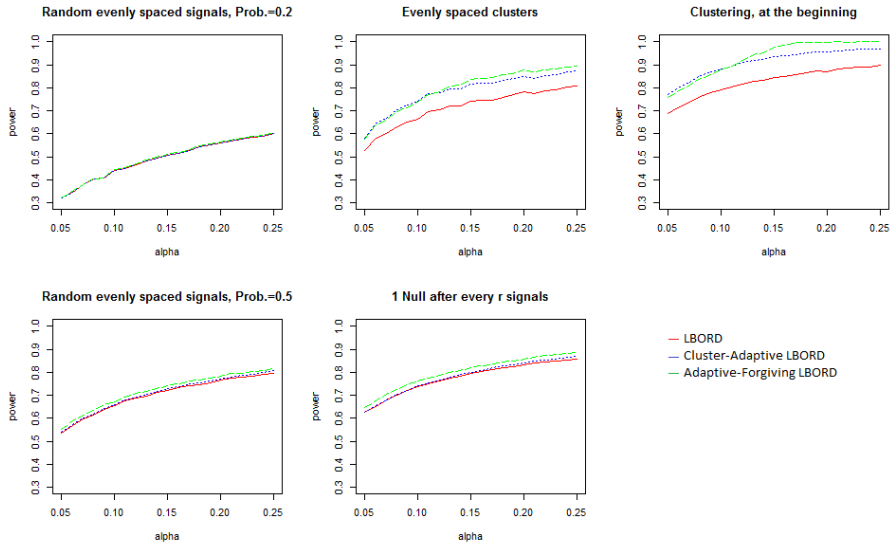
- Pattern 1: **Random evenly spaced signals**. Every hypothesis has a 0.2 probability of being a signal. The proportion of true signals is therefore 0.2.
- Pattern 2: **Random evenly spaced signals**. Every hypothesis has a 0.5 probability of being a signal. The proportion of true signals is therefore 0.5.
- Pattern 3: **Evenly spaced clusters**. There is a cluster of signals for every 500 hypotheses. The proportion of true signals is 0.2.

- Pattern 4: **Clustering at the beginning**. All true signals are beginning, followed by all nulls. The proportion of true signals is 0.2.
- Pattern 5: **1 null after every r signals**. To test the robustness of Cluster-Adaptive LBORD, we get a null whenever there are consecutive r signals. True signal proportion is $\frac{r}{r+1}$

Effects of α on FDR and Power

We first examine the effects of different cutoff values α on the FDR and power of the 5 patterns. We generate different values of α from 0.05 to 0.25. For each value of α, we run LBORD, cluster-adaptive LBORD, and adaptive-forgiving LBORD each 100 times and take the average of their false discovery proportion and power values. Then we plot the alpha values against the FDR and power values for each of the 5 patterns.

FIGURE 1. alpha values and power

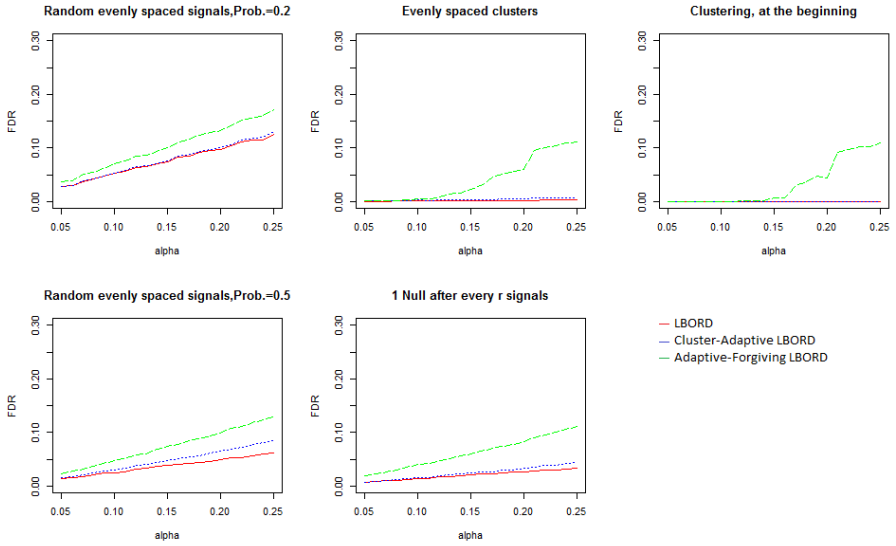


In Figure 1, we can see that when there are clusters present (evenly spaced and at the beginning), both the cluster-adaptive LBORD and the adaptive-forgiving LBORD have better power than the original LBORD. The adaptive-forgiving LBORD, being more liberal, has higher power than the cluster-adaptive LBORD.

In the cases where the signals are random with 0.2 and 0.5 probabilities, the 3 methods have roughly the same power, because there are no significant clusters present.

In the case where there is 1 null after r signals, the cluster-adaptive LBORD and LBORD have similar power, but the adaptive-forgiving LBORD has higher power. This is because when there is a null after r signals, the cluster-adaptive LBORD does not find a cluster and behaves the same as LBORD. However, the adaptive-forgiving LBORD can still accumulate the r signals and have a higher threshold for the next null.

FIGURE 2. alpha values and FDR

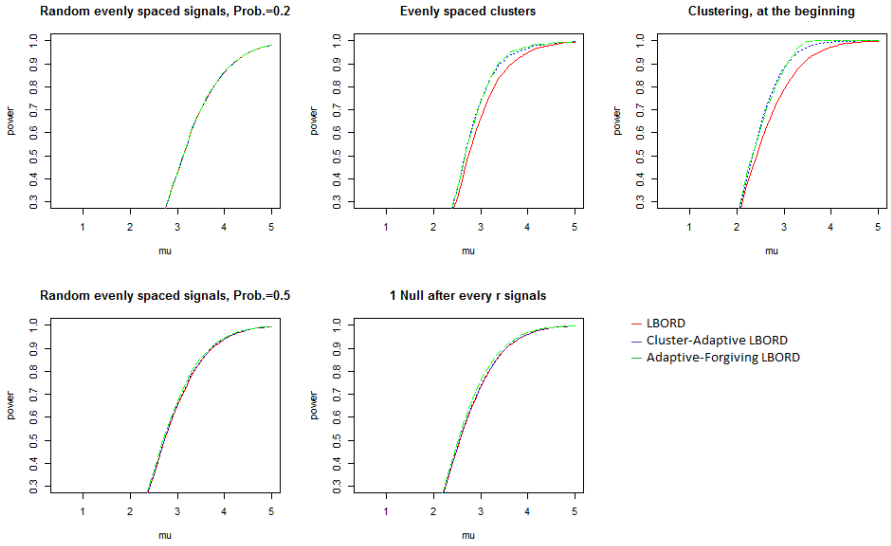


In Figure 2, the FDR values of LBORD and cluster-adaptive LBORD are very close. The adaptive-forgiving method has the highest FDR, since it has the most liberal threshold when dealing with clusters. Even though the adaptive-forgiving method's FDR is noticeably higher than that of the other two methods, it is still well below the α threshold. In the 5 plots above, the highest FDR is around 0.15, which is still much lower than the corresponding α value 0.25.

Effects of μ on FDR and Power

Next, we inspect the effects of signal strength μ on the power and FDR on the 3 methods. We generate μ values between 0.5 and 5. For each μ, we run LBORD, cluster-adaptive LBORD, and adaptive-forgiving LBORD 100 times and take the average values of false discovery proportion and power. Then the alpha values are plotted against the FDR and power values for the 5 patterns.

FIGURE 3. μ values and power



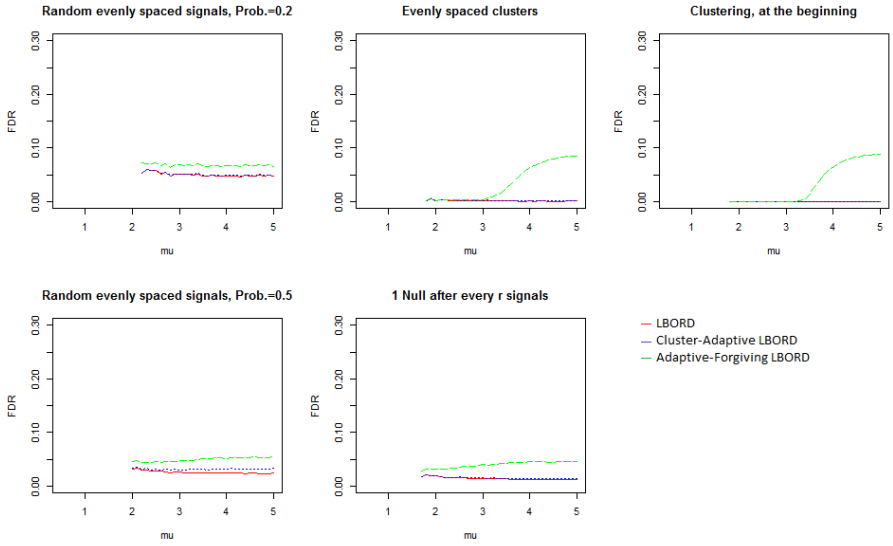
In Figure 3 [3], in the patterns where either the signals are randomly distributed or there is a null every r signals, there are no underlining true clusters and the 3 methods basically have the same power throughout different values of μ.

When there are underlining clusters (evenly spaced or at the beginning), both the cluster-adaptive LBORD and the adaptive-forgiving LBORD have higher power than the original LBORD. The two methods have similar power.

In Figure 4 below, the original LBORD and the cluster-adaptive LBORD have very similar FDR which is lower than $\alpha = 0.1$ for all values of μ. The adaptive-forgiving method has significantly higher FDR when there are clusters present. In the evenly spaced and the beginning placed clusters patterns, the FDR of the forgiving methods, represented by the green line, approaches $\alpha = 0.1$ for large values of μ. This is because when μ is large, most of the signals are detected, and the false discoveries are made at the end of each cluster, which are all false positives. The number these false positives is αL, so we have the overall FDR= $\frac{\alpha L}{L} = \alpha = 0.1$. In this case, the feature of forgiving, which would preserve the length of long clusters when signal is not extremely strong, does not increase power.

[3]The power of the three methods are almost the same when signal is weak. So our graph starts with power = 0.3

FIGURE 4. μ values and FDR



8. Comparison of the Two Methods

Cluster-adaptive LBORD controls FDR, and performs reliably under any condition that LBORD does. We have tested the worst case for cluster-adaptive LBORD, i.e., the case that there would be no "body" part for any rejection cluster, and the false clusters outnumbered observed true clusters. This method gives more power than the unflavored LBORD. It would not detect more clusters than LBORD, but would preserve long clusters which LBORD might recognize as several separate clusters.

The adaptive-forgiving method does not have exact FDR control. Another problem is in some cases, even though it controls FDR, it might not be preferred. For example, when signals are very strong, as shown above. The main strength of this method is that it often "forgives" very large p-values in the middle of the cluster, thus preserving the cluster as a whole. However, when signal is very strong, there is no such need to forgive, because all p-values are small already. In this case adaptive-forgiving method still controls FDR, and has power almost equal to 1, but makes α proportion of false discoveries.

The reason the adaptive-forgiving method might be useful, however, is that it preserves length of cluster. In the case that all 1000 signals are clustered together in one cluster, as we test above, LBORD usually divides the cluster into 80-100 smaller clusters, while cluster-adaptive gives 30-50, and adaptive-forgiving usually gives no more than 4. For most of the time, it would recognize the 1000 signals as in a single cluster. If, in some cases, we want an accurate detection of signals as well as the number of signal clusters, adaptive-forgiving method FDR would be a very good method, for it preserves the length of clusters in an often "unconditional" fashion.

In the general case where we may expect some cluster to exist, we can apply cluster-adaptive LBORD.

9. Discussion

Some future study can be done in the following directions. First, the adaptive-forgiving method can be improved if we find a way to deal with the stacked "forgiving"s in the end of huge clusters. For example, if we add the rule that *the chances to "forgive" will be discarded if they are used consecutively twice.* This is saying that we would tolerate extreme p-values from signal in the middle of clusters, which seem like nulls, but would not tolerate if they happens twice consecutively. This limits the false positives in the very end of the tail to be bounded by 2, instead of αL, as in the naive version.

A better way of choosing r and m might be applied in the cluster-adaptive method. Now the method to pick this two variables is based on intuition about the data, the FDR-control criterion, and conventions that $r < 10$ and $m\alpha < 0.5$. There might be ways to adaptively choose r and m, based on the length of clusters seen, and signal strength estimated. However, the method now works better under mixed data pattern (where non-above is fixed).

The most important improvement to this study would be testing on real data. We do not have access to longer online-testing data, which limits our ability to test the empirical behavior of the two methods. Also we would like the methods to be tested in larger data sizes.

10. Conclusion

In this paper, we have provided two adaptations of the LBORD method to have greater power when there are clusters of signals in online testing. First, the Cluster Adaptive method and second, the Adaptive Forgiving method. We have given a proof that the first method controls the FDR under the desired level α. We also showed that though the second method does not control the FDR strictly, its FDR is still reasonably bounded. Both methods should have greater power then LBORD when there are clusters of signals present.

Our simulations further support our theory. Running all 3 methods under 5 different data patterns with different parameters, we have found that both the Cluster Adaptive method and the Adaptive Forgiving method have higher power than LBORD when there are signal clusters. The FDR of the Cluster Adaptive method is strictly controlled and the FDR of the Adaptive Forgiving method also behaves under control. Therefore, these two methods can be useful for gaining more power when clusters are prevalent in online testing settings.

11. Acknowledgement

We thank Professor Rina Barber for her great support and insight. We also thank Yuancheng Zhu for his helpful comments and feedback which greatly assisted our progress.

References

[1] Javanmard, A., and Montanari, A. On online control of false discovery rate, 2015.