Outline

August 8, 2011

1 Main Result

1.1 Ideal Procedure

In this section, we restrict our discussion to the one-sided testing. Similar conclusions hold for the two-sided testing with appropriate adjustment.

Since the introduction of False Discovery Rate in Benjamini and Hochberg [1995], there are quite a large amount of literature in talking about various procedure controlling the FDR, and various forms of error rate, such as the marginal FDR, FNR, etc. Genovese and Wasserman [2002] has shown that asymptotically $mFDR = mFNR + O(1/m)$. Assuming that the density of the p-value is concave under the alternative hypothesis, they have shown that asymptotically the cutoff of BH's procedure is equivalent to choose a maximum u such that

$$u/G(u) \leq q, \tag{1.1}$$

where $G(u)$ is the CDF of the p-values.

Let $f_0(x)$ be the density of X under the null hypothesis H_0 and $f_1(x)$ be the density of X under the alternative. The choice of u in (1.1) is equivalent to choose a cutoff $T_q(F)$ where

$$T_q(F) = argmin_t\{\int_t^{+\infty} f_0(x)dx / \int_t^{+\infty} dF(x) \leq q.\}$$

Here $F(x) = (1-\epsilon)F_0(x) + \epsilon F_1(x)$ and $dF(x) = (1-\epsilon)f_0(x) + \epsilon f_1(x)$. Thereafter, we reject a hypothesis H_i if the corresponding observation X_i is greater than or equal to $T_q(F)$. Equivalent, we reject H_i is $X_i \in I = [T_q(F), +\infty)$. We call such an interval I the ideal BH interval. Note that the ideal BH interval would reject all the observations that exceed a

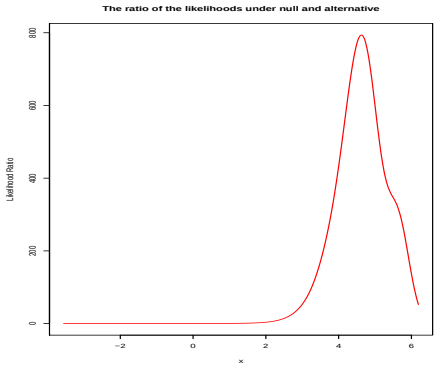

Figure 1: Plot of the likelihood ratio for the Golden Spike-in dataset.

certain threshold. This rejection rule would be problematic in both theory and applications.

Consider the Golden Spike-in data set of Choe et al. [2005], which will be revisited in Section 3. After estimating the empirical null (Cai and Jin [2010] and Jin [2008]) and the density function $f_1(x)$ by using the kernel method, we plot the estimated likelihood ratio $h(x) = \frac{\hat{f}_1(x)}{\hat{f}_0(x)}$ in Figure 1. It is clearly seen that the likelihood function is not monotone increasing. Consequently, the extremely large observations are very likely from the null distribution and they don't deserve more chance to be rejected. Therefore, we can not rely on the BH ideal rejection interval I in deciding when to reject/accept a null hypothesis.

This example motivates us to generalize the BH's procedure as selecting an ideal rejection region $R_q(F)$ as

$$R_q(F) = \text{argmax}_R \left\{ \int 1_R dF : q \int 1_R dF \geq \int 1_R d\Phi \right\}. \tag{1.2}$$

The definition $R_q(F)$ in (1.2) is a distribution free procedure. When a reliable information of ϵ is available, one can incorporate such information and derive a less conservative procedure as

$$R_q^*(F) = \text{argmax}_R \left\{ \int 1_R dF : q \int 1_R dF \geq (1-\epsilon) \int 1_R f_0 \right\}. \tag{1.3}$$

The region $R_q^*(F)$ will maximize the average number of rejection while controlling the mFDR at q-level exactly. Later in this article, we will only consider the distribution free procedure $R_q(F)$ which indeed controls the mFDR at the $q(1-\epsilon)$ level. If there is reliable

information of ϵ, one can simply replace q by $\frac{q}{1-\epsilon}$ to control the mFDR at q-level.

The criteria as maximizing the average rejection while controlling the mFDR relates to the Neyman-Pearson Lemma. Defined as in (1.3), $R_q(F)$ is determined by the likelihood ratio function $h(x) = \frac{f_1(x)}{f_0(x)}$, or equivalently, the local fdr.

Theorem 1 *For a given q, if there exits a $c(q)$ such that the set $S = \{x : \frac{f_1(x)}{f_0(x)} \geq c(q)\}$ satisfies*

$$\int_S f_0(x) = q \int_S dF(x).$$

Then for any other set T such that $\int_T f_0(x) \leq q \int_T dF(x)$,

$$\int_S dF(x) \geq \int_T dF(x).$$

Theorem 1 shows that among all the region which offers the control of mFDR at a certain level, the one determined by cutting the likelihood ratio at some threshold $c(q)$ maximizes the average number of rejections.

Paralleling to the definition of (1.3), one can define the ideal rejection interval $I_q(F)$ as

$$I_q(F) = \text{argmax}_{I_{\{t,s\}}} \left\{ \int 1_{\{I_{\{t,s\}}\}} dF : q \int 1_{\{I_{\{t,s\}}\}} dF \geq \int 1_{\{I_{\{t,s\}}\}} f_0 \right\}. \tag{1.4}$$

We can easily prove the following corollary and the proof is omitted.

Corollary 1.1 *(1) When $h(x)$ has a uni-mode, the interval based on (1.4) is optimal; when $h(x) \to 0$ as $x \to \infty$, the ideal BH interval is not optimal;*

(2) When $h(x)$ is monotone increasing, then the rejection region (1.3), the ideal interval (1.4), and the ideal BH interval are the same.

The non-monotonicity of the likelihood ratio is common in applications. As in the Golden Spike-in data set we mentioned earlier, the estimated likelihood ratio function behaves strongly as a unimodal function. The non-monotonicity is also prevalent in the common distributions we used. For instance, consider a location family where $f_1(x) = \phi(x - \mu)$ where μ is a location parameter and $f_0(x) \in C^1(R)$. Then $h(x)$ is monotone increasing for any $\mu > 0$ if and if $\log(f_0(x))$ is a concave function. For many cases such

as $f_0(x)$ being the density of a T distribution, Cauchy distribution, and many others, the assumption of monotonicity of $h(x)$ does not hold.

When putting the scale parameter into consideration, the monotonicity condition is even easier to be violated. For the very simple case where $\phi(x)$ is the density function of standard normal distribution, and $f_1(x) \sim N(\mu, \sigma^2)$ where $\sigma < 1$. Then the function $h(x)$ is no longer monotone increasing (See Figure 2). It will be problematic if ignoring this issue and simply use the ideal BH interval.

To better illustrate the problem, we consider the following simulation study. Let $p = 100,000$ and $\epsilon = p^{-\beta}$ be the proportion of non-null hypothesis. Assume that $f_0(x) = \phi(x)$ and $f_1(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$. For various choice of (β, μ, σ), we randomly generate the sequence $X_1, X_2, \cdots, X_n$ and order them decreasing as $X_{(1)} \geq X_{(2)} \geq \cdots X_{(p)}$. We then record the position of the very first observation in the ordered sequence which is generated from the alternative hypothesis. We replicate this step 100 times and calculate the average positions and report this number in the third column of Table 1. For instance, when $\beta = 0.6$ and there are $p\epsilon = 100$ non-null hypothesis. Setting $\mu = 1.5$ and $\sigma = 0.8$, on average, the first largest observation generating from the alternative hypothesis appears at the 47.12-th location. In other words, the first 46 largest observations are generated from the null hypothesis. Using BH ideal rejection interval $I = [T_q(F), +\infty)$ results in at least 46 false rejection on the right tail if $T_q(F) \leq X_{(47)}$. It is obviously better to choose an rejection interval which ends at a finite number, hopefully at $X_{(477)}$. Consequently, it is advantageous to consider a rejection interval without fixing the right ending at $+\infty$.

Theorem 1 indicates that the optimal set is given by cutting the likelihood ratio. Most existing literature talks about how to identify such set and study its property. There is little discussion about the existence of such an interval. Recently, Zhang et al. [2011] brought out an phenomenon called "lack of identification" which describes situations when there is no non-trivial procedures can control the FDR at a certain level. In what follows in this section, we will show conditions under which an optimal region/interval does/doesn't exist.

Theorem 2 *Let $h(x) = \frac{f_1(x)}{f_0(x)}$ be the function of likelihood ratio.*
(1). If $\max_x h(x) < \frac{1}{q'}$, where $q' = \frac{q\epsilon}{1-q(1-\epsilon)}$, then there exists no non-trivial interval I such

β	σ	μ	Position
0.7	0.8	2.0	38.77
0.7	0.5	2.5	32.5
0.6	0.8	1.5	47.12

Table 1: Assume that $F(x) = (1-\epsilon)\phi(x) + \epsilon\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$ where $\epsilon = p^{-\beta}$ and $p = 100,000$. For each parameters setting, we generate the random sample and sort the data. This table report the average number of the location of the largest observation generated from the alternative hypothesis after ordering the data decreasingly. In all the settings, the numbers are non-zero, indicating that including the very tail as the ideal BH interval suggests leads to many false discoveries. We replicate the simulation 100 times for each setting.

that

$$\int_I f_0(x)dx \le q\int_I dF(x).$$

(2). If $h(x)$ has a unimodal and $h(x) \to 0$ as $x \to \pm\infty$. When $\max_x h(x) > 1/q'$, let c_1 and c_2 be the solutions of $h(x) = \frac{1}{q'}$, then $[c_1, c_2] \subset I_q(F)$.

Since q' is increasing with respect to q, smaller q requires a larger mode of the likelihood ratio $h(x)$. Theorem 2 indicates that a sufficiently large mode of $h(x)$ is necessary for the existence of the ideal interval I. When this condition doesn't hold, there exists none but trivial rejection interval which controls the mFDR at the pre-specified level q. Any FDR controlling procedures fail in a way that it either has no power or fails to control the FDR.

When $h(x)$ is monotone increasing, intuitively, one would conject that the mFDR will be small, saying smaller than q for any pre-specified q-level, when the left endpoint of BH interval is suffcient large. However, if $\max_x h(x) < \frac{1}{q'}$, the only procedure that can control the mFDR at q-level is the trivial procedure, i.e. the one that will accept all the null hypothesis. In other words, no matter how large the left endpoint of BH interval is, the mFDR level can never be controlled at the q-level. One of such examples is when $f_0 \sim T_d$ and $f_1 \sim T_d(\mu)$. The likelihood function is monotone increasing with an upper limit. Consequently, there is a lower limit of the FDR level that one can possibly control. When setting the FDR level to be smaller than this, all procedures fail.

1.2 Empirical Procedure

The ideal interval in (1.4) relies on the mixture CDF $F(x)$ which is unknown in application. We can thus replace $F(x)$ by the empirical CDF $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \le x)$ and define the

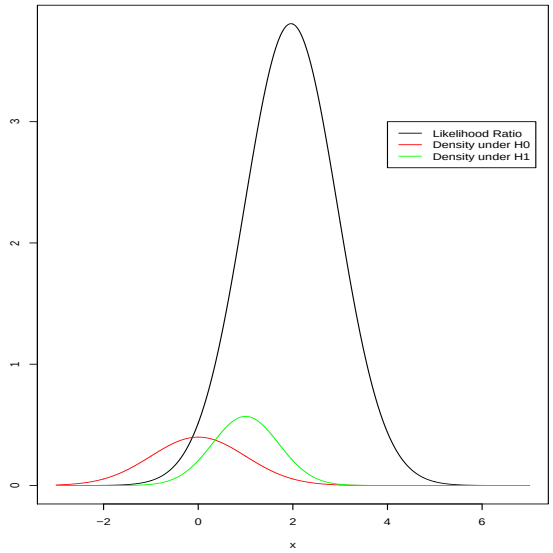

Figure 2: Let $\mu = 1$ and $\sigma = 0.7$. We plot the density function $f_0(x)$, $f_1(x)$, and the likelihood ratio $h(x) = \frac{f_1(x)}{f_0(x)}$. It is clearly seen that when $\sigma < 1$, the likelihood function $h(x)$ is not monotone. Ideally, one would like to reject the null hypothesis when x is within a finite interval.

empirical rejection interval I_n as

$$I_n = argmax_I\{\int_I dF_n, length(I) > 0.5, \int_I f_0(x)dx \leq q\int_I dF_n(x)\}, \tag{1.5}$$

As argued in Cai et al. [2007] (**Is This right?**), the interval I_n can not be too short and we put a restriction as $Length(I_n) > 0.5$. The choice of the constant 0.5 is not critical. Any value between 0.1 and 1 will work well.

Direct searching for an interval requires a $O(p^2)$ calculation which takes much time when p is large. The following algorithm reduces the computation complexity to $O(p \log p)$. The code is available at

Algorithm

Let $X_1, \cdots, X_n$ be the observations. Generate two sequences $F_n(X_i)$ and $F_0(X_i)$ based on the empirical CDF $F_n(x)$ and the CDF under the null $F_0(x)$, $i = 1, 2, \cdots, n$. Let $C_i = qF_n(X_i) - F_0(X_i)$. Then $C_i < C_j$ implies that $q\int_{X_i}^{X_j} dF_n \geq \int_{X_i}^{X_j} dF_0$.

1. Order C_i as $C_{(i)}$;

2. Set $I = 1$, $J = 1$, and $MAX = 0$. Let $i' = \arg_i\{C_i = C_{(1)}\}$;

3. For k in 2:n, let $j' = \arg_j\{C_j = C_{(k)}\}$.

   If $F_n(X_{j'}) - F_n(X_{i'}) > MAX$ and $|X_{j'} - X_{i'}| > 0.5$, set $MAX = F_n(X_{j'}) - F_n(X_{i'}), I = i', J = j'$;

   If $F_n(X_{j'}) < F_n(X_{i'})$, set $i' = j'$;

4. Then $[X_I, X_J]$ maximizes $F_n(X_j) - F_n(X_i)$ under the restriction that $q(F_n(X_j) - F_n(X_i)) \geq (F_0(X_j) - F_0(X_i))$. Reject X_k if $X \in [X_I, X_J]$.

1.3 Convergence rate of the generalized BH procedure

Our empirical approach is based on the empirical CDF F_n. DKW's theorem guarantees that $||F_n - F|| = O_p(n^{-1/2})$. Therefore, we would expect that the empirical interval mimics the ideal interval for large sample size n. Theorem 3 shows that the two ending points of the empirical rejection intervals converge to the ideal ones with the rate $n^{-1/4}$.

Before stating the theorem, we will give some necessary definitions. Let $S(a, b) = \int_a^b f_0 - q\int_a^b dF$, then according to the proof of Theorem 2,

$$\frac{\partial s}{\partial b}\begin{cases} > 0 & b < c_1 \\ < 0, & c_1 < b < c_2 \\ > 0 & b > c_2 \end{cases} \quad \text{and} \quad \frac{\partial s}{\partial a}\begin{cases} < 0 & a < c_1 \\ > 0, & c_1 < a < c_2 \\ < 0 & a > c_2 \end{cases}$$

where c_1 and c_2 are given in 3.

Assume that $f_0, f_1 \in C^1$. Let $b_a(F) = argmax_b\{b : s(a, b) \leq 0\}$. Define $g_n(a) = F_n(b_a(F_n)) - F_n(a)$ and $g(a) = F(b_a(F)) - F(a)$. Assume that the function $g(a)$ attains the maximum at $a = a_0$. Then according to Theorem 2, $a_0 < c_1$ and $f_0'(a_0) - qF'(a_0) > 0$, and $f_0(b_{a_0}(F)) - qF'(b_{a_0}(F)) > 0$. Further, there exists a neighborhood A of a_0 such that $f_0(x) - qF'(x) > 0\ \forall x \in b_A(F)$. Consequently, $b_a(F)$ has second derivative with respect to a within a neighborhood of a_0.

Theorem 3 *Assume that the likelihood ratio function $h(x)$ is unimodal and $h(x) \to 0$ as $x \to -\infty$ and ∞ and $\max_x h(x) > 1/q'$. Assume that $g_n(a)$ attains the maximum at $a = a_n$.*

Further,

$$f_1'(b_{a_0}(F))(b_{a_0}'(F))^2 + f_0(b_{a_0}(F))b_{a_0}''(F) - f_1'(a_0) \neq 0. \tag{1.6}$$

Then

$$|a_n - a_0| = O(n^{-1/4}), \textit{and} \quad |b_{a_n}(F_n) - b_{a_0}(F)| = O(n^{-1/4}). \tag{1.7}$$

The proof, given in the appendix, relies heavily on the modulus of continuity and the DKW's theorem which guarantees the uniform control of the empirical CDF.

1.4 Compare to locfdr

Some alternative approaches in searching for the rejection interval are based on the likelihood function $h(x) = \frac{f_1(x)}{f_0(x)}$ or the local fdr (See Efron [2005, 2007, 2008, 2010], Sun and Cai [2007].) Giving the exact information of the density, the locfdr based approach chooses the rejection region in a way to cut the likelihood ratio function at some threshold. Theorem 1 indicates that these methods lead to the optimal rejection interval ideally. However, to get a reliable interval, the curve of $h(x)$ at the ending point can not be too flat. In other words, assume that $[a, b]$ is the rejection interval and $h'(b)$ (or $h'(a)$) is close to zero, a small change of $h(x)$ (due to the estimation error for the density, for instance) might result in a dramatic change of the rejection interval. In Sections 2 and 3, it can be seen that the locfdr based approach might have a extremely unexpected large number of (false) rejection.

Let $\phi(x)$ be the density function of the standard normal distribution and assume the mixture model as

$$G(x) = (1 - \epsilon)\phi(x) + \epsilon g(x). \tag{1.8}$$

Under the alternative, the observation is a combination of two truncated normal. To be specific, generate Y as $N(\mu, \sigma)$ first. We pertain those observations that are greater than a_0. For the rest, we replace them by an observation generated from the standard normal distribution truncated at a_0. Consequently, the density function $g(x)$ can be written as

$$g(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu^2)}{2\sigma^2}) & if x > a_0 \\ \frac{\Phi_{\mu,\sigma^2}(a_0)}{\Phi(a_0)} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) & if x < a_0 \end{cases}$$

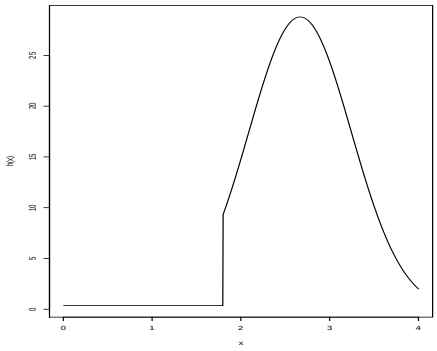

Figure 3: The plot of the likelihood ratio function.

Consequently, the likelihood ratio $h(x)$ is

$$h(x) = \begin{cases} \frac{1}{\sigma}\exp(\frac{x^2}{2} - \frac{(x-\mu)^2}{2\sigma^2}) & if x > a_0 \\ \frac{\Phi_{\mu,\sigma^2}(a_0)}{\Phi(a_0)} & if x < a_0 \end{cases}$$

As shown in Figure 3, the likelihood ratio is flat up to a_0 and discontinuous at a_0. Similar phenomenon can be observed when the function $h(x)$ on $(-\infty, a_0]$ is nearly flat. We use this extreme example just for the reason of convenience.

Let r_R be the solution of $h(x) = \frac{\Phi_{\mu,\sigma^2}(a_0)}{\Phi(a_0)}$ which is greater than a_0. Then for any $q > \frac{(1-\epsilon)(\Phi(r_R)-\Phi(a_0))}{(1-\epsilon)(\Phi(r_R)-\Phi(a_0))+\epsilon\int_{a_0}^{r_R} g(x)dx}$, a locfdr based approach will choose the rejection region as $[a_0, +\infty)$. For a smaller q, the ideal rejection region becomes $[a_1, r_R]$ where $a_1 < a_0$. When decreasing the level q, there exists a sudden change when determining the rejection region especially at the left ending point of the rejection interval by cutting the function $h(x)$ horizontally. It is problematic for the locfdr even for this ideal case. In real application, this will be exaggerated due to the error of the density estimation. This can be explained clearly in the following simulation.

Setting $p = 10,000$, $\epsilon = 0.1$, $(\mu, \sigma, a_0, r_L) = (2, 0.5, 1.8, 1.25)$. Then $r_R = 3.46$ and the $q = 0.583$. After generating the random numbers, the adjusted BH procedure and the locfdr procedure are applied in searching for significance observations. We calculate the left ending point of the rejection intervals and the simulated False Discovery Proportion among 100 replications. In Figure 4, we have plotted the histogram of these two quantities. Top two figures correspond to the left ending point of the rejection interval. When there is no

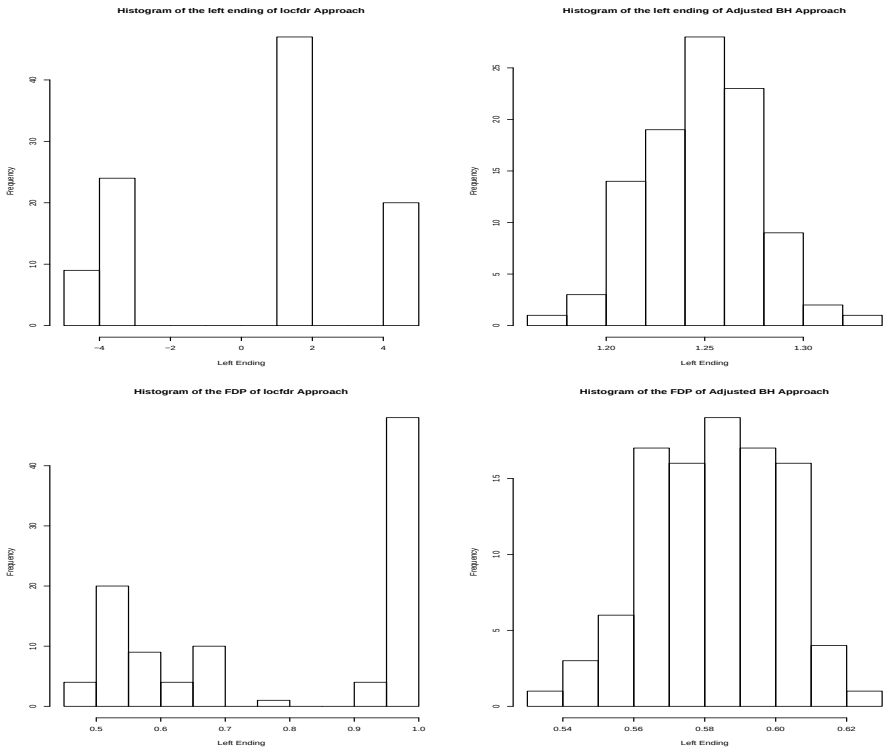

Figure 4: Histogram of the simulated left ending of the rejection interval and the FDP.

rejection, we set the left ending of the rejection interval as 5. It turns out that the locfdr approach is very unstable among these 100 replications. With a high probability, the left ending of the rejection interval is significantly different than the oracle value $r_L = 1.25$. On the other hand, the empirical left endings based on the adjusted BH procedure center around the oracle value closely. Intuitively, the proposed procedure care more about the area under the curve and thus provides more reliable result even though the curves at the ending points are flat.

The bottom two figures are the histograms of the simulated FDP. Once again, the simulated FDP of the adjusted BH procedure center around 0.583, the pre-specified level we are aiming at. The locfdr approach did a very poor job.

Furthermore, the fdr based approaches rely on the estimation of the density which is a more difficult problem than the estimation of the CDF. DKW's theorem guarantees the controlling of Kolmogorov-Smirnov norm for the empirical cdf. This is the key to the arguments of the rate of convergence. However, to the best of the authors' knowledge, there is no such result regarding to the density estimation.

In summary, when compared with the locfdr, our approach converge to the ideal interval under weaker regularity conditions. Consequently, we strongly recommend our approach in both theory and applications.

1.5 Multi-mode

The previous discussion, aiming at searching for the best interval, converges to the oracle rejection region when the oracle region is an interval. In a more general case, when the likelihood function is indeed multimodal, and the oracle region consists of several intervals, our approach can be easily generalized to deal with this.

For instance, if the likelihood ratio consists of two modes. We can firstly apply the algorithm in Section 1.2 to search for an interval I_1. Next, we mask all the observation within the interval I_1 and search for another interval I_2. In other words, in the step 3 of the algorithm, we update the subscripts when imposing an additional condition that $X_{i'}$ and $X_{j'}$ are not in I_1. Then $I = I_1 \cup I_2$ will be our final rejection interval.

This algorithm also have the computation costs $O(p \log p)$. It can be easily generalized to multi-modes. However, there is little practical benefit in searching for a rejection region as a union of more than two intervals.

1.6 Dependence

Previously, we have assumed the independence among all observations X_i's and shown that the proposed procedure will indeed converge to the ideal optimal rejection region under certain regularity conditions. By using the modulus of continuity, we can indeed prove the convergence rate of the ending point of the empirical rejection region. In this section, we will extend such a result to the case when assuming certain dependence among the observations.

We will use the α-mixing to describe the dependence structure. The proof of the convergence rate relies on the convergence rate of the empirical CDF which is stated in the following theorem. For the simplicity, we assume that $f_0(x) = 1, 0 \leq x \leq 1$ and the support of $f_1(x)$ is $[0, 1]$.

Theorem 4 *Assume that the α-mixing coefficient $\alpha(m)$ satisfies $\alpha(m) \asymp m^{-k}$ for some $k > 0$. Let F_n be the empirical CDF. Then*

$$P(\sup_x |F_n(x) - F(x)| < \epsilon) \to 1, where \epsilon = O(n^{-\frac{2k}{4k+7}}). \tag{1.9}$$

Based on Theorem 4, we can show that the convergence rate of the two ending points of the empirical rejection interval under this dependence structure is $O(n^{\frac{-k}{4k+7}})$.

2 Simulation

In this section, we will use the simulation to compare three approaches, BH's procedure, the procedure of Sun and Cai [2007] which is locfdr based, and the adjusted BH procedure. We are interested in three quantities, FDR, the average number of true positives and false positives.

Simulation Setting

Under H_0: $X \sim N(0,1)$, and under H_1: $X \sim g(x)$ where

$$g(x) \sim p_1 \epsilon N(\mu, \sigma^2) + (1 - p_1)\epsilon N(-\mu, \sigma^2).$$

Here, μ is according to Theorem 2, guaranteeing the existence of the ideal rejection interval. The number of signals is $n\epsilon$ with $\epsilon = p^{-\beta}$ with $0 \leq \beta \leq 1$. Among all the $p\epsilon$ non-null hypothesis, $100p_1\%$ of the signals are on the right tail and the rest on the left. After generating the sequence, we applied various procedures to the data and calculate the FDP, the number of true positives and false positives. We replicate the simulation 100 times and calculate the average of the three quantities and report them in Tables 2 and 3. The FDR level q we are aiming at controlling is 0.1, 0.3, 0.5, and 0.7.

The original procedure of Sun and Cai [2007] involves the estimation of non-null distribution $f_0(x)$ and proportion of non-zero ϵ when estimating the Z-value $\frac{(1-\hat{\epsilon})\hat{f}_0(x)}{\hat{F(x)}}$. In our simulation, we don't estimate $f_0(x)$ and use $N(0,1)$ as the theoretical null for all proce-

$n = 100,000, p_1 = 0.8, \epsilon = 0.1\%, n\epsilon = 100, \sigma = 1/3.$

$FDRLevel$	μ	BH	Sun and Cai	Adjusted BH
0.1	3.53	0.128/10.6/1.6	0.117/16.3/2.2	0.117/16.6/2.2
0.3	3.16	0.325/11.5/5.53	0.353/18.8/10.3	0.338/18.4/9.4
0.5	2.90	0.503/12.7/12.9	0.556/21.5/27.0	0.546/20.0/24.0
0.7	2.61	0.740/13.9/39.7	0.791/23.8/90.4	0.771/21.8/73.5

Table 2: In this study, $p_1 = 0.8$. In other words, among 100 nonzero parameters, 20 of them appear on the left tail and the rest on the right tail. Two-sided p-values are used for BH's procedure. The procedure of Sun and Cai [2007] rejects parameters on both sides. The rejection interval of our procedure is a union of two intervals.

$n = 100,000, p_1 = 0.9, \epsilon = 0.1\%, n\epsilon = 100, \sigma = 1/3.$

$FDRLevel$	μ	BH	Sun and Cai	Adjusted BH
0.1	3.95	0.104/24.2/2.8	0.110/58.5/7.3	0.111/59.1/7.4
0.3	3.61	0.315/22.8/10.5	0.321/61.1/28.9	0.303/60.9/26.5
0.5	3.39	0.502/18.8/19.0	0.537/64.5/74.8	0.512/64.0/67.2
0.7	3.15	0.714/27.5/68.6	0.756/71.8/222.9	0.726/70.4/186.6

Table 3: $p_1 = 0.9, \beta = 0.6$. Among 100 nonzero parameters, 10 of them apear on the left tail and the rest on the right tail.

dures. We also ignore the estimation of the proportion ϵ for two reasons. Firstly, ϵ is very small. When $n = 100,000$ and $\beta = 0.6$, $\epsilon = 0.1\%$. Secondly, employing the information of the estimator $\hat{\epsilon}$ is equivalent to enlarge the q-level to $\frac{q}{1-\hat{\epsilon}}$ for all procedures.

In this simulation, we consider two-sided test and the alternative appears on both sides. Therefore, the reject region of the proposed procedure is a union of two intervals, appearing on both the positive and negative side. Two-sided p-values are used for BH procedure. In Table 2 when $p_1 = 0.8$, it is clearly seen that the BH procedure has much less power in detecting nonzero parameters than the other two. For instance, in Table 3 when $p_1 = 0.9$ and $q = 0.1$, BH procedure discoveries 24.2 true positive on average, and the other two methods declare 58.5 and 59.1 true positives respectively. The procedure of Sun and Cai [2007] have essentially the same number of true positives as the proposed method. However, the latter one commits much smaller false positives.

For instance, in the last row of Table 2 where $\sigma = 1/3$, $\epsilon = 0.1\%$, and $q = 0.7$. The average number of rejection of Sun and Cai [2007] is 114.2 while that of ours is only 95.3. However, the majority of the difference is contributed by the number of false positives. Sun and Cai [2007] commits 16.9 more false positive than ours on average, 16.9% of the number of true signals. Consequently, Sun and Cai [2007] tends to have a larger FDR than ours.

FDR level(q)	BH	Sun and Cai	Adjusted BH
0.4	0.511/16/19	0.556/19/29	0.500/20/19
0.5	0.561/17/21	0.667/21/47	0.588/20/25
0.6	0.621/19/27	0.771/21/76	0.632/21/32
0.7	0.634/22/33	0.876/21/151	0.661/21/40

Table 4: This table summaries the data analysis result of three testing procedures for the HGU133A data set when using the theoretical null. In each cell, three numbers correspond to the median False Discovery Proportion, median number of true positives and false positives among all rejections. We set $f_0(x)$ as the standard normal distribution.

FDR level(q)	BH	Sun and Cai	Adjusted BH
0.4	0.368/8/2	0.478/11/11	0.467/9/7
0.5	0.469/12/7	0.586/12/17	0.486/12/8
0.6	0.485/12/8	0.685/13/31	0.538/13/16
0.7	0.486/14/12	0.782/13/62	0.551/17/23

Table 5: This table summaries the data analysis result of three testing procedures for the HGU133A data set when using the empirical null. In each cell, three numbers correspond to the median of False Discovery Proportion, median number of true positives and false positives among all rejections. We use the empirical null hypothesis $f_0(x) = \frac{1}{\hat{\sigma}}\phi(\frac{x-\hat{\mu}}{\hat{\sigma}})$.

The same conclusion holds in Table 3 when $p_1 = 0.9$.

In summary, the adjusted BH procedure has more power in detecting the true significance than that of the BH while committing less false positives than that of Sun and Cai [2007].

3 Real Data Analysis

In this section, we apply various procedure to two datasets. In the previous section, it is shown that the proposed method can identify many true significance and commit the least number of false positives. This leads to a number of rejection compared to the local fdr based approach, such as the one of Sun and Cai [2007]. It remains difficult to compare the two methods if we don't know which hypothesis are null and which are non-null. We therefore apply these methods to the spike-in data where the hypothesis are preset and known. We can thus compare various procedures.

Example 1: Affymetrix's Spike-in HGU133A Experiment

This is a spike-in data set performed on the HGU133A platform. In each experiments, 42 genes are spiked in at 14 concentrations raging from 0pM to 412pM. Three genes are spiked

in at each concentration and the Latin Square design is applied. There are three technical replicates in each array group, resulting in 42 arrays in total. For more information of the data, please read Irizarry et al. [2003].

We analyze the data using RMA and examine the log fold changes in the expression between the first experiments and the rest 13 experiments. Three testing procedures as used in Section 2 are applied by setting various q level, varying among 0.4, 0.5, 0.6, and 0.7. For each pair of experiments, we can calculate the number of true/false rejection and FDP, and we then calculate the median. We prefer the median to the mean because of its robust to the outliers. We then summarize the data in both Table **??** and 5, where we use the theoretical null $f_0(x) = \phi(x)$ in Table 4 and the empirical null $f_0(x) = \frac{1}{\hat{\sigma}}\phi(\frac{x-\hat{\mu}}{\hat{\sigma}})$ in Table 5.

When considering the theoretical null as applied in Table 4, it is clearly seen that the proposed approach has more number of true rejection than the BH procedure. For instance, when $q = 0.5$, we declare 20 true significance while BH procedure can only detect 17 true significance. The price we pay is a slightly more false rejections, 25 vs 21 when $q = 0.5$. We can observe a similar pattern when using the empirical null as in Table 5.

On the other hand, the procedure of Sun and Cai [2007], which is local fdr based, has the similar power to the proposed procedure. However, it commits much more false significance than both the proposed approach and the BH procedure. For instance, when using the theoretical null, and setting q as 0.6, both their approach and the proposed one declare 21 true significance. However, The median number of false rejection of Sun and Cai [2007] is 76, more than as twice as that number of the proposed approach, which is only 32. This data analysis demonstrates that the proposed approach is quite similar to the BH procedure and is much better than that of Sun and Cai [2007].

Example 2. Golden Spike-in Data Set. We consider another spike in data set, the golden spike in data set of Choe et al. [2005].

We download the data from `http://www.elwood9.net/spike` and preprocess the data according to Hwang et al. [2009]. After taking the $\log_2$-transformation, we fit a gene-specific one-way ANOVA models to the data of size 14010x6. Among all these genes, 1331 of them

FDR level(q)	BH	Sun and Cai	Adjusted BH
0.01	0.052/202/11	0.033/353/12	0.036/349/13
0.05	0.151/602/107	0.115/721/94	0.109/728/89
0.10	0.247/760/249	0.223/848/243	0.189/859/201
0.15	0.322/851/406	0.310/908/408	0.330/922/312

Table 6: This table summaries the data analysis result of three testing procedures. In each cell, three numbers correspond to the False Discovery Proportion, number of true positives and false positives among all rejections. We use the theoretical null $f_0(x) = \phi(x)$.

have differential expression being non-zero. There are three replicates in both the treatment and control groups, denoted as y_{tij} and y_{cij}, where $i = 1, 2, \cdots, p(= 14010)$, $j = 1, 2, 3$. Let

$$x_i = \bar{y}_{ti\cdot} - \bar{z}_{ci\cdot}, s_i = (s_{ti}^2/3 + s_{ci}^2/3)^{1/2},$$

where s_{ti} and s_{ci} are the sample variance of observations corresponding to the i-th in both the treatment and control group. Let $t_i = x_i/s_i$ be the T statistic with the degrees freedom d_i taken to be the Satterthwaite approximation. At the end, define the Z-statistic as

$$z_i = \Phi^{-1}(P(T_{d_i} \leq t_i)),$$

where Φ is the CDF of the standard normal distribution. Now, we apply three approaches: Adjusted BH, BH (Benjamini and Hochberg [1995]), SW(Sun and Cai [2007]) to this Z statistic. The FDR level we are aiming at controlling is $q = 0.05, 0.1, 0.15$, and 0.20. The rejection region appears on both tails. We have reported the result in Table **??**.

It is clearly seen that both the adjusted BH procedure and Sun and Cai [2007] perform better than Benjamini and Hochberg [1995]. The adjusted BH procedure has much less false positives and significant more true positives compared with the procedure of Sun and Cai [2007].

Unfortunately, all the procedures fail to control the FDR at the pre-specified level. Nevertheless, our approach always has the smallest estimated FDR. For different q level, the number of the true rejection of our procedure is similar to that of Sun and Cai [2007] and much larger than that of BH's procedure. On the other hand, our procedure has the smallest number of false positives. Consequently, our approach is strongly recommended

FDR level(q)	BH	Sun and Cai	Adjusted BH
0.01	0/0/0	0/0/0	0.5/1/1
0.05	0.034/228/8	0.048/535/27	0.049/539/28
0.10	0.048/539/27	0.124/771/109	0.123/768/108
0.15	0.138/676/70	0.203/883/225	0.202/873/221

Table 7: This table summaries the data analysis result of three testing procedures. In each cell, three numbers correspond to the False Discovery Proportion, number of true positives and false positives among all rejections. We use the empirical null $f_0(x) = \frac{1}{\hat{\sigma}}\phi(\frac{x-\hat{\mu}}{\hat{\sigma}})$.

for applications.

4 More of the Ideal Interval

In this section, we will continue discussing the conditions for the existence of the ideal interval (1.4) and the ideal BH interval.

Under an additional weak condition, this necessary condition is also sufficient for the existence for the ideal interval I.

Theorem 5 *(a) Assume that* $\max_x h(x) > \frac{1}{q'}$ *and there exists an* c_0*, such that* $h(c) \leq \frac{1}{q'}, \forall c > c_0$ *and* $h(c) \geq \frac{1}{q'}, \forall c_0 - \delta < c < c_0$ *for some* $\delta > 0$*, then there exists an interval* $I = [a, b]$ *such that*

$$\int_a^b f_0(x)dx = q\int_a^b dF(x).$$

(b) If $h(x)$ *is monotone increasing function of* x *and* $\max_x h(x) > \frac{1}{q'}$*, then the same conclusion as in (b) holds.*

A large amount of models satisfy the conditions in this theorem, such as those models with $h(x)$ being unimodal. Once the mode of $h(x)$ is greater than $1/q'$, it is almost guaranteed the existence of the ideal interval.

The very natural question we might ask is whether the necessary condition in Theorem 2 is also sufficient for the existence of ideal BH interval. The answer is yes and no. It is sufficient when $h(x)$ is a monotone increasing function of x and is not otherwise. The next theorem will describe the sufficient condition for the existence of the ideal BH procedure when $h(x)$ is either monotone or unimodal.

Theorem 6 *(a) If $h(x)$ is monotone increasing and $\max h(x) \geq \frac{1}{q'}$, then the ideal BH interval $I = [t, +\infty)$ exists;*

(b) Assume that $h(x)$ is unimodal with the mode $x^* > 0$ and $\max_x h(x) \geq \frac{1}{q'}$. Let c_1 and c_2 $(c_1 \leq c_2)$ be two constants such that $h(c_1) = h(c_2) = \frac{1}{q'}$ where $q' = \frac{q\epsilon}{1-(1-\epsilon)q}$. Then there exists an interval $I = [a, \infty)$ satisfying $\int_I f_0(x)dx \leq q \int_I dF(x)$ if and only if

$$\int_{c_1}^{\infty} f_0(x)dx \leq q \int_{c_1}^{\infty} dF(x). \tag{4.10}$$

Both Theorem 2 and 6 tell us that there exists an ideal BH interval if and only if $\max_x h(x) \geq \frac{1}{q'}$ when assuming $h(x)$ being monotone increasing.

Essentially, the existence of the ideal interval only requires the mode of $h(x)$ to exceed a threshold level $\frac{1}{q'}$. Since the ideal BH interval $I = [t, +\infty)$ contains a subinterval $I' = (t_1, +\infty)$ which might contribute more to $\int_{I'} f_0(x)dx$ than $\int_{I'} f_1(x)dx$ when $h(x)$ is non-monotone, it requires an additional condition (4.10). To explain the discrepancy between these two conditions, consider two special example.

5 Appendix

Proof of Theorem 1:

For any set T, the deinition of S indicates that

$$(1_S - 1_T)(f_1(x) - C(q)f_0(x)) \geq 0.$$

Let $\alpha(S) = \int_S f_0, \beta(S) = \int_S f_1$ and similarly define $\alpha(T)$ and $\beta(T)$. Then

$$0 \leq \int (1_S - 1_T)(f_1(x) - C(q)f_0(x))dx = \beta(S) - \beta(T) + C(q)(\alpha(T) - \alpha(S)). \tag{5.11}$$

Since

$$\int_S f_0(x) - q\int_S dF(x) = (1-q(1-\epsilon))\alpha(S) - q\epsilon\beta(S), \int_T f_0(x) - q\int_T dF(x) = (1-q(1-\epsilon))\alpha(T) - q\epsilon\beta(T),$$

then

$$(1-q(1-\epsilon))(\alpha(T)-\alpha(S)) = [(\int_T f_0(x)-q\int_T dF(x))-(\int_S f_0(x)-q\int_S dF(x))]+q\epsilon(\beta(T)-\beta(S)).$$

Combining this with equation (5.11), one knows that

$$0 \leq (1-\frac{C(q)q\epsilon}{1-q(1-\epsilon)})(\beta(S)-\beta(T)) + [(\int_T f_0(x) - q\int_T f_1(x)) - (\int_S f_0(x) - q\int_S f_1(x))].$$

The choice of T satisfies $\int_T f_0(x) - q\int_T f_1(x) \leq 0 = \int_S f_0(x) - q\int_S f_1(x)$. Therefore,

$$0 \leq (1-\frac{C(q)q\epsilon}{1-q(1-\epsilon)})(\beta(S)-\beta(T)).$$

Since $\int_S f_0(x) = q\int_S dF(x)$, then

$$\int_S f_0(x) = \frac{q\epsilon}{1-q(1-\epsilon)}\int_S f_1(x) \geq \frac{C(q)q\epsilon}{1-q(1-\epsilon)}\int_S f_0(x). \tag{5.12}$$

Consequently, $1-\frac{C(q)q\epsilon}{1-q(1-\epsilon)} \geq 0$, which implies that $\beta(S) \geq \beta(T)$. Further

$$\int_S [(1-\epsilon)f_0(x)+\epsilon f_1(x)] = [\frac{(1-\epsilon)q\epsilon}{1-q(1-\epsilon)}+\epsilon]\beta(S).$$

On the other hand, $\int_T f_0(x) \leq q\int_T f_1(x)$, and we know that

$$\int_T [(1-\epsilon)f_0(x)+\epsilon f_1(x)] \leq [\frac{(1-\epsilon)q\epsilon}{1-q(1-\epsilon)}+\epsilon]\beta(T).$$

In summary, one knows that

$$\int_T dF(x) \leq \int_S dF(x).$$

□

Proof of Theorem 2:

(1) For any interval $I=[a,b]$, let $s(a,b) = \int_a^b f_0(x)dx - q'\int_a^b f_1(x)dx$. Then

$$\frac{\partial s}{\partial b} = f_0(b) - q'f_1(b) = q'f_0(b)(1-q'\frac{f_1(b)}{f_0(b)}) > 0.$$

Consequently, for any fixed a, $s(a,b)$ is increasing with respect to b for any $b > 0$. Since $s(a,a) = 0$, therefore, $s(a,b) > 0, \forall b > a$. This implies that

$$\int_a^b f_0(x)dx > q' \int_a^b f_1(x)dx.$$

Consequently, there exists no interval I such that $\int_I f_0(x)dx \leq q \int_I dF(x)$.

(2) Let $I_q(F) = [a_0, b_0]$. Then according to Theorem 1, $h(a_0) = h(b_0)$. In order to show that $[c_1, c_2] \subset [a_0, b_0]$, it sufficies to prove that $s(c_1, c_2) < 0$. Indeed, $s(c_1, c_1) = 0$ and $\frac{\partial s}{\partial b} < 0, \forall b \in [c_1, c_2]$ implies that $s(c_1, c_2) < 0$. □

Proof of Theorem 3:

Before we prove the theorem, we will state the following lemmas.

Lemma 5.1 *Let F_n be the empirical cdf, then $\forall a$, if $F'(b_a(F)) - \frac{1}{q} f_0(b_a(F)) \neq 0$, then*

$$b_a(F_n) \to b_a(F), \text{ and } \quad g_n(a) \to g(a). \tag{5.13}$$

If $F'(b_a(F)) - \frac{1}{q} f_0(b_a(F)) = 0$, then $\overline{\lim} g_a(F_n) \leq g_a(F)$.

Lemma 5.2 *There exists a neighborhood A of a_0 such that $\forall \xi \in b_A(F)$,*

$$f_0(\xi) - qF'(\xi) > L, \tag{5.14}$$

for some positive constant L.

Lemma 5.3 *There exists a sub-interval B of A, such that for all $a \in B$, $|b_a(F_n) - b_a(F)| \leq C(L)\epsilon$ provided that $||F_n - F|| < \epsilon$.*

Lemma 5.4 *The function $g_n(a)$ can not achieve the maximum at B^c.*

Lemma 5.5 *For any $a \in B$, then $|g_n(a) - g(a)| < \epsilon$.*

Lemma 5.6 $g''(a_0) \neq 0$.

Proof of Theorem 3: Assume that $g_n(a)$ attains the maximum at $a = a_1$, then according to Lemma 5.4, $a_1 \in B$. Since $g(a_1) - g(a_0) \leq 0$, and

$$g(a_1) - g(a_0) = g(a_1) - g_n(a_1) + g_n(a_1) - g_n(a_0) + g_n(a_0) - f(a_0) > -2\epsilon. \tag{5.15}$$

Consequently, $|g(a_1) - g(a_0)| < 2\epsilon$. In other words,

$$(a_1 - a_0)^2 \frac{g''(a_0)}{2} < 2\epsilon. \tag{5.16}$$

Therefore $|a_1 - a_0| \leq C(L)\epsilon^{-1/2}$ and furthermore $F'(b_{a_0}(F)) - \frac{1}{q} f_0(b_{a_0}(F)) \neq 0$ implies that $b_{a_1}(F) - b_{a_0}(F) = O(\epsilon^{-1/2})$. Consequently,

$$|b_{a_1}(F_n) - b_{a_0}(F_n)| \leq |b_{a_1}(F) - b_{a_0}(F)| + 2\epsilon = O(\epsilon^{-1/2}). \tag{5.17}$$

All in together, we know that the two ending points of the interval a_1 and $b_{a_1}(F_n)$ converge to the ideal ending points with the rate $O(\epsilon^{-1/2}) = O(n^{-1/4})$. □

Proof of the Lemmas.

Lemma 5.1:

Since F_n is the empirical CDF, DKS's theorem guarantees that $\forall \epsilon > 0$, with high probability

$$F(x) - \epsilon \leq F_n \leq F(x) + \epsilon, \forall x. \tag{5.18}$$

Consider the function

$$F_U(x) = \begin{cases} F(x) + \epsilon & \forall x > a \\ F(x) & \forall x \leq a \end{cases}$$

Then by the definition of $b_a(F_n)$ and F_U,

$$\frac{1}{q} \leq \frac{F_n(b_a(F_n)) - F_n(a)}{F_0(b_a(F_n)) - F_0(a)} \leq \frac{F_U(b_a(F_n)) - F_U(a)}{F_0(b_a(F_n)) - F_0(a)}. \tag{5.19}$$

Consequently, $b_a(F_n) \leq b_a(F_U)$. Similarly define

$$F_L(x) = \begin{cases} F(x) - \epsilon & \forall x > a \\ F(x) & \forall x \leq a \end{cases}$$

Then one can similarly show that $b_a(F_L) \leq b_a(F_n)$. As a result,

$$b_a(F_L) \leq b_a(F_n) \leq b_a(F_U). \tag{5.20}$$

When $\epsilon \to 0$ and $f_0(b_a(F)) - qF'(b_a(F)) \neq 0$, one know that $\lim_{\epsilon \to 0} b_a(F_L) = \lim_{\epsilon \to 0} b_a(F_U) = \lim_{\epsilon \to 0} b_a(F)$. Therefore

$$b_a(F_n) \to b_a(F). \tag{5.21}$$

Further,

$$\begin{aligned} & |g_n(a) - g(a)| = |F_n(b_a(F_n)) - F_n(a) - F(b_a(F)) + F(a)| \\ \leq \; & |F_n(b_a(F_n)) - F(b_a(F_n))| + |F(b_a(F_n) - F(b_a(F))| + |F_n(a) - F(a)| \\ \leq \; & 2\epsilon + |F(b_a(F_n) - F(b_a(F))| \to 0. \end{aligned}$$

If $f_0(b_a(F)) - qF'(b_a(F)) = 0$, then there exists an neighborhood C of $b_a(F)$ such that $s(a, x) > \delta > 0, \forall x \in C, x > b_a(F)$. Then $b_a(F_n)$ is bounded by the maximum of $b_a(F_L)$ and $b_a(F)$. Consequently,

$$\overline{\lim}\, g_a(F_n) \leq g_a(F). \tag{5.22}$$

Lemma 5.3: Let B be a sub-interval of A which contains a_0 and satisfy the distance between $\bar{B}$ and $\bar{A}$ is greater than a positive quantity.

For any $a \in B$, let $t_U = argmax_t\{(F_0(t) - F_0(a)) - q(F_n(t) - F_n(a)) \leq 0\}$. If $t_U > b_a(F)$, we will show that $t_U \in b_A(F)$. Otherwise, $F_0(t_U) - F_0(a) - q(F(t_U) - F(a)) > 0$. The fact that

$$f_0(x) - qF'(x) > 0, \forall x > b_{a_0}(F)$$

imples that the previous quantity actually has a lower bound Δ. By the definition of t_U,

$$F_0(t_U) - F_0(a) - q(F_n(t_U) - F_n(a)) \leq 0. \tag{5.23}$$

Consequently,

$$q(F(t_U) - F_n(t_U)) \leq -\Delta + q(F(a) - F_n(a)) \leq -\Delta + q\epsilon < -q\epsilon, \tag{5.24}$$

which leads to a contradiction. Therefore $t_U \in b_A(F)$.

Similarly, let

$$t_L = argmax_t\{t < b_a(F_n) : (F_0(t) - F_0(a)) - q(F_n(t) - F_n(a)) \leq 0\}.$$

Let $s(t) = F_0(t) - F_0(a) - q(F(t) - F(a))$, then $s(b_a(F)) = 0$ and

$$s'(t)|_{t=b_a(F)} = f_0(b_a(F)) - qF'(b_a(F)) < 0. \tag{5.25}$$

Then we can find $t_0 < b_a(F), t_0 \in b_A(F)$, such that

$$F_0(t_0) - F_0(a) - q(F(t_0) - F(a)) = -\delta < 0 \tag{5.26}$$

Therefore for sufficiently small ϵ,

$$F_0(t_0) - F_0(a) - q(F_n(t_0) - F_n(a)) < -\delta + 2\epsilon < 0 \tag{5.27}$$

which implies that $t_L > t_0$. Consequently, $t_L < b_a(F_n) < t_U$ and $t_L, t_U \in h_A(F)$.

Next, we will prove that

$$|t_L - b_a(F)| \leq L\epsilon, |t_U - b_a(F)| \leq L\epsilon. \tag{5.28}$$

Indeed, since $F_0(t_U) - F_0(a) - q(F_n(t_U) - F_n(a)) \leq 0$ and $F_0(b_a(F)) - F_0(a) - q(F(b_a(F)) -$

$F(a)) = 0$, then

$$q(F_n(t_U) - F(b_a(F))) - (F_0(t_U) - F_0(b_a(F))) > q(F_n(a) - F(a)). \tag{5.29}$$

As a result,

$$q(F(t_U) - F(b_a(F))) - (F_0(t_U) - F_0(b_a(F))) > q(F_n(a) - F(a)) + q(F(t_U) - F_n(t_U)) > -2q\epsilon. \tag{5.30}$$

By the definition of t_U,

$$F_0(t_U^+) - F_0(a) - q(F_n(t_U^+) - F_n(a)) > 0. \tag{5.31}$$

Therefore

$$q(F(t_U^+) - F(b_a(F))) - (F_0(t_U^+) - F_0(b_a(F))) < q(F(t_U^+) - F_n(t_U^+)) + q(F_n(a) - F(a)) < 2q\epsilon. \tag{5.32}$$

All in together, we know that

$$-2q\epsilon < q(F(t_U) - F(b_a(F))) - (F_0(t_U) - F_0(b_a(F))) < 2q\epsilon. \tag{5.33}$$

Therefore

$$|(t_U - b_a(F))(qF'(\xi) - f_0(\xi))| \leq 2q\epsilon. \tag{5.34}$$

Since $|qF'(\xi) - f_0(\xi)| > L$, then $|t_U - b_a(F)| \leq C(L)\epsilon$. We can similarly proof the statement for t_L. Consequently, $|b_a(F_n) - b_a(F)| = O(\epsilon)$.

Lemma 5.4

Firstly, we will show that there exists a positive constant Δ such that $g(a_1) - g(a_0) < -\Delta$, $\forall a_1 \notin B$.

Let $s(a,b) = \int_a^b f_0(x)dx - q' \int_a^b f_1(x)dx$. Then

$$\frac{\partial s}{\partial b} = f_0(b)(1 - q'\frac{f_1(b)}{f_0(b)}) \begin{cases} > 0 & if b < c_1, \\ < 0 & if c_1 < b < c_2, \\ > 0 & if b > c_2 \end{cases}$$

Consequently, for fixed a, $s(a,b)$ decareses with respect to b when $b < c1$ or $b > c_2$; it increases with respect to b if $c_1 < b < c_2$. Simiarly,

$$\frac{\partial s}{\partial a} = f_0(a)(q'\frac{f_1(a)}{f_0(a)} - 1) \begin{cases} < 0 & if a < c_1, \\ > 0 & if c_1 < a < c_2, \\ < 0 & if b > c_2 \end{cases}$$

For fixed b, $s(a,b)$ decreases for $a < c_1$ or $a > c_2$; it inreases for $c_1 < a < c_2$.

(a) Since $s(-\infty, c_2) = \int_{-\infty}^{c_2}(f_0(x) - q'f_1(x))dx > \frac{1}{2} = q' > 0$ and $s(a, c_2)$ decreases when $a < c_2$ and increases when $c_1 < a < c_2$. Combining this with the fact that $s(c_2, c_2) = 0$, one knows that there exists a unique $a^* < c_1$ such that $s(a^*, c_2) = 0$.

(b) Consider $s(c_1, b)$, then $s(c_1, c_1) = 0$ and $s(c_1, \infty) > 0$. Since $s(c_1, b)$ decreases when $c_1 < b < c_2$ and increases when $b > c_2$, then there exits a unique c^* such that $s(c_1, c^*) = 0$.

(c) Let $\mathcal{I} = \{[a,b] : s(a,b) \leq 0\}$ and $A = \{a : \text{there exists } b > a \text{ such that } I = [a,b] \in \mathcal{I}\}$. First, we prove that $A = [a^*, c_2]$. Indeed if $a' > c_2$, then

$$s(a', b) > s(a', a') = 0;$$

if $a' < a^*$, then $s(a', b) > s(a^*, b) \geq 0, \forall b > a^*$. Consequently $A \subset [a^*, c_2]$.

On the other hand, for any $a^* \leq a \leq c_1$, $s(a, c_2) \leq s(a^*, c_2) = 0$ and $s(a, c_2) \leq s(c_2, c_2) =$

$0, \forall c_1 < a \leq c_2$ Consequenlty, $A = [a^*, c_2]$. For any $a \in A$, let

$$b(a) = argmax_b\{\int_a^b dF(x)\}.$$

Then $b(a)$ is a continuous function. If $c_1 < a \leq c_2$, then $b(a) < b(c_1)$ implying $\int_a^b(a)dF(x) < \int_{c_1}^{b(c_1)} dF(x)$. Consequently, we only consider $A = [a^*, c_1]$.

Let $f(a) = \int_a^{b(a)} dF(x)$. Then $f : A \to [0, 1]$ is a continuous function on a closed set and it attains the maximum on the boundary or local maximam. Based on Theorem 1, the function $g(a)$ attains the unique maximum at $a = a_0$. Therefore, we can find a positive constant Δ such that

$$g(a_1) - g(a_0) < -\Delta, \forall a_1 \in B^c. \tag{5.35}$$

Now, we will prove the lemma. For any $a_1 \in B^c$, if a_1 satisfies $f_0(b_{a_1}(F)) - qF'(b_{a_1}(F)) = 0$, Lemma 5.1 implies that f_n can not acheive the maximum at a_1.

If $F'(b_{a_1}(F)) - \frac{1}{q}\phi(b_{a_1}(F)) \neq 0$, then

$$g_n(a_1) - g_n(a_0) = g_n(a_1) - g(a_1) + g(a_1) - g(a_0) + g(a_0) - g_n(a_0) < -\Delta + g_n(a_1) - g(a_1) + g(a_0) - g_n(a_0). \tag{5.36}$$

According to Lemma 5.1, $g_n(a_1) \to g(a_1), g_n(a_0) \to g(a)$, then $g_n(a_1) < g_n(a_0)$. Consequently, g_n attains the maximum in A.

Lemma 5.5

$$\begin{aligned}
& |g_n(a) - g(a)| = |F_n(b_a(F_n)) - F_n(a) - F(b_a(F)) + F(a)| \\
= \; & |F_n(b_a(F_n)) - F(b_a(F_n)) + F(b_a(F_n)) - F(b_a(F)) - (F_n(a) - F(a))| \\
\leq \; & 2\epsilon + F(b_a(F_n)) - F(b_a(F)) \leq 2\epsilon + |b_a(F_n) - b_a(F)||F'(\xi)| \\
\leq \; & 2\epsilon + |F'(\xi)| \max\{b_a(F) - b_a(F_L), b_a(F_U) - b_a(F)\},
\end{aligned}$$

where $b_a(F_L)$ and $b_a(F_U)$ are defined in the proof of Lemma 5.1. Since $|f_0(b_a(F)) -$

$qF'(b_a(F))| > L$, $\forall a \in B$, then

$$b_a(F) - b_a(F_L) = O(\epsilon), b_a(F_U) - b_a(F) = O(\epsilon).$$

Consequently,

$$|g_n(a) - g(a)| \leq C\epsilon.$$

Lemma 5.6

According to the definition of $b_a(F)$,

$$(1 - q(1 - \epsilon))(F_0(b_a(F)) - F_0(a)) = q\epsilon(F_1(b_a(F)) - F_1(a)). \tag{5.37}$$

Take the second derivative on both sides of the above equation, one knows that

$$(1-q(1-\epsilon))(f_0'(b_a(F))(b_a'(F))^2+f_0(b_a(F))b_a''(F)-f_0'(a)) = q\epsilon(f_1'(b_a(F))(b_a'(F))^2+f_1(b_a(F))b_a''(F)-f_1'(a)). \tag{5.38}$$

The definition of $g(a)$ implies that $g(a) = (1-\epsilon)(F_0(b_a(F))-F_0(a))+\epsilon(F_1(b_a(F))-F_1(a))$.

Thus

$$\begin{aligned} & g''(a_0) \\ = \; & (1-\epsilon)(f_0'(b_{a_0}(F))(b_{a_0}'(F))^2 + f_0(b_{a_0}(F))b_{a_0}''(F) - f_0'(a)) + \epsilon(f_1'(b_{a_0}(F))(b_{a_0}'(F))^2 + f_1(b_{a_0}(F))b_{a_0}''(F) \\ = \; & \frac{1}{q}(f_0'(b_{a_0}(F))(b_{a_0}'(F))^2 + f_0(b_{a_0}(F)) - f_0'(a_0) \neq 0. \end{aligned}$$

□

Proof of Theorem 4

Let $A = \{\sup_{t\in[0,1]} |F_n(t) - F(t)| > \epsilon\}$ and $A_{m,i} = \{\sup_{t\in[\frac{i-1}{m},\frac{i}{m}]} |F_n(t) - F(t)| > \epsilon\}$ where $i = 1, 2, \cdots, m$. Then $\forall m$,

$$P(A) = P(\cup_i A_{m,i}) \leq \sum_i P(A_{m,i}). \tag{5.39}$$

Let $t_i = \frac{i}{m}$. Then for any $t \in [t_{i-1}, t_i]$, one knows that

$$\begin{aligned} & |F_n(t) - F(t)| \leq |F_n(t_i) - F(t_i)| + |F_n(t) - F(t) - F_n(t_i) + F(t_i)| \\ \leq \; & |F_n(t_i) - F(t_i)| + F_n(t_i) - F_n(t_{i-1}) + |F(t_i) - F(t)| \\ \leq \; & 2|F_n(t_i) - F(t_i)| + |F_n(t_{i-1}) - F(t_{i-1})| + |F(t_i) - F(t_{i-1})| + |F(t_i) - F(t)| \\ \leq \; & \frac{C}{m} + 2|F_n(t_i) - F(t_i)| + |F_n(t_{i-1}) - F(t_{i-1})|. \end{aligned}$$

Consequently,

$$P(A_i) \leq P(|F_n(t_i) - F(t_i)| > (\epsilon - \frac{C}{m})/3) + P(|F_n(t_{i-1}) - F(t_{i-1})| > (\epsilon - \frac{C}{m})/3). \quad (5.40)$$

According to Theorem 1.3 of **?**, one knows that for each integer $q \in [1, n/2]$, and each $\epsilon > 0$,

$$P(|F_n(t_i) - F(t_i)| > n\epsilon) \leq 4\exp(-\frac{\epsilon^2}{8b^2}q) + 22(1 + \frac{4b}{\epsilon})^{1/2} q\alpha([\frac{n}{2q}]). \quad (5.41)$$

Assume that $\epsilon = n^x$, $m = n^{-x}$, and $q = n^z$ where $x > -\frac{2k}{4k+7}$, and $z = -2x + \delta$ where $\delta > 0$ satisfies $\delta > \frac{4k+7}{2(k+1)}(x - \frac{2k}{4k+7})$. Then

$$P(A) \leq \sum_i P(A_i) \asymp m(4\exp(-n^{2x+z}) + 22(n^{-x/2+z-k(1-z)})) \asymp n^{-3x/2-k+(k+1)z} \to 0. \quad (5.42)$$

□

Proof of Theorem 5:

Part (a). Let $c_0 - \delta < a < c_0$ and consider an interval $I = [a, b]$. Let

$$s(a, b) = \int_a^b f_0(x)dx - q' \int_a^b f_1(x)dx.$$

Then $\frac{\partial s}{\partial b} = f_0(b) - q' f_1(b) = f_0(b)(1 - q' \frac{f_1(b)}{f_0(b)})$. Consequently, $s(a, b)$ decreases when $a < b < c_0$ and increases when $b > c_0$. Since $s(a, a) = 0$, it suffices to show that there exists $b > a$ such that $s(a, b) = 0$. In other words, one only needs to show that there exists an $c_1 < a < c_2$ such that $s(a, \infty) > 0$.

Since $s(c, \infty) = \int_c^\infty f_0(x)dx - q' \int_c^\infty f_1(x)$ decreases for $c > c_0$, then $s(c_0, +\infty) > 0$. The

continuity of $s(c, +\infty)$ indicates that there exits a $c_0 - \delta < a < c_0$ and $b > c_2$, such that $s(a, b) = 0$. In other words, the interval $I = [a, b]$ satisfies

$$\int_I f_0(x)dx = q \int_I dF(x).$$

Part (b). WLOG, assume that $\frac{f_1(x)}{f_0(x)}$ is monotone increasing. Let $\frac{f_1(c_1)}{f_0(c_1)} = \frac{1}{q'}$. For any interval $I = [a, b]$, let

$$s(a, b) = \int_a^b f_0(x)dx - q' \int_a^b f_1(x)dx.$$

Then similarly as in the proof of part (b),

$$\frac{\partial}{\partial b} s(a, b) = f_0(b)(1 - q' \frac{f_1(b)}{f_0(b)}),$$

which is less than 0 for $b > c_1$. Consequently, $s(c_1, b)$ decreases with respect to b, implying $s(c_1, \infty) < 0$. Now consider $s(a, \infty)$. One knows that $h(-\infty) = 1 - q' > 0$ and $h(c_2) < 0$. Therefore, there exists a a such that $s(a, \infty) = 0$. In other words, there exists an interval $I = [a, \infty)$ such that

$$\int_a^\infty f_0(x)dx = q' \int_a^\infty f_1(x)dx.$$

□

Proof of Theorem 6:

(a) Let $s(a, +\infty) = \int_a^\infty f_0(x)dx - q' \int_a^\infty f_1(x)dx$. Then

$$\frac{\partial}{\partial a} s(a, \infty) = q' f_0(a)(h(a) - \frac{1}{q'}).$$

Choose t_0 such that $s(t, +\infty) > \frac{1}{q'}, \forall t > t_0$. Then $s(t, +\infty)$ is a monotone increasing function for $t > t_0$. Since $s(+\infty, +\infty) = 0$, then $s(t, +\infty) < 0, \forall t > t_0$.

(b)

"if" part is obvious.

"only if":

Assume that there exists an interval $I = [a^*, +\infty)$ satisfying $\int_I f_0(x)dx \leq q \int_I dF(x)$.

Since $s(+\infty, +\infty) = 0$, and $s(a, +\infty)$ is decreasing for $a > c_2$, then $s(a, +\infty) > 0, \forall a \geq c_2$. If $s(c_1, +\infty) > 0$, then $s(a, +\infty) > 0, \forall c_1 \leq a < c_2$ because $s(a, +\infty)$ increases for $c_1 < a < c_2$ and $s(a, +\infty) > 0, \forall a < c_1$ because $s(a, +\infty)$ decreases when $a < c_1$. Consequently, $s(a, +\infty) > 0, \forall a$ which leads to a contradiction.

Consequently, $s(c_1, +\infty) \leq 0$, i.e.

$$\int_{c_1}^{+\infty} f_0(x)dx \leq q \int_{c_1}^{+\infty} dF(x).$$

□

References

Y. Benjamini and Y. 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.

T. Tony Cai, Jiashun Jin, and Mark G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.*, 35(6):2421–2449, 2007. ISSN 0090-5364.

T.T. Cai and J. Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38 (1):100–145, 2010.

S. E. Choe, M. Bouttros, A. M. Michelson, G. M. Chruch, and M.S. Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16.1–16, 2005.

B. Efron. Local false discovery rates. 2005.

Bradley Efron. Size, power and false discovery rates. *Ann. Statist.*, 35(4):1351–1377, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001460. URL `http://dx.doi.org/10.1214/009053606000001460`.

Bradley Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23 (1):1–22, 2008.

Bradley Efron. *Large-scale inference, empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2010.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):499–517, 2002. ISSN 13697412. URL `http://www.jstor.org/stable/3088785`.

J. T. Hwang, J. Qiu, and Z. Zhao. Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(1):265–285, 2009.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15, 2003.

Jiashun Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(3):461–493, 2008. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00645.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2007.00645.x`.

Wenguang Sun and T. Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.*, 102(479):901–912, 2007. ISSN 0162-1459.

C. Zhang, J. Fan, and T. Yu. Multiple testing via fdrl for large-scale imaging data. *The Annals of Statistics*, 39(1):613–642, 2011.