Lecture 3 — April 5

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Logan Bell*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by John Cherian, Paula Gablenz, and David Ritzwoller.*

3.1 Recap: Bonferroni and Fisher Have Drawbacks

In the previous two lectures, we have introduced and discussed multiple hypothesis testing. We have been interested in testing the global null, and to this end we considered Bonferroni's method and Fisher's combination test, but each of these has its pros and cons. By examining their performance in the independent Gaussian model,[1] we determined the following:

- **If our data exhibit strong, sparsely distributed signals**, then Bonferroni's method excels and is optimal in the "needle in a haystack" setting in which one μ_i is nonzero, but Fisher's method performs very poorly.
- **If our data exhibit small, widely distributed signals**, Fisher's method excels and is optimal as expected noise power overtakes expected signal power, but Bonferroni's method is powerless.

In light of these facts, we would like a test that combines the strengths of Bonferroni and Fisher. This lecture will examine a few such methods. We will proceed by considering n hypothesis $H_{0,1}, \dots, H_{0,n}$ and p-values $p_1, \dots, p_n$ such that $p_i \sim \text{Unif}(0,1)$ under $H_{0,i}$, and we will be testing the global null $H_0 = \bigcap_{i=1}^n H_{0,i}$.

3.2 Simes Test

Simes test, introduced independently by Simes [8] in 1986 and by Eklund [3] in the early 1960s, is a modification of Bonferroni's test that is less conservative. Suppose we choose a significance level α and we order our p-values as $p_{(1)} \leq \cdots \leq p_{(n)}$. Bonferroni's method considers only the smallest p-value, rejecting H_0 if $p_{(1)} \leq \frac{\alpha}{n}$. Simes test extends this idea, rejecting if any $p_{(i)} \leq \alpha \frac{i}{n}$. Equivalently, we compute the **Simes statistic**,

$$T_n = \min_{i=1}^n \left(p_{(i)} \frac{n}{i} \right),$$

and reject if $T_n \leq \alpha$. Note that the factor of n/i is an adjustment factor that is 1 for the largest p-value, but n for the smallest p-value. The sketch in Figure 3.1 illustrates the test.

[1] That is, our data are sampled from $Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1)$ random variables with the global hypothesis $H_0^n = \bigcap_{i=1}^n H_{0,i}$, where $H_{0,i} : \mu_i = 0$, and alternative hypothesis H_1^n that there exists an i such that $\mu_i = \mu > 0$.

In the left panel, no p-value falls below the critical line, so H_0 would not be rejected. In the right panel, there exists a p-value which falls below the critical line, and therefore Simes test would reject H_0.

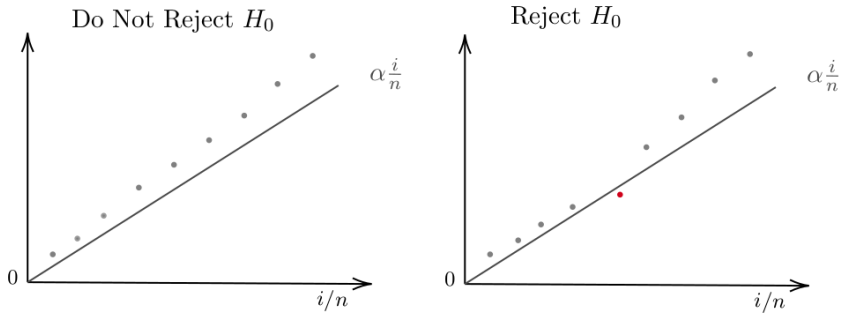

Figure 3.1. Sketch of an example where the Simes test would not reject H_0 (left panel) and where it would reject (right panel). p-values are displayed as grey dots, the critical line $\alpha\frac{i}{n}$ is shown in grey. The p-value below the critical line in the right panel is shown as a red dot.

As we observed earlier, Simes test extends Bonferroni's method to consider every p-value rather than just the smallest one. As such, its rejection region contains Bonferroni's rejection region. We can visualize the case $n = 2$, shown in Figure 3.2 below, in which Simes's rejection region is shaded in gray while Bonferroni's rejection region is contained within, depicted by hatching lines.

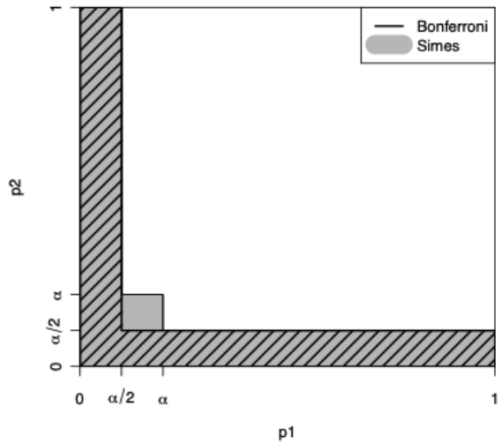

Figure 3.2. Rejection regions for Simes and Bonferroni procedures in the case $n = 2$.

In the above example, the Simes test rejects if $p_{(1)} \leq \frac{\alpha}{2}$ or $p_{(2)} \leq \alpha$, while Bonferroni's method rejects if $p_{(1)} \leq \frac{\alpha}{2}$. Though the Simes test is less conservative than Bonferroni's test, it is not too liberal. The following theorem shows that the size of Simes test is at most α when the p_i are independent under H_0.

Theorem 1. Under H_0 and independence of the p_i's, $T_n \sim \text{Unif}(0, 1)$.

Proof. We will show this by induction on n, following Simes [8]. In the base case $n = 1$, under H_0, it is clear that $T_1 = p_1 \sim \text{Unif}(0, 1)$. Now assume that $T_{n-1} \sim \text{Unif}(0, 1)$ for some

n. We will show that $\mathbb{P}(T_n \le \alpha) = \alpha$ for $\alpha \in [0,1]$. First, we split the probability into the case $p_{(n)} \le \alpha$ and $p_{(n)} > \alpha$.

$$\mathbb{P}(T_n \le \alpha) = \mathbb{P}\left(\min_{i=1}^{n}\left(p_{(i)}\frac{n}{i}\right) \le \alpha\right) = \mathbb{P}(p_{(n)} \le \alpha) + \mathbb{P}\left(\min_{i=1}^{n-1}\left(p_{(i)}\frac{n}{i}\right) \le \alpha, p_{(n)} > \alpha\right) \quad (3.1)$$

Now recall that the density of the n-th order statistic $p_{(n)}$ is given by $f(t) = nt^{n-1}$ for $t \in [0,1]$. We can thus compute each term individually. The first term is straightforward:

$$\mathbb{P}(p_{(n)} \le \alpha) = \int_0^\alpha nt^{n-1}\,dt = \alpha^n.$$

For the second term, we observe that whenever $t \in (0,1]$,

$$\min_{i=1}^{n-1} p_{(i)}\frac{n}{i} \le \alpha \iff \min_{i=1}^{n-1}\frac{p_{(i)}}{t}\cdot\frac{n-1}{t} \le \frac{\alpha}{t}\cdot\frac{n-1}{n}.$$

Moreover, conditional on $p_{(n)} = t$, the other p-values are i.i.d. $\text{Unif}(0,t)$ random variables. Thus, using the inductive hypothesis,

$$\begin{aligned}
\mathbb{P}\left(\min_{i=1}^{n-1}\left(p_{(i)}\frac{n}{i}\right) \le \alpha, p_{(n)} > \alpha\right) &= \mathbb{E}\left(\mathbb{P}\left(\min_{i=1}^{n-1}\left(p_{(i)}\frac{n}{i}\right) \le \alpha \mid p_{(n)}\right)\mathbb{I}\left(p_{(n)} > \alpha\right)\right) \\
&= \int_\alpha^1 \mathbb{P}\left(T_{n-1} \le \frac{\alpha}{t}\cdot\frac{n-1}{n}\right) f(t)\,dt \\
&= \int_\alpha^1 \frac{\alpha}{t}\cdot\frac{n-1}{n} nt^{n-1}\,dt \\
&= \alpha\int_\alpha^1 (n-1)t^{n-2}\,dt \\
&= \alpha(1-\alpha^{n-1}) \\
&= \alpha - \alpha^n.
\end{aligned}$$

Thus, (3.1) may be continued as

$$\begin{aligned}
\mathbb{P}(T_n \le \alpha) &= \mathbb{P}(p_{(n)} \le \alpha) + \mathbb{P}\left(\min_{i=1}^{n-1}\left(p_{(i)}\frac{n}{i}\right) \le \alpha, p_{(n)} > \alpha\right) \\
&= \alpha^n + \alpha - \alpha^n \\
&= \alpha.
\end{aligned}$$

□

We will see next week that the Simes test also has size at most α under a sort of positive dependence [7]. We also note that, with respect to power, Simes test is similar to that of Bonferroni. It is powerful for data which exhibit a few strong effects, and it has moderate power for many mild effects.

3.3 Tests Based on Empirical CDF's

Another approach to aggregating the effects of the p-values is to consider the resulting **empirical CDF** and compare it to the theoretical CDF one would observe under H_0.

Definition 1. The empirical CDF of $p_1, \dots, p_n$ is

$$\hat{F}_n(t) = \frac{1}{n}\#\{i : p_i \le t\}.$$

Intuitively, the empirical CDF gives the fraction of the p-values that are at most some threshold value t. In our case, under H_0 we have that $\mathbb{I}(p_i \le t) \sim \text{Ber}(t)$, so we should have $n\hat{F}_n(t) \sim \text{Bin}(n, t)$ and thus $\mathbb{E}_{H_0}(\hat{F}_n(t)) = t$. To test H_0, we would then measure the difference between this theoretical distribution and what we actually observe. If the difference between our empirical observation $\hat{F}_n(t)$ and the expected observation t is too large, then we would reject H_0. We consider three tests based on the empirical CDF: the Kolmogorov-Smirnov test, the Anderson-Darling test, and Tukey's second-level significance test.

3.3.1 Kolmogorov-Smirnov Test

The idea behind the Kolmogorov-Smirnov test is to scan over the empirical CDF and reject if the distance between $\hat{F}_n(t)$ and t is ever "too large." More precisely, we define the **Kolmogorov-Smirnov (KS) test statistic** as follows.

Definition 2. The Kolmogorov-Smirnov (KS) test statistic is defined as

$$\mathsf{KS} = \sup_t |\hat{F}_n(t) - t|.$$

We reject H_0 if KS exceeds some threshold. Another option is to consider a one-sided statistic, $\mathsf{KS}^+ = \sup_t(\hat{F}_n(t) - t)$. This statistic is used when we care especially about small p-values (as we often do), as if many small p-values are present, then $\hat{F}_n(t) - t$ will be large and positive.

In real life, finding the correct threshold is tricky since we need to know the theoretical distribution of KS under the global null hypothesis. One might use simulations or asymptotic calculations instead. A useful inequality developed by Massart [6] shows that the tail of the KS statistic is usually sub-Gaussian, and thus decays fast.

Theorem 2 (Massart's Inequality)**.** Under H_0 and independence,

$$\mathbb{P}(\mathsf{KS}^+ \ge u) \le e^{-2nu^2}$$

for $u \ge \sqrt{\frac{\log 2}{2n}}$.

A lot of work went into establishing this very nice bound, and we will not prove it here. Interested students are referred to Massart [6].

3.3.2 Anderson-Darling Test

Instead of looking at the maximum gap between the empirical CDF and the expected CDF under the global null hypothesis, as in the case of the KS statistic, another idea is to consider the cumulative gap. This idea can be expressed by a **quadratic test statistic**.

Definition 3. A quadratic test statistic is one of the form

$$\mathsf{A}^2 = n \int_0^1 (\hat{F}_n(t) - t)^2 \omega(t)\, dt,$$

where $\omega(t) \geq 0$ is a weight function.

Note that such a statistic uses the squared difference rather than the absolute difference, meaning small deviations are treated as practically negligible compared to larger gaps.[2] The weight function $\omega(t)$ exists to grant greater significance to gaps at certain thresholds. Common weight functions are $\omega(t) = 1$, which yields the **Cramér-von Mises statistic**, and $\omega(t) = [t(t-1)]^{-1}$, which gives the **Anderson-Darling statistic**. The Anderson-Darling statistic,

$$\mathsf{A}^2 = n \int_0^1 \frac{(\hat{F}_n(t) - t)^2}{t(t-1)}\, dt,$$

was introduced by Anderson and Darling [1] in 1954, and it puts more weight on small and large p-values when compared with the Cramér-von Mises statistic. For statistical intuition, one can think of the Anderson-Darling statistic as "averaging" the squared z-score over t. This is because, under the global null, $n\hat{F}_n(t) \sim \text{Bin}(n, t)$, and thus $\text{Var}(\hat{F}_n(t)) \propto t(1-t)$, so the integrand $(\hat{F}_n(t) - t)^2[t(t-1)]^{-1}$ is similar to a squared z-score.

Notice that, since $\hat{F}_n$ is piecewise constant, we can explicitly compute the Anderson-Darling statistic A^2 as

$$\begin{aligned}
\mathsf{A}^2 &= n \int_0^1 \frac{(\hat{F}_n(t) - t)^2}{t(t-1)}\, dt \\
&= n \left(\int_0^{p_{(1)}} \frac{t^2}{t(t-1)}\, dt + \sum_{k=1}^{n-1} \int_{p_{(k)}}^{p_{(k+1)}} \frac{(k/n - t)^2}{t(t-1)}\, dt + \int_{p_{(n)}}^1 \frac{(1-t)^2}{t(t-1)}\, dt \right) \\
&= n \Bigg(p_{(1)} + \log(1 - p_{(1)}) \\
&\qquad + \sum_{k=1}^{n-1} \left[(p_{(k+1)} - p_{(k)}) + \left(1 - \frac{2k}{n}\right) \log\left(\frac{p_{(k+1)} - 1}{p_{(k)} - 1} \right) + \frac{k^2}{n^2} \log\left(\frac{p_{(k)}(p_{(k+1)} - 1)}{p_{(k+1)}(p_{(k)} - 1)} \right) \right] \\
&\qquad + 1 - p_{(n)} + \log(p_{(n)}) \Bigg) \\
&= -n - \sum_{k=1}^{n} \frac{2k-1}{n} \left[\log(p_{(k)}) + \log(1 - p_{(n+1-k)}) \right]. \qquad (3.2)
\end{aligned}$$

[2] A quadratic statistic is thus akin to a weighted L^2-norm of $\hat{F}_n(t) - t$, while the KS statistic is like the L^∞-norm.

This formula lets us connect the Anderson-Darling test to test statistics we have seen previously. Recall that Fisher's test statistic is $T_F = -2\sum_{i=1}^{n}\log(p_i)$, and Pearson's test statistic is $T_P = 2\sum_{i=1}^{n}\log(1-p_i)$. The formula (3.2) shows that the Anderson-Darling statistic is a combination of Fisher's test and Pearson's test. Compared to Fisher's test, the Anderson-Darling test assigns greater weight to the p-values that are in the bulk because it re-weights the p-values depending on their rank, something that Fisher's test does not do. This alleviates the high sensitivity to small p-values that Fisher's test experiences.

3.3.3 Tukey's Second-Level Significance Testing: Higher-Criticism Statistic

As we have seen, the Kolmogorov-Smirnov test looks for the maximum distance between the empirical CDF and its expected value under the global null hypothesis, while the Anderson-Darling test integrates the differences instead. We now combine the two approaches.

When testing n hypotheses at level α, we would expect $n\alpha$ tests to be significant, while the observed number of significant tests would be $n\hat{F}_n(\alpha)$ and the standard deviation would be given by $\sqrt{n\alpha(1-\alpha)}$. Thus, as previously, we can construct a z-score, and the overall significance at level α would be

$$\frac{(\#\text{ significant tests at level }\alpha) - \text{expected}}{\mathsf{SD}} = \frac{n\hat{F}_n(\alpha) - n\alpha}{\sqrt{n\alpha(1-\alpha)}}.$$

According to Donoho and Jin [2], Tukey [9] proposed in his class notes at Princeton to use a second-level significance test. Tukey's test combines the Kolmogorov-Smirnov test and the Anderson-Darling test by taking the maximum of the difference $\hat{F}_n(t) - t$, as in the Kolmogorov-Smirnov test, but also standardizes the difference like in the Anderson-Darling test. Specifically, Tukey proposed using the **higher-criticism statistic**.

Definition 4. The higher-criticism statistic is

$$\mathsf{HC}_n^* = \max_{0<\alpha\le\alpha_0} \frac{\hat{F}_n(\alpha) - \alpha}{\sqrt{\alpha(1-\alpha)/n}}.$$

Theoretical Analysis

Donoho and Jin [2], based on Jin [5] and Ingster [4], provide a theoretical anaylsis of Tukey's higher-criticism statistic. The higher-criticism statistic "scans" across the significance levels for departures from H_0. Hence, a large value of HC_n^* indicates significance of an overall body of tests. To understand the power of the higher-criticism statistic and to compare it to Bonferroni's method, we will next study sparse mixtures.

3.4 Sparse Mixtures

Donoho and Jin [2], based on Jin [5] and Ingster [4], consider n tests of $H_{0,i}$ vs. $H_{1,i}$ with independent test statistics X_i. In the original model, the hypotheses are

$$\begin{aligned} H_{0,i} &: X_i \sim N(0,1) \\ H_{1,i} &: X_i \sim N(\mu_i, 1) \quad \mu_i > 0. \end{aligned}$$

We are interested in possibilities within H_1 with a small fraction of non-null hypotheses. Rather than directly assuming that there is some amount of nonzero means under H_1, we assume that our samples follow a mixture of $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$, with $\mu > 0$ fixed and with some mixture parameter ε. This simple model with equals means can be written as

$$\begin{aligned} H_{0,i} &: X_i \overset{\text{i.i.d.}}{\sim} N(0,1) \\ H_{1,i} &: X_i \overset{\text{i.i.d.}}{\sim} (1-\varepsilon)N(0,1) + \varepsilon N(\mu,1). \end{aligned}$$

The expected number of non-nulls under H_1 is $n\varepsilon$. If $\varepsilon = 1/n$, then the above would become the "needle in a haystack" problem: on average, there would be one coordinate with μ nonzero.

If ε and μ were known, then the optimal test would be the likelihood ratio test. The likelihood ratio test under the sparse mixture model is

$$L = \prod_{i=1}^{n} \left[(1-\varepsilon) + \varepsilon e^{\mu X_i - \mu^2/2} \right].$$

The asymptotic analysis of Ingster [4] and Jin [5] specifies the dependency of ε and μ on n as follows:

$$\begin{aligned} \varepsilon_n &= n^{-\beta} \quad \frac{1}{2} < \beta < 1 \\ \mu_n &= \sqrt{2r \log n} \quad 0 < r < 1 \end{aligned}$$

The parameter β controls the set of non-nulls in the range 1 to $\sqrt{n}$, and thus the sparsity of the alternatives, while r paramterizes the mean shift. If β were large, then our problem would be very sparse, while if β were small, it would be mildly sparse. If $r = 1$, then we get the detection threshold we have seen for Bonferroni. Hence, the "needle in a haystack" problem corresponds to $\beta = 1$ and $r = 1$.

Ingster [4] and Jin [5] find that there is a threshold curve for r of the form

$$\rho^*(\beta) = \begin{cases} \beta - 1/2 & 1/2 < \beta < 3/4 \\ (1 - \sqrt{1-\beta})^2 & 3/4 \le \beta \le 1 \end{cases}$$

such that

1. If $r > \rho^*(\beta)$, then we can adjust the NP test to achieve

$$\mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \to 0;$$

2. if $r \leq \rho^*(\beta)$, then for *any* test,

$$\liminf_n \mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \geq 1.$$

Unfortunately, we generally cannot use the NP test since we do not know ε or μ. However, Donoho and Jin [2], based on Ingster [4] and Jin [5], show that Tukey's higher-criticism statistic, which does not require knowledge of ε or μ, *asymptotically* achives the optimal detection threshold, with

$$p_i = \overline{\Phi}(X_i) = \mathbb{P}(\mathcal{N}(0,1) > X_i), \quad \text{and} \quad \mathsf{HC}_n^* = \max_{\alpha \leq \alpha_0} \frac{\sqrt{n}(\hat{F}_n(\alpha) - \alpha)}{\sqrt{\alpha(1-\alpha)}}.$$

To better understand the results, Figure 3.3 visualizes the detection thresholds for NP and Bonferroni.

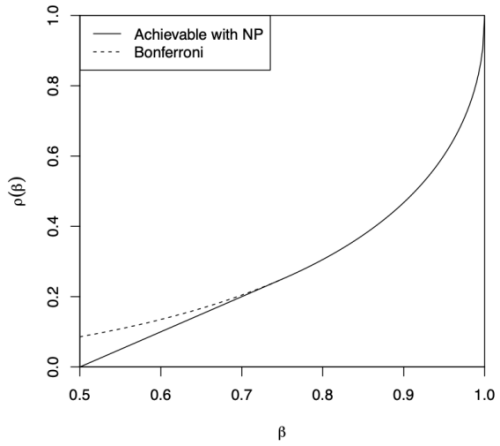

Figure 3.3. Detection thresholds for NP and Bonferroni tests.

If the amplitude of the signal is above the solid black curve (achievable with NP), then the NP test has full power, that is, we asymptotically separate. However, if it is below the curve, we asymptotically merge, that is, every test is no better than flipping a coin. The dashed black curve in Figure 3.3 shows the detection threshold for Bonferroni. Bonferroni's method achieves the optimal threshold for $\beta \in [3/4, 1]$, corresponding to the sparsest setting, but has suboptimal threshold if $\beta \in [1/2, 3/4]$, which has less sparsity. This is seen in the figure as the dashed Bonferroni curve and the solid NP curve align for $\beta \geq 3/4$, but separate below. Hence, in the area between the Bonferroni and the NP curve, the NP test has full power, while Bonferroni is no better than a coin toss.

Bonferroni's threshold for $1/2 \leq \beta \leq 1$ is

$$\rho_{\text{Bon}}(\beta) = (1 - \sqrt{1-\beta})^2.$$

Bonferroni is powerless if $r < \rho_{\text{Bon}}$. Bonferroni correctly detects non-nulls if the maximum of non-nulls is greater than that of nulls, i.e., roughly

$$\begin{aligned}\max_{\text{non null}} X_i &\simeq \sqrt{2r\log n} + \sqrt{2\log n^{1-\beta}} > \sqrt{2\log n} \\ &\iff \sqrt{r} + \sqrt{1-\beta} > 1 \\ &\iff r > (1-\sqrt{1-\beta})^2 = \rho_{\text{Bon}}(\beta).\end{aligned}$$

By comparison, the higher-criticism test rejects when HC_n^* is large, i.e., when the p-values tend to be a bit too small. Next, we will consider "how small" the p-values should be, which will be discussed in more detail in the next lecture.

For now, consider the empirical process

$$W_n(t) = \frac{\sqrt{n}(\hat{F}_n(t) - t)}{\sqrt{t(1-t)}},$$

where $W_n(t)$ converges in distribution to $\mathcal{N}(0,1)$ for each t. Empirical process theory tells us that

1. $\{\sqrt{n}(F_n(t) - t)\}_{0 \le t \le 1}$ converges in distribution to a Brownian bridge, and
2. $\max_{1/n \le t \le \alpha_0} W_n(t)/\sqrt{2\log\log n} \xrightarrow{p} 1$

This suggests the threshold $\sqrt{2\log\log n}$ for the HC statistic.

Theorem 3 (Donoho and Jin [2])**.** If we reject when $\mathsf{HC}_n^* \ge \sqrt{(1+\varepsilon)2\log\log n}$ then for any alternative H_1 in which $r > \rho^*(\beta)$,

$$\mathbb{P}_0(\text{Type I Error}) + \mathbb{P}_1(\text{Type II Error}) \to 0$$

In practice, however, the asymptotic approximation to the critical value of the higher-criticism statistic may not be accurate in finite samples. Furthermore, in finite samples, the behavior of the process $W_n(t)$ may not be well approximated by a Brownian bridge for small values of t; the process $W_n(t)$ is heavy-tailed near 0. In particular, for n large and t small, a $\text{Poi}(np)$ distribution provides a more accurate approximation to $\hat{F}_n(t)$ than a suitably centered and scaled Gaussian. As a result, the tails of $W_n(t)$ will be much heavier for t close to zero than for t away from zero, and so the behavior of $\sup_{1/n \le t \le \alpha_0} W_n(t)$ will be highly dependent on the behavior of $W_n(t)$ for t close to zero, which we can see in Figure 3.4. Therefore, critical values computed with simulation may be conservative and result in a test that performs more similarly to Bonferroni's method than to a test using Fisher's statistic.

Variations of Tukey's higher-criticism test that, in part, adjust for the heavy tails near 0 have been proposed and recently studied. One variation is the **Berk-Jones statistic**, which standardizes the binomial counts using a log-likelihood ratio transformation rather than a normal approximation.

Definition 5. The Berk-Jones statistic is given by

$$\mathsf{BJ}_n^+ = \max_{1 \le i \le n/2} nD(p_{(i)}, i/n),$$

where $D(p_0, p_1) = p_0 \log(p_0/p_1) + (1-p_0)\log((1-p_0)/(1-p_1))$ gives the Kullback-Leibler distance between $\text{Ber}(p_0)$ and $\text{Ber}(p_1)$ distributions.

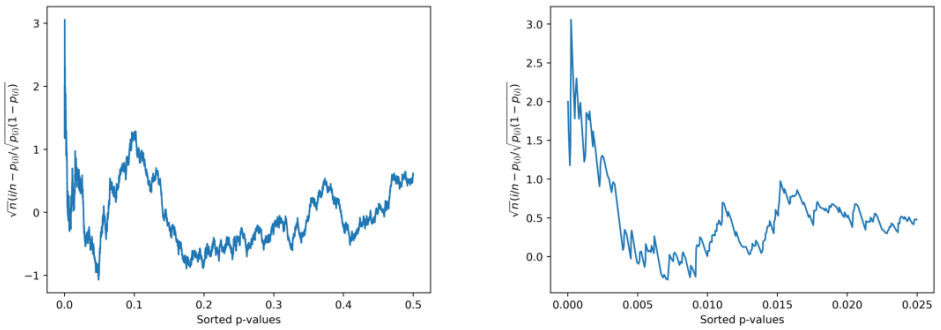

Figure 3.4. Finite sample behavior of $W_n(t)$.

The Berk-Jones statistic also achieves the optimal detection boundary asymptotically, and it produces subexponential tails under H_0, regardless of t. Walther ran simulations to compare the performance of HC_n^* and BJ_n^+ and plotted the results in the graphs shown in Figure 3.5. The plots demonstrate that HC_n^* outperforms BJ_n^+ when $\beta \in [3/4, 1]$, the sparsest setting, while BJ_n^+ is better when β is small.

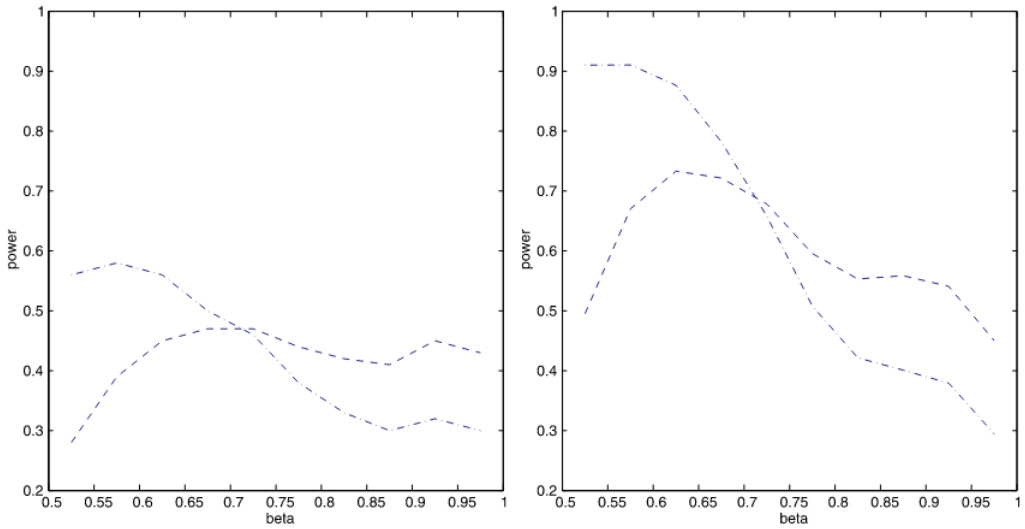

Figure 3.5. Power of HC_n^* (dashed) and BJ_n^+ (dash-dot) as a function of the sparsity parameter β. The left plot shows power for sample size $n = 10^4$, the right plot for $n = 10^6$. In the range $\beta \in (1/2, 3/4)$, we see that HC_n^* has low, but increasing power, while BJ_n^+ has high, but decreasing power. Their power curves cross at $\beta = 3/4$, and in the range $\beta \in [3/4, 1]$, we see that HC_n^* outperforms BJ_n^+

.

Another variation, proposed by Walther [10], reintroduces an element of integration. This variation combines the power of HC_n^* when $\beta \in [3/4, 1)$ by looking at the smallest p-value (namely, by performing the likelihood ratio test over the interval $(0, p_{(1)}]$), and the power of BJ_n^+ when $\beta \in (1/2, 3/4)$ by employing a likelihood ratio test on the interval $(0, n^{-4(\beta-1/2)})$. This gives the **average likelihood ratio (ALR)**.

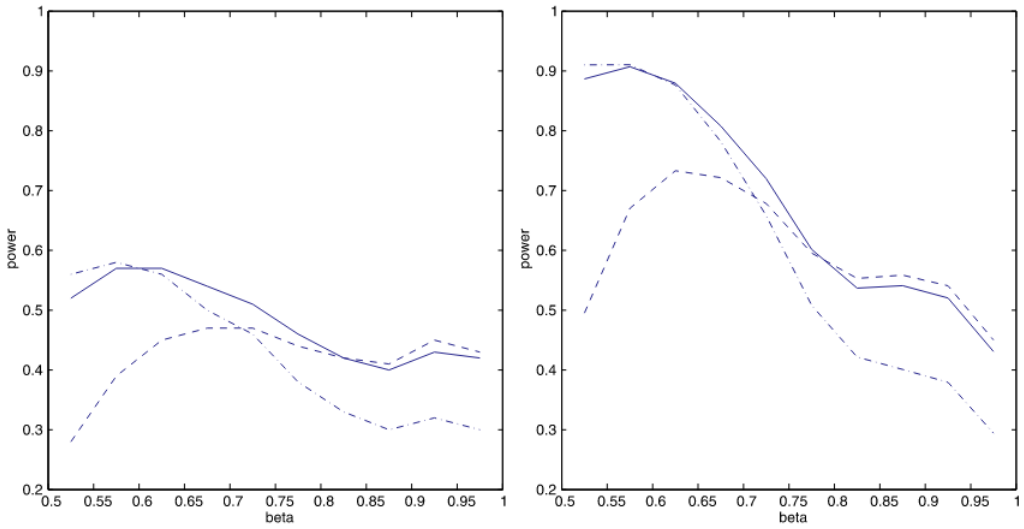

Figure 3.6. Power of ALR (solid), HC_n^* (dashed), and BJ_n^+ (dash-dot) as a function of the sparsity parameter β. The left plot shows power for sample size $n = 10^4$, the right plot for $n = 10^6$.

Definition 6. The average likelihood ratio is given by

$$\mathsf{ALR} = \sum_{i=1}^{n} w_i \mathsf{LR}_i,$$

where $\mathsf{LR}_i = \exp\left(nD(p_{(i)}, i/n)\right)$ and the weights $w_i = \frac{1}{2i\log(n/3)}$.

As with HC_n^* and BJ_n^+, ALR also achieves the optimal detection threshold asymptotically. Plotting the power of ALR against HC_n^* and BJ_n^+ gives Figure 3.6. We see that it achieves essentially the better of HC_n^* and BJ_n^+ for any β.

Bibliography

[1] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954.

[2] David Donoho and Jiashun Jin. Higher criticism for detecting spare homogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[3] G Eklund. Massignifikansproblemet. *Unpublished seminar papers, Uppsala University Institute of Statistics*, 1961.

[4] Yu I Ingster. Minimax detection of a signal for l_n^p-balls. *Mathematical Methods of Statistics*, 7:401–428, 1999.

[5] Jiashun Jin. *Detecting and estimating sparse mixtures*. PhD thesis, 2003.

[6] Pascal Massart. The tight constant in the dvortezky-kiefer-wolowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

[7] Sanat K Sarkar and Chung-Kuei Chang. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Assocation*, 92(440):1601–1608, 1997.

[8] R J Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[9] J W Tukey. T13 n: The higher criticism. *Course Notes, Statistics 411, Princeton Univ.*, 1976.

[10] Guenther Walther. The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. *Institute of Mathematical Statistics Collections*, 9:317–326, 2013.