HIGHER CRITICISM FOR HETEROGENEOUS MIXTURES

by

David Donoho
Department of Statistics
Stanford University

Jiashun Jin
Department of Statistics
Stanford University

Technical Report No. 2002-12
July 2002

This research was supported in part by National Science Foundation grants DMS-00-77261 and DMS-95-05151

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Higher Criticism for Heterogeneous Mixtures

David Donoho and Jiashun Jin
Department of Statistics
Stanford University
July 8, 2002

Abstract

Higher Criticism, or second-level significance testing, is a multiple comparisons concept mentioned in passing by Tukey (1976). It concerns a situation where there are many independent tests of significance and one is interested in rejecting the joint null hypothesis. Tukey suggested to compare the fraction of observed significances at a given $\alpha$-level to the expected fraction under the joint null, in fact he suggested to standardize the difference of the two quantities and form a z-score; the resulting z-score tests the significance of the body of significance tests.

We consider a generalization, where we maximize this z-score over a range of significance levels $0 < \alpha < \alpha_0$. We are able to show that the resulting Higher Criticism statistic is effective at resolving a very subtle testing problem: testing whether $n$ normal means are all zero versus the alternative that a small fraction is nonzero.

The subtlety of this ‘sparse normal means’ testing problem can be seen from work of Ingster (1999) and Jin (2002), who studied such problems in great detail. In their studies, they identified an interesting range of cases where the small fraction of nonzero means is so small that the alternative hypothesis exhibits little noticeable effect on the distribution on the p-values either for the bulk of the tests or for the few most highly significant tests. In this range, when the amplitude of nonzero means is calibrated with the fraction of nonzero means, the likelihood ratio test for a precisely-specified alternative would still succeed in separating the two hypotheses.

We show that the higher criticism is successful throughout the same region of amplitude vs. sparsity where the likelihood ratio test would succeed. Since it does not require a specification of the alternative, this shows that Higher Criticism is in a sense optimally adaptive to unknown sparsity and size of the non-null effects. While our theoretical work is largely asymptotic, we provide simulations in finite samples, and suggest some possible applications. We also show that Higher Criticism works well over a range of non-Gaussian cases.

Keywords: Multiple Comparsions, Combining Many p-values, Sparse Normal Means, Thresholding, Normalized Empirical Process.

Acknowledgements: The authors would like to thank Yoav Benjamini and Iain Johnstone for extensive discussions, references, and encouragement, they also would like to thank Howard Wainer, Betsy Becker, Henry Braun, Art Owen and David Siegmund for useful references and pointers. This work has been partially supported by National Science Foundation grants DMS 00-77261 and DMS 95-05151.

1 Introduction

In his Class Notes for Statistics 411 at Princeton University in 1976 [23], John Tukey introduced the notion of the Higher Criticism by means of a story. A young psychologist administers many hypothesis tests as part of a research project, and finds that, of 250 tests, 11 were significant at the 5% level. The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher (Tukey himself?) suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance. In that sense, finding only 11 significant results is actually somewhat disappointing!

Tukey used this story as a way to make vivid the notion of the higher criticism of such situations of multiple testing. He then proposed a sort of second-level significance testing, based on the statistic

$$HC_{05,n} = \sqrt{n}[(\text{Fraction Significant at .05}) - .05]/\sqrt{.05 \times .95},$$

and suggested that values (say) 2 or greater indicate a kind of significance of the overall body of tests. (The same statistic was proposed and applied in a psychometric trial by Brożek and Tiede [6] even earlier; without the catchy name).

Although Tukey's discussion turned to other topics at that point, we may, if we like, imagine that it had continued in this vein. We might then consider, not only significances at the .05 level, but perhaps at all levels between (say) 0 and $\alpha_0$, and to define

$$HC_n^* = \max_{0 < \alpha \leq \alpha_0} \sqrt{n}[(\text{Fraction Significant at } \alpha) - \alpha]/\sqrt{\alpha \times (1 - \alpha)}.$$

In this paper we will analyze a statistic of this kind in a setting where there are a small fraction of non-null hypotheses and derive an adaptive optimality for it.

In our setting, there are $n$ independent tests of unrelated hypotheses, $H_{0,i}$ vs $H_{1,i}$, where the test statistics $X_i$ obey

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

In the overwhelming majority of the tests, the corresponding null hypothesis is true (i.e. the corresponding normal mean $\mu_i = 0$), but some small fraction may concern tests where the null hypothesis is false (and so the mean $\mu_i > 0$). So the fraction of false null hypotheses, if nonzero, is small. Can we tell reliably whether the fraction is actually zero or not?

We mention three application areas where situations like this might arise.

- **Early Detection of Bio-weapons use.** Suppose that there are $n$ observational units in a certain geographical region, and for each one we have a $z$-score associated with the presence of a certain symptom at rates higher than background. If a bio-weapon has been used in that region, then in early stages we don’t expect all observational units to be affected, we don’t know which ones might be affected, and we don’t want to wait until some observational unit begins to display wildly elevated rates. We want to detect while a small fraction begin to show individually significant but no unit yet shows jointly significant results.

- **Detection of Covert Communications.** In a signals intelligence setting, we suppose that a small fraction of the signal spectrum in a certain situation may be used for covert communications, which would mean that a few frequencies exhibit increased power. However, we don’t know what frequencies those might be, and the specific frequencies being used might change randomly from one epoch to another, so that we never get a very definite indication that we are seeing definitely increased power in any one specific frequency. Nevertheless, we might still want to detect the presence of a small fraction of frequencies with slightly increased power.

- **Meta-Analysis with Heterogeneity.** We have results from $n$ experiments testing a certain treatment. It turns out that an unidentified experimental factor is crucial to success,
but that only in a small fraction of experiments is this factor fortuitously chosen so that
the experimental performance follows a non-null distribution. Can we reliably detect the
presence of a small fraction of well-laid out experiments among many hopeless ones, when
we don’t know which ones may be well-laid out?

1.1 The Model, and the Asymptotic Detection Boundary

Translating our problem into precise terms, we begin by scrutinizing a special case where all the
nonzero $\mu_i$ are equal, and we can then model our data as providing $n$ iid observations from one
of two possible situations:

$$H_0 : X_i \overset{i.i.d.}{\sim} N(0,1), \quad 1 \leq i \leq n,$$

$$H_1^{(n)} : X_i \overset{i.i.d.}{\sim} (1 - \epsilon)N(0,1) + \epsilon N(\mu, 1), \quad 1 \leq i \leq n.\quad (1.3)$$

Here, $H_0$ denotes the global intersection null hypothesis, and $H_1^{(n)}$ denotes a specific element
in its complement. Under $H_1^{(n)}$, a fraction $\epsilon$ of the data comes from a normal with common
non-null mean. Here $\epsilon = \epsilon_n$ and the mean $\mu = \mu_n$ will be chosen to make the problem very hard,
but (just barely) still solvable.

Obviously, in this situation, with $\epsilon_n$ and $\mu_n$ fixed and known, the optimal procedure is simply
the likelihood ratio test; a careful analysis of its performance in [17] (compare also [14, 13], etc.)
tells us the following. Suppose we let $\epsilon_n = n^{-\beta}$ for some exponent $\beta \in (1/2, 1)$, so that the
fraction of nonzero means is small but not vanishingly small. In this range the number of nonzero means is too small to be noticeable in any sum which is in expectation of order $n$, so it
cannot noticeably affect the behavior of the bulk distribution of the $p$-values. Let

$$\mu_n = \sqrt{2r \log(n)}, \quad 0 < r < 1.\quad (1.5)$$

As $\mu_n < \sqrt{2 \log(n)}$, the nonzero means are, in expectation, smaller than the largest $X_i$ coming
from the true component null hypotheses, so the nonzero means cannot have a visible effect on the
upper extremes. Clearly, this is a rather subtle testing problem.

It turns out that there is a threshold effect for the likelihood ratio test: the sum of Type I and
Type II errors tends to zero or one depend on whether $\mu$ exceeds a so-called detection boundary
or not. In detail, there is a function $\rho^*(\beta)$ so that

- If $r > \rho^*(\beta)$, $H_0$ and $H_1^{(n)}$ separate asymptotically,
- If $r < \rho^*(\beta)$, $H_0$ and $H_1^{(n)}$ merge asymptotically.

In short $\rho^*(\beta)$ defines a precise demarcation between what is possible and impossible in this
problem, i.e. how big the nonzero effect must be to be detectable as a function of the rarity of nonzero effects. Hence the term detection boundary. In detail, [17, 14],

$$\rho^*(\beta) = \begin{cases} 
\beta - \frac{1}{2}, & 1/2 < \beta \leq 3/4, \\
(1 - \sqrt{1 - \beta})^2, & 3/4 < \beta < 1.
\end{cases}\quad (1.6)$$

If we think of the $(r, \beta)$ plane, $0 < r < 1$, $1/2 < \beta < 1$, we are saying that throughout the region $r > \rho^*(\beta)$, the alternative can be detected reliably using the likelihood ratio test (LRT).
Unfortunately, the likelihood ratio requires a precise specification of $r$ and $\beta$, and misspecification of $(r, \beta)$ may lead to failure of the LRT; see [17] for remarks. Naturally, in any practical situation we would like to have a procedure which does well throughout this whole region without knowledge of $(r, \beta)$. This is where $HC_n$ comes in.

1.2 Performance of Higher Criticism

To apply the Higher Criticism, let’s convert the individual test statistics into another form. Let
$p_i = P\{N(0,1) > X_i\}$ be the $p$-value for the $i$-th component null hypothesis, and let the $p_{(i)}$
denote the \( p \)-values sorted in increasing order, so that under the intersection null hypothesis the \( p(i) \) behave like order statistics from a uniform distribution.

With this notation, we can write

\[
HC_n^* = \max_{1 \leq i \leq \infty - n} \sqrt{n}[i/n - p(i)]/\sqrt{p(i)(1 - p(i))}.
\]

Despite the closeness of our statistic to the one in [2], note that what we are doing is not arbitrary goodness of fit, instead we are dealing with a specific kind of multiple hypothesis testing, outside of which the problem wouldn’t be interesting.

In order to use \( HC_n^* \) to conduct a level-\( \alpha \) test, we must find a critical value \( h(n, \alpha) \):

\[
P_{H_0}\{HC_n^* > h(n, \alpha)\} \leq \alpha.
\]

Adapting asymptotic theory for the normalized empirical process as in Shorack and Wellner [19] Chapter 16 gives us the following information on the size of \( h(n, \alpha) \):

**Theorem 1.1** Under the null hypothesis \( H_0 \),

\[
\frac{HC_n^*}{\sqrt{2 \log \log(n)}} \rightarrow_p 1, \quad n \rightarrow \infty.
\]  

(1.7)

It follows that for fixed \( \alpha > 0 \), \( h(n, \alpha) \approx \sqrt{2 \log \log(n)} \). For asymptotic analysis, it is convenient to consider a sequence of problems indexed by \( n \), with critical values \( \alpha_n \to 0 \). We will say that the level \( \alpha_n \to 0 \) slowly enough if

\[
h(n, \alpha_n) = \sqrt{2 \log \log(n)}(1 + o(1)).
\]

**Theorem 1.2** Consider the Higher Criticism test that rejects \( H_0 \) when

\[
HC_n^* > h(n, \alpha_n)
\]

where the level \( \alpha_n \to 0 \) slowly enough. For every alternative \( H_{2(n)}^* \) defined above where \( r \) exceeds the detection boundary \( \rho^*(\beta) \) – so that the likelihood ratio test would have full power – Higher Criticism also has full power:

\[
P_{H_{1}^*\{\text{Reject } H_0\}} \rightarrow 1, \quad n \rightarrow \infty.
\]

Roughly speaking, everywhere in the amplitude/sparsity (\( r, \beta \)) – plane where the likelihood ratio test would completely separate the two hypotheses asymptotically – the Higher Criticism will also completely separate the two hypotheses asymptotically. Of course, in the cases where the amplitude/sparsity relation falls below the detection boundary, all methods fail. More precisely, we only claim that Higher Criticism works in the interior of this region. Just at the critical point where \( r = \rho^*(1 + o(1)) \), our result says nothing; this would be an interesting (but very challenging) area for future work.

### 1.3 Which Part of The Sample Contains the Information?

Underlying our results is a set of insights about ‘where to look’ for evidence against \( H_0 \), how the evidence may be not in the ‘obvious’ place, and how the adaptation in \( HC_n^* \) automatically ensures that the best evidence will be included in making ‘case’ against \( H_0 \).

To get started, note that, in the null case, \( HC_n^* \) is closely related to well-known functionals of the standard uniform empirical process. This is because under \( H_0 \), the \( n \) \( p \)-values can be viewed as iid samples from \( U(0, 1) \).

Formalizing, given \( n \) independent random samples \( U_1, \ldots, U_n \) from \( U(0, 1) \), with empirical distribution function

\[
F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{U_i \leq t\}},
\]
the uniform empirical process is denoted by

\[ U_n(t) = \sqrt{n}[F_n(t) - t], \quad 0 < t < 1, \]

and the normalized uniform empirical process by

\[ W_n(t) = \frac{U_n(t)}{\sqrt{t(1-t)}}. \]

Note that \( W_n(t) \) is asymptotically \( N(0, 1) \) for each fixed \( t \in (0, 1) \).

Since, under \( H_0 \), the \( p \)-values are iid \( U(0, 1) \), we have, in that case, the representation

\[ HC_n^* = \max_{0 < t < \alpha_0} W_n(t); \]

note the use of \( \max \) instead of \( \sup \), since the maximum value of \( W_n(t) \) is attained.

We will extend usage below by letting \( W_n(t) \) also stand for the normalized empirical process starting from \( F_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1(p_i \leq t) \), where the \( p_i \) are the \( n \) given \( p \)-values, but which are iid \( U(0, 1) \) only in the null case. Accordingly \( W_n(t) \) will be asymptotically \( N(0, 1) \) at each fixed \( t \) only under \( H_0 \), and one anticipates a different limiting behavior under the alternative.

We now look for the value of \( t \) at which \( W_n(t) \) has the most dramatically different behavior under the null and the alternative.

We introduce the notation \( z_n(q) = \sqrt{2q \log(n)} \), for \( 0 < q \leq 1 \), and look for the value of \( q \) for which \( \text{Prob}\{X_i > z_n(q)\} \) best differentiates between null and alternative. Recall that under the alternative \( H_1^{(n)} \) the data have a sparse component with nonzero mean at \( \mu_n = z_n(\tau) \). It might seem that the most informative part of the sample would be in the vicinity of \( \mu_n \), where data from the alternative are most common, and that therefore the most informative value of \( q \) is \( q = \tau \). Surprisingly, this is not the case.

To find this most informative \( q \), we introduce some notations. Let \( p_{n,q} = \text{Prob}\{X_i > z_n(q)\} \), \( q > 0 \), and note that

\[ \max_{0 < t < \frac{1}{2}} W_n(t) = \max_{0 < q < \infty} W_n(p_{n,q}). \]

It will be immediately clear that it's only necessary to consider for \( 0 < q \leq 1 \).

Let \( N_n(q) \) count the observations exceeding \( z_n(q) \):

\[ N_n(q) = \#\{i : X_i \geq z_n(q)\}. \]

Define then

\[ V_n(q) = \frac{N_n(q) - np_{n,q}}{\sqrt{np_{n,q}(1-p_{n,q})}}. \]

In many calculations, we will find factors with polylog behavior, \( \sim \text{Const} \cdot \log^\alpha(n) \), where \( \alpha \) may be positive or negative depending on the case. When such factors are multiplying terms \( n^\gamma \) for \( \gamma \neq 0 \), the polylog factors have a weak influence on the eventual growth rate of the resulting product. To focus on main ideas, we introduce the notation \( L_n \) for a generic polylog factor, which may change from occurrence to occurrence. When we do this, we think of \( L_n \) essentially as if it were constant.

Now

\[ \text{Prob}\{X_i > z_n(q)\} = L_n \cdot n^{-q}, \quad \text{Prob}\{X_i > z_n(q)\} = L_n \cdot n^{-q}, \quad r < q \leq 1. \]

It follows that under the alternative \( H_1^{(n)} \), we have

\[ EV_n(q) = L_n \cdot \frac{n^{1-\beta} \cdot n^{-\sqrt{q-\sqrt{r}}} \cdot n^{-\sqrt{q-\sqrt{r}}}}{\sqrt{n} \cdot n^{-q}} = L_n \cdot n^{\frac{1+\beta}{2} - \sqrt{q-\sqrt{r}}} \cdot n^{-q}. \]
while, under the null, \( EV_n(q) = 0 \). The most informative value of \( q \) will optimize the growth rate of \( n^{\left(1 + \frac{\beta}{2} - \frac{1}{2} + \sqrt{r - \beta - \left(\sqrt{r} - \sqrt{r - \beta}\right)^2}\right)} \) to \( \infty \). The most informative value of \( q \) satisfies

- if \( r < \frac{1}{4} \), then \( q = 4r \) and \( EV_n(q) = L_n \cdot n^{2q - \left(\beta - \frac{1}{2}\right)} \),
- if \( r \geq \frac{1}{4} \), then \( q = 1 \) and \( EV_n(q) = L_n \cdot n^{\left(1 - \beta\right) - \left(1 - \sqrt{r}\right)^2} \).

In fact, there can be considerable latitude in choosing \( q \) so that \( EV_n(q) \) goes to \( \infty \) under \( H_{1(n)}^1 \). This requires

\[
2\sqrt{r} - \sqrt{2(r - \beta + \frac{1}{2})} < \sqrt{q} < 2\sqrt{r} + \sqrt{2(r - \beta + \frac{1}{2})}.
\]

Notice that \( r > \rho^*(\beta) \) implies \( r - \beta + \frac{1}{2} > 0 \). Within this interval, the “most informative choice” of \( q \) is the center: \( \sqrt{q} = 2\sqrt{r} \) in case \( r < \frac{1}{4} \).

Translating this into the original \( z \)-scale yields the surprise mentioned earlier. When \( r \leq \frac{1}{4} \) the most informative place on the original \( z \)-scale is not at \( \mu_n \), as we might suppose, but at \( 2\mu_n \). By going out ‘in the tails’ farther than \( \mu_n \), observations from both \( H_0 \) and \( H_{1(n)}^1 \) are becoming extremely rare, but the ones from \( H_{1(n)}^1 \) are far more frequent in a relative sense.

The story when \( r > 1/4 \) is somewhat different. There, when looking for a discrepancy, we had best look near \( \sqrt{2\log(n)} \), which is less than \( 2\mu_n \). The point is that observations from \( H_0 \) almost never get substantially larger than \( \sqrt{2\log(n)} \), so there’s no need to look much farther out in the tails.

### 1.4 Comparison to Several Multiple Comparison Procedures

The Higher Criticism is just one specific approach to combining many \( p \)-values in search of an overall test; many other tools are available from the field of multiple comparisons [25, 12] and meta-analysis [4]. How do these other tools perform?

In this section we describe several specific procedures and their range of detectability. See Figure 1.

#### 1.4.1 Range/Maximum/Bonferroni

One of the most classical, and frequently-used tools in multiple comparisons, also associated with Tukey, is the studentized range:

\[
R_n = \frac{(\max X_i - \min X_i)}{S_n},
\]
where $S_n$ is the sample standard deviation. This is frequently used in testing for homogeneity of a set of normal means, and could well be used in the setting we have defined here. For our theoretical purposes, it is convenient to analyze the simpler statistic

$$M_n = \max_i X_i,$$

where we focus attention on one-sided deviations only, and have no need to estimate the (known) standard deviation under the null. (We remark that in the field of meta-analysis, this is the same thing as combining several $p$-values by taking the minimum $p$-value [4]; another equivalent term for this is Bonferroni-based inference). The maximum statistic has a critical value $m(n, \alpha)$ which obeys

$$m(n, \alpha) \sim \sqrt{2 \log(n)}, \quad n \to \infty.$$

In comparison to $HC^*$, it follows that the test focuses entirely on whether there are observations exceeding $\sqrt{2 \log(n)}$. We again say that $\alpha_n$ goes to zero slowly enough (now for use with $M_n$) if

$$m(n, \alpha_n) \sim \sqrt{2 \log(n)}.$$

The following result summarizes the behavior of $M_n$.

**Theorem 1.3** Define

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

Suppose $r > \rho_{\text{Max}}(\beta)$, and consider a sequence of level $\alpha_n$-tests based on $M_n$ with $\alpha_n \to 0$ slowly enough. Then

$$P_{H_0^{(n)}} \{M_n \text{ rejects } H_0 \} \to 1;$$

while

$$P_{H_1^{(n)}} \{M_n \text{ rejects } H_0 \} \to 0.$$

In short, $\rho_{\text{Max}}$ defines the detection boundary for $M_n$. This compares to the ‘efficient’ boundary as follows:

$$\rho^*(\beta) = \rho_{\text{Max}}(\beta), \quad \beta \in [3/4, 1)$$

so that $M_n$ is effective in the range of very sparse alternatives; while

$$\rho^*(\beta) < \rho_{\text{Max}}(\beta), \quad \beta \in [1/2, 3/4)$$

so that $M_n$ is inefficient if $\beta < 3/4$. In particular, note that

$$0 = \rho^*(1/2) < \rho_{\text{Max}}(1/2) = (2 - \sqrt{2})^2 / 4 \approx .0858$$

We interpret this as saying that $M_n$ (and the studentized range $R_n$ as well) can be dramatically outperformed when there may be about $n^{1/2}$ nonzero means among the $n$ observations. Compare Figure 1.

### 1.4.2 FDR-controlling Methods

Recently, considerable interest has been focused on False Discovery Rate (FDR) controlling methodology for simultaneous inference. In one example of this so-called FDR approach [3], one considers, for $k = 1, 2, 3, \ldots$, the $k$ most significant $p$-values. These are compared with $\alpha_k^\beta$, where $\alpha$ is a critical value (e.g. .05). If the $p$-values in the group are all smaller than the standard for comparison, then $H_0$ is accepted at that stage. If some are larger than the standard for comparison, then $H_1^{(n)}$ is accepted, no further $k$ are considered, and that specific group
of \( k \) hypothesis is identified as containing likely non-null hypotheses. Viewed as a hypothesis testing procedure for the intersection null hypothesis, Benjamini and Hochberg [3] show that this procedure has level \( \leq \alpha \). They also show that procedure controls the FDR, which means, roughly speaking, that in expectation, a fraction of at least \( (1 - \alpha) \) of the rejected null hypotheses should be truly non-null hypotheses.

How does such a procedure behave in the current setting? We begin by pointing out that Abramovitch et al. [1] have analyzed the behavior of FDR in exactly the kind of mixture model described above in eqs (1.3)-(1.4), but where \( \mu_n \) is calibrated differently than (1.5), and they found an asymptotic minimaxity of FDR in that setting. That is, they considered a situation where one observed data \( X_i = \theta_i + Z_i \) and where the \( Z_i \) are iid \( N(0, 1) \). They supposed that only a fraction \( \epsilon_n \) of the means \( \theta_i \) might be nonzero. This is similar to our model above.

As it turns out, for our measure of performance, the behavior of FDR-controlling procedures is not different than that of the maximum \( M_n \). To articulate this, consider the detection boundary for the FDR-controlling procedure above. This is the function \( \rho_{FDR}(\beta) \) such that, if \( \mu_n = \sqrt{2r \log(n)} \), and if we use a sequence of levels \( \alpha = \alpha_n \) tending to zero slowly enough, then, for \( r > \rho_{FDR}(\beta) \) the procedure has power tending to one as \( n \to \infty \), while for \( r < \rho_{FDR} \) the procedure has power tending to zero.

**Theorem 1.4**

\[
\rho_{FDR}(\beta) = \rho_{Max}(\beta) = (1 - \sqrt{1 - \beta})^2, \quad 1/2 < \beta < 1.
\]  

(1.8)

From our discussion of the behavior of the maximum \( M_n \), the FDR-controlling procedure is effective in the range \( \beta \in [3/4, 1) \), while it is relatively inefficient for \( \beta < 3/4 \). Compare Figure 1.

### 1.5 Classical Methods for Combining \( p \)-values

In the literature of meta-analysis [4] one also faces the problem of combining several \( p \)-values to achieve an overall test of significance. In that literature, the component non-null hypotheses are all the same ("homogeneity") whereas in our discussion they vary widely ("heterogeneity"). A classical approach to combining \( p \)-values is Fisher’s method [10], which refers \( F_n = -2 \sum_{1 \leq i \leq n} \log(p_i) \) to the \( \chi^2_n \) distribution. In our setting of extreme heterogeneity, Fisher’s method is unable to function well asymptotically:

**Theorem 1.5** If \( \epsilon = n^{-\beta} \), \( \beta > \frac{1}{2} \) and \( \mu_n \leq \sqrt{2 \log(n)} \), asymptotically \( F_n \) is unable to separate \( H_1^{(n)} \) and \( H_0 \).

### 1.6 Relation to Goodness-of-fit Testing

Of course, the method we are discussing may be viewed as application of a goodness-of-fit measure, comparing the empirical distribution of \( p \)-values to the uniform distribution. As such, it may be compared to many goodness-of-fit procedures where distribution under \( H_1 \) differs from that \( H_0 \) in one tail.

Thus, Anderson and Darling [2] defined a goodness-of-fit measures which involves the maximum of the normalized empirical process. Translated into the current setting, this is very close to \( HC^* \). The main difference is that we focus attention near \( p = 0 \), while Anderson and Darling maximize over \( a < p < b \) with \( a > 0, b < 1 \).

In a similar vein, R. H. Berk and D. H. Jones [5] proposed a goodness-of-fit measure:

\[
BJ^+_n = n \cdot \max_{1 \leq i \leq n/2} K^+(i/n, p(i)),
\]  

(1.9)

where \( K^+ \) is defined by:

\[
K^+(t, x) = \begin{cases} 
  t \log \frac{\frac{t}{x}}{1 - t} + (1 - t) \log \frac{1 - x}{1 - t} & \text{if } 0 < x < t < 1, \\
  0 & \text{if } 0 \leq t \leq x \leq 1, \\
  +\infty & \text{otherwise},
\end{cases}
\]
and is derived from large deviation theory. Roughly speaking, \( K^+(t, x) \) behaves as \( \frac{1}{2}(t - x)^2 \), so it should not be surprising that for our measure of performance, the detection boundary of \( BJ^+_n \) is the same as that of \( HC^* \). To articulate this, there is a function \( \rho_{BJ}(\beta) \) such that, if \( \mu_n = \sqrt{2r \log(n)} \), and if we use a sequence of levels \( \alpha = \alpha_n \) tending to zero slowly enough, then, for \( r > \rho_{BJ}(\beta) \) the procedure has power tending to one as \( n \to \infty \), while for \( r < \rho_{FDR} \) the procedure has power tending to zero.

**Theorem 1.6** \( \rho_{BJ}(\beta) \equiv \rho^*(\beta), \frac{1}{2} < \beta < 1. \)

However \( HC^* \) is still better than \( BJ^+_n \) in important ways; we will discuss this in the Appendix, where we prove Theorem 1.6.

### 1.7 Generalizations

As we have defined it, the Higher Criticism statistic can obviously be used in a wide variety of situations; there is no need that the \( p \)-values derive from normally distributed \( Z \)-scores, for example. In consequence, numerous other settings for its deployment can be considered. We have found that in a wide variety of settings where one has data which are ‘sparsely non-null’, the \( HC^* \) statistic has an adaptive optimality.

To give the flavor of one of these, we consider a model deriving from the covert communications example mentioned earlier. The problem is one of noncooperative spread-spectrum signal detection. Here one observes \( n \) periodogram ordinates \( X_i \). In the ‘covert data absent’ case, these represent periodogram ordinates of a white noise, while in the ‘covert data present’ case a small fraction of periodogram ordinates are inflated by the presence of covert signal. The formal model takes the form:

\[
H_0 : X_i \sim \text{Exp}(2), \quad 1 \leq i \leq n, \\
H_1(n) : X_i \sim (1 - \epsilon)\text{Exp}(2) + \epsilon \chi^2(\delta), \quad 1 \leq i \leq n.
\]

(to avoid misunderstanding, \( \text{Exp}(2) \) denotes the exponential distribution with mean 2). Here the data are nonGaussian, \( n \) is large, and the sparsity parameter \( \epsilon = \epsilon_n = n^{-\beta} \) as before. The strength of the covert signal is measured by the noncentrality parameter \( \delta \) of the \( \chi^2 \) distribution, which we take as \( \delta = \delta_n = 2r \log(n) \) for an underlying amplitude parameter \( r \) having much the same interpretation as before. (In a cooperative signal detection setting, we would know a priori which of the coordinates \( X_i \) will exhibit the presence of the covert signal; in the noncooperative case, we would not know this).

We can again apply the principle of Higher Criticism in this setting, defining \( p \)-values through the component null hypotheses:

\[
p_i = \text{Prob} \{ \text{Exp}(2) > X_i \}, \quad i = 1, \ldots, n.
\]

We can also define the detection boundary for this test, as before. This is the function \( \rho_{HC,\text{Exp}}(\beta) \) such that, if \( \delta_n = 2r \log(n) \), and if we use a sequence of levels \( \alpha = \alpha_n \) tending to zero slowly enough, then, for \( r > \rho_{HC,\text{Exp}}(\beta) \) the procedure has power tending to one as \( n \to \infty \), while for \( r < \rho_{HC,\text{Exp}} \) the procedure has power tending to zero. We can also define the intrinsic detection boundary, the function \( \rho^*_{\text{Exp}}(\beta) \) such that if \( r < \rho^*_{\text{Exp}} \), the two hypotheses merge asymptotically, while if \( r > \rho^*_{\text{Exp}} \), the two hypotheses separate asymptotically.

**Theorem 1.7**

\[
\rho_{HC,\text{Exp}}(\beta) = \rho^*_{\text{Exp}}(\beta) = \begin{cases} 
\beta - \frac{1}{2}, & \frac{1}{2} < \beta < 3/4, \\
\left( 1 - \sqrt{1-\beta} \right)^2, & 3/4 < \beta < 1.
\end{cases} \tag{1.10}
\]

In words, the higher criticism statistic achieves the optimal detection region in the \((r, \beta)\) plane; interestingly, this region is the same as we had in the Gaussian case, see Figure 3 also compare to Figure 1.

Other nonGaussian settings are discussed in Sections 5.1 and 5.2 below. In each case, the higher criticism statistic achieves the (interior of) the optimal detection region.
1.8 Contents of this Paper

In this paper, we establish the key results referred to so far, and then consider various generalizations. Section 2 develops Theorems 1.1-Theorem 1.2; Section 3 discusses \( HC^+ \) as a variant of \( HC_n^* \) but with better performance in finite samples; Section 4 describes some simulation experiments. Section 5 considers numerous non-Gaussian settings and shows consistently the superiority of Higher Criticism over Bonferroni and other ideas from the field of multiple comparisons, the proof of Theorem 1.7 is also in this section. Section 6, the Appendix provides proofs of Theorems 1.3, 1.4, 1.5 and 1.6 as well as Lemma 5.1.

2 Main Results

We now turn to proofs of the main results of this paper.

2.1 Proof of Theorem 1.1

The idea behind Theorem 1.1 is to simply apply known results from the theory of empirical processes. We recall the normalized uniform empirical process \( W_n(t) \) introduced in Section 1.3, and remind the reader that under \( H_0 \):

\[
HC^* = \max_{0 < t < 1} W_n(t).
\]

The normalized empirical process has been studied carefully by a number of authors, and a summary of results can be found in [19], Chapter 16. There, in equation (16.20) they show that

\[
\frac{\max_{0 < t < a_0} W_n(t)}{\sqrt{2 \log \log(n)}} \to_p 1, \quad n \to \infty,
\]

by an argument that depends on work of Jaeschke (1979) [16]. This in turn depends on the approximation of \( W_n \) by a Brownian Bridge, and on a result of Darling and Erdös (1956) [7] which says that if \( B(t) \) is standard Brownian motion starting at \( B(0) = 0 \), then

\[
\sup_{[1, u]} \frac{B(t)}{\sqrt{t}} \cdot \frac{1}{\sqrt{2 \log \log u}} \to_p 1, \quad u \to \infty.
\]

Of course (2.11) implies Theorem 1.1. \( \square \)

2.2 Proof of Theorem 1.2

We begin with a simple observation.

**Lemma 2.1** Let \( X_1, \ldots, X_n \) be i.i.d Bernoulli(\(\pi_n\)) and \( a_1, \ldots, a_n \) be a sequence of real numbers. If \( n\pi_n \to \infty \), and \( a_n/\sqrt{n\pi_n} \to \infty \), then \( \lim_{n \to \infty} P(\sum_{i=1}^n (X_i - \pi_n) \leq -a_n) = 0 \).

**Proof.** Since \( \text{Var}(\sum_{i=1}^n X_i) \leq n\pi_n \), use Chebyshev's inequality:

\[
P\{\sum_{i=1}^n (X_i - \pi_n) \leq -a_n\} \leq \frac{n\pi_n}{a_n^2} \to 0, \quad n \to \infty.
\]

\( \square \)

To prove the Theorem itself, we note that it is enough to show

\[
\lim_{n \to \infty} P_{H^*_n}(HC^*_n \leq \sqrt{4 \log \log(n)}) = 0.
\]

(2.1)

Now, recalling the definition of \( V_n(q) \), we have

\[
HC^*_n \geq \sup_{0 < q \leq 1} V_n(q).
\]
For $0 < q \leq 1$, recall $P_{n,q} = P\{N(0,1) \geq z_n(q)\}$, and we also let

$$p'_{n,q} = P\{(1 - \epsilon_n)N(0,1) + \epsilon_nN(\mu_n,1) \geq z_n(q)\}.$$  

We now consider two cases. First, suppose that $r > \rho(\beta)$ and $r \geq \frac{1}{4}$; then the hypothesis is detectable, if at all, merely by looking at the maximum of the $X_i$. Note that $\sqrt{r} + \sqrt{1 - \beta} > 1$. Now as $V_n(1) \leq \sup_0<q\leq 1 V_n(q)$ and $V_n(1) = D (N_n(1) - n\pi_n,1)/\sqrt{n\pi_n(1 - n\pi_n)},$

$$P_{H_1^{(n)}}\{\sup_0<q\leq 1 V_n(q) \leq \sqrt{4\log\log(n)}\} \leq P_{H_1^{(n)}}\{V_n(1) \leq \sqrt{4\log\log(n)}\}$$

$$\leq P_{H_1^{(n)}}\{N_n(1) \leq \sqrt{n\pi_n}\sqrt{4\log\log(n)} + n\pi_n\}. $$

Under $H_1^{(n)}$, $N_n(1)$ is sum of independent Bernoulli$(p'_{n,1})$, by direct calculations,

$$p_{n,1} = o(1),$$

$$p'_{n,1} = \frac{1}{\sqrt{4\pi\log(n)(1 - \sqrt{r})}} n^{-\beta -(1-\sqrt{r})^2} (1 + o(1)).$$

With Lemma 2.1 in mind, let $\pi_n = p'_{n,1}$ and

$$a_n = np'_{n,1} - [\sqrt{n\pi_n}\sqrt{4\log\log(n)} + n\pi_n] = np'_{n,1} (1 + o(1)), $$

so $n\pi_n \to \infty$ and $a_n/\sqrt{n\pi_n} \to \infty$, the desired result (2.1) in this case follows from Lemma 2.1.

In the second case, we have $r > \rho(\beta)$ and $r < \frac{1}{4}$. It will turn out that the hypothesis is detectable based on $HC^*$ but not on the maximum of the $X_i$. Note that $\beta + r < 1$.

$$P_{H_1^{(n)}}\{\sup_0<q\leq 1 V_n(q) \leq \sqrt{4\log\log(n)}\} \leq P_{H_1^{(n)}}\{V_n(4r) \leq \sqrt{4\log\log(n)}\}$$

$$\leq P_{H_1^{(n)}}\{N_n(4r) \leq \sqrt{n\pi_{4r}}\sqrt{4\log\log(n)} + n\pi_{4r}\}.$$ 

Under $H_1^{(n)}$, $N_n(4r)$ is sum of $n$ independent Bernoulli$(p'_{n,4r})$, and

$$p_{n,4r} = \frac{1}{4\sqrt{\pi r}\log(n)} n^{-4r} (1 + o(1)),$$

$$p'_{n,4r} = p_{n,4r} + \frac{1}{\sqrt{4\pi r\log(n)}} n^{-(\beta + r)} (1 + o(1)).$$

With Lemma 2.1 in mind, let $\pi_n = p_{n,4r}$ and

$$a_n = n[p'_{n,4r} - p_{n,4r}] - \sqrt{n\pi_{4r}}\sqrt{4\log\log(n)}.$$ 

Direct calculation shows

$$\frac{a_n}{\sqrt{n\pi_n}} = O(n^{r-(\beta - \frac{1}{2})}/(\log(n))^{\frac{1}{2}}), \quad \text{if } \beta > 3r,$$

$$\frac{a_n}{\sqrt{n\pi_n}} = O(n^{\frac{1}{2}[-(\beta + r)]}/(\log(n))^{\frac{1}{2}}), \quad \text{if } \beta \leq 3r,$$

so $n\pi_n \to \infty$, $a_n/\sqrt{n\pi_n} \to \infty$, the desired result (2.1) follows for this case from Lemma 2.1.

\[ \square \]

3 A Refinement

In the most challenging cases, where $0 < r < 1/4$, the analysis above tells us that the real information about the presence of $H_1^{(n)}$ is located away from the extreme values, and so, perhaps unexpectedly, the few smallest $p_i$ are not relevant to the detection problem.
This suggests that we will still be able to reach the full interior of the detection region if we work with a modified statistic involving the maximum over all \( p \)-values \( \geq 1/n \), which is

\[
HC^*_n = \max_{1 < i \leq n/2, \ p(i) \geq 1/n} \sqrt{n} \frac{i/n - p(i)}{\sqrt{p(i)(1 - p(i))}}.
\]

Of course this restriction is not necessary from the viewpoint of asymptotic analysis, since our results show that \( HC^*_n \) is effective without any adjustment. However, our experience with moderate-sized samples shows this adjustment to be quite valuable.

The empirical process viewpoint is helpful. Under \( H_0 \), \( HC^*_n \) behaves as

\[
\sup_{0 < t < 1/2} W_n(t).
\]

For comparison, \( HC^+_n \) behaves as

\[
\sup_{1/n < t < 1/2} W_n(t).
\]

The limitation of the range in (3.2) is important. The quantity in (3.1) can be very much dominated by behavior in the vicinity of 0, or more particularly, by the smallest observation. Since the smallest observation (smallest \( p \)-value) is not where the information for detection resides, this emphasis is misplaced.

Speaking quantitatively, the statistic can have a heavy-tailed distribution under the null hypothesis, as we will show in a moment. Since ordinarily, one does not want to use test statistics with heavy-tailed distributions – because tests at stringent levels would have low power in small samples – the quantitative motivation for \( HC^*_n \) is clear.

The heavy tails under the null hypothesis come from the following effect. Recall that under \( H_0 \), the minimum \( p \)-value has a distribution that is \( \text{Exp}(1/n) \). It follows that the first component in the maximum defining \( HC^*_n \),

\[
HC^*_{n,1} = \sqrt{n} \frac{1/n - p(1)}{\sqrt{p(1)(1 - p(1))}},
\]

has the asymptotic distribution

\[
HC^*_{n,1} \Rightarrow D \frac{1}{\sqrt{E}} - \sqrt{E},
\]

where \( E \) is exponentially-distributed with mean 1. Now for large \( t \),

\[
P(\frac{1}{\sqrt{E}} - \sqrt{E} > t) = P(\sqrt{E} < (\sqrt{t^2 + 4} - t)/2) \sim \frac{1}{t^2}.
\]

Hence \( HC^*_{n,1} \) has “heavy tails” under \( H_0 \).

Numerical simulations show that unusually large values of \( HC^*_n \) under the null hypothesis are most frequently caused by \( HC^*_{n,1} \), suggesting that if we restrict the range of the maximum as in \( HC^*_n \) this “heavy tail” is thinned out considerably. Numerical experiments support this analysis.

4 Simulation

We have conducted a small-scale empirical study of the performance of \( HC^*_n \) and \( HC^+_n \). Our idea was to select a few interesting \((r, \beta)\) pairs in the detectable region – above, but close to the boundary, to create samples from null and alternative – and to study the behavior of the proposed statistics.

We took \( n = 10^4, \epsilon = \frac{1}{1000}, \mu = \sqrt{2 \times 0.15 \times \log(n)} \approx 2.04 \) for the following experiment:

1. Step 1. Draw \( n \) samples from \( N(0,1) \) to represent \( H_0 \), calculate \( HC^* \) and \( HC^+ \).
Table 1: $E V_n(4r)$ under $H_1^{(n)}$ for various $n$. The parameters $(r, \beta)$ are $r = 0.1$, $\beta = 1/2$ and $r = 0.05$, $\beta = 1/2$ respectively; values of $\sqrt{2 \log \log(n)}$ are also included for comparison.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$10^6$</th>
<th>$10^7$</th>
<th>$10^8$</th>
<th>$10^9$</th>
<th>$10^{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{2 \log \log(n)}$</td>
<td>2.2916</td>
<td>2.3579</td>
<td>2.4139</td>
<td>2.4622</td>
<td>2.5046</td>
</tr>
<tr>
<td>$E V_n(4r)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r = 0.1$, $\beta = 1/2$</td>
<td>2.7582</td>
<td>3.3411</td>
<td>4.0680</td>
<td>4.9728</td>
<td>6.0976</td>
</tr>
<tr>
<td>$r = 0.05$, $\beta = 1/2$</td>
<td>1.6439</td>
<td>1.7748</td>
<td>1.9259</td>
<td>2.0982</td>
<td>2.2931</td>
</tr>
</tbody>
</table>

2. Step 2. Replace 1000 of the previous samples by same number of samples from $N(\mu, 1)$ to represent $H_1^{(n)}$, calculate $HC^*$ and $HC^+$.

3. Repeat Steps 1 and 2 100 times, and make histograms of the simulated $HC^*$ and $HC^+$.

See Figure 2 for the results. As can be seen from the theory above, detectability requires increasingly large samples as one approaches the detection boundary. In fact, depending on how close one goes to the boundary, the required sample sizes can become enormous. To have an idea how large $n$ may be needed, we consider the case $r < 1/4$, $\frac{1}{2} < \beta < \frac{3}{4}$. We recall that

$$EV_n(4r) \begin{cases} = 0, \\ \sim n^{r-(\beta-1/2)}/\sqrt{\pi \tau \log(n)}, \end{cases} \text{ under } H_0, \text{ under } H_1^{(n)}.$$  (4.1)

Numerically we calculate the values of $EV_n(4r)$ under $H_1^{(n)}$ for various $n$ with $r = 0.1$, $\beta = 1/2$ and $r = 0.05$, $\beta = 1/2$ respectively; the results are summarized in Table 1 together with the values of $\sqrt{2 \log \log(n)}$ for comparison; notice the variance of $V_n(4r)$ is roughly 1. This of course raises the issue of whether our theory adequately describes practice in small samples, and in particular whether the motivating examples of the introduction really offer credible scenarios for deployment of these results. We leave such discussion to another occasion.

The large-sample issue also raises the question of how to efficiently sample from the normal distribution with very large $n$. Our approach goes as follows.

1. Pick a small number $\epsilon$, such as $\epsilon = 10^{-3}$ or $\epsilon = 10^{-6}$.

2. Simulate samples from uniform distribution at quantiles $> 1 - \epsilon$.
   - Sample a number $K$ from Poisson distribution with mean $n \cdot \epsilon$.
   - Generate $K$ samples $(U_1, \ldots, U_K)$ from the uniform distribution on $(1 - \epsilon, 1)$.
   - Generate $K$ samples of $(z_1, \ldots, z_K)$ by letting
     $$y_i = -2 \log(1 - U_i) - \log 2\pi, \quad z_i = \sqrt{y_i - \log y_i}, \quad 1 \leq i \leq K.$$  (4.2)

   Approximately, the $(z_1, \ldots, z_K)$ can be viewed as if a sample of size $n$ has been taken from the Normal distribution, and then only the $(1 - \epsilon)n$ largest sample values were retained. Compared to brute force, the algorithm requires only $\epsilon \cdot n$ flops, rather than $n$ flops. Obviously, the accuracy of the approximation in (4.2) depends on how small $\epsilon$ is; the smaller the $\epsilon$, the more accurate the approximation, and thus more accurate the simulation.

5 Other Settings

The principle of Higher Criticism can be applied in a much wider series of situations than the Gaussian model studied so far, as we now show.
Figure 2: Histograms for $HC^*$ and $HC^+$ Top row: behavior under $H_0$. Bottom row: behavior under $H_1^{(n)}$. Left column: $HC^*$. Right column: $HC^+$. Here $n = 10^8, \epsilon = 10^{-3}, \mu = \sqrt{0.3 \log(n)} \approx 2.04$.

5.1 Chi-squared

Let $\chi^2_\nu(\delta)$ denote the usual chi-square distribution with $\nu$ degrees of freedom and non-centrality parameter $\delta$. Consider the problem of testing between these hypotheses:

$$H_0 : X_i \overset{i.i.d.}{\sim} \chi^2(0), \quad 1 \leq i \leq n, \quad (5.1)$$

$$H_1^{(n)} : X_i \overset{i.i.d.}{\sim} (1 - \epsilon)\chi^2(0) + \epsilon\chi^2(\delta), \quad 1 \leq i \leq n. \quad (5.2)$$

Here $\epsilon = \epsilon_n = n^{-\beta}$ as before. Owing to the representation of chi-squared r.v.’s as sums of squares of standard normals, the problem can be rewritten in terms of arrays of normals $z_{ij} \overset{i.i.d.}{\sim} N(0, 1)$:

$$H_0 : X_i \overset{i.i.d.}{\sim} \sum_{j=1}^{\nu} z_{ij}^2, \quad 1 \leq i \leq n, \quad (5.3)$$

$$H_1^{(n)} : X_i \overset{i.i.d.}{\sim} (1 - \epsilon)\sum_{j=1}^{\nu} z_{ij}^2 + \epsilon[(z_{i1} + \delta)^2 + \sum_{j=2}^{\nu} z_{ij}^2], \quad 1 \leq i \leq n. \quad (5.4)$$

where $z_{ij} \overset{i.i.d.}{\sim} N(0, 1)$. Roughly speaking, a small fraction of the normals may have nonzero mean, and we must base our decisions on sums of squares rather than on the normals themselves. The problem is obviously related to the Gaussian problem discussed above, only the observations with nonzero means occur in groups, and while the nonzero means are not equal within a group, the sum of squares of those means within a group must be equal.

Now obviously, if $\nu = 1$, then we are equivalently dealing with squares of individual normals, and so are just seeing a two-sided variant of the original one-sided normal testing problem considered so far. As it turns out, all the detection boundary and attainability results for the two-sided normal problem are the same as in the one-sided case.

If $\nu = 2$, we view this as modelling the covert communications problem of Section 1.5 of the introduction. Indeed, the real and imaginary components of the discrete Fourier transform of Gaussian white noise are normal and independent; the sum of squares of those two components is just the periodogram. In a frequency where there is no signal, only Gaussian noise, the periodogram has a $\chi^2_2(0)$ (or $\exp(2)$) distribution, that is precisely the exponential mentioned earlier; while in the signal-present case, the periodogram has a $\chi^2_2(\delta_k)$, where $\delta_k$ is the signal energy at that frequency.

If $\nu > 2$, we can think of agricultural trials, where a treatment is attempted with replications, and in many different blocks. The alternative hypothesis is that, in a few special blocks, the treatment has a substantial effect, but in most blocks it has no effect.

As it turns out, for any fixed, constant number of degrees of freedom, the results will be similar. Let the noncentrality parameter obey

$$\delta = \delta_n = 2r \log(n), \quad 0 < r < 1.$$
Figure 3: Three regions in $\beta - r$ plane for the $\chi^2$ model. The detection boundary separates the white and blue regions. In the yellow region, it is possible not only to detect the presence of nonzero means, but also to estimate those means.

In this problem we again have an $(r, \beta)$ plane, in which there is a region of detectability, and a detection boundary. There are two key facts:

- The optimal detection boundary $\rho^*_\chi^2(\beta)$ is the same as in the Gaussian case:

$$\rho^*_\chi^2(\beta) = \begin{cases} (1 - \sqrt{1 - \beta})^2, & 3/4 \leq \beta \leq 1, \\ \beta - 1/2, & 1/2 < \beta < 3/4. \end{cases}$$

This is proven in Jin (2002).

- $HC^*_n$ is able to separate the null and alternative hypotheses throughout the interior of the detection region, and thus its detection boundary $\rho_{HC, \chi^2}(\beta)$ obeys

$$\rho_{HC, \chi^2}(\beta) = \rho^*_\chi^2(\beta), \quad 1/2 < \beta \leq 1.$$}

The key ideas are illustrated in Figure 3.

The analysis supporting the performance of $HC^*_n$ is as follows. Define $x_n^2(q) = 2q \log(n)$. We then have

$$P\{\chi^2_n(0) > x_n^2(q)\} \sim L_n \cdot n^{-q}, \quad P\{\chi^2_n(\delta_n) > x_n^2(q)\} \sim L_n \cdot n^{-(\sqrt{q} - \sqrt{r})^2}, \quad r < q \leq 1. \quad (5.5)$$

The left-hand relation is proved as follows:

$$P\{\chi^2_n(0) > x_n^2(q)\} = \frac{2^{1-q}}{\Gamma\left(\frac{q}{2}\right)} \int_{\sqrt{2q \log(n)}}^{\infty} \rho^{q-1} e^{-\rho^2/4} d\rho = \frac{(q \log(n))^{q/2-1}}{\Gamma(\frac{q}{2})} n^{-q}(1 + o(1)).$$

It takes a little more effort to check that the second equality is also true; this follows from the following lemma.

**Lemma 5.1** Let $0 < r < q < 1, \quad \delta_n = 2q \log(n)$, then as $n \to \infty$,

$$P\{\chi^2_n(\delta_n) > x_n^2(q)\} = \frac{1}{\sqrt{2\pi \log(n)}} \left(\frac{r}{q}\right)^{1/4} \frac{1}{\sqrt{2q - \sqrt{2r}}} \cdot n^{-(\sqrt{q} - \sqrt{r})^2}(1 + o(1)).$$

The proof of Lemma 5.1 is given in the Appendix. Once the relations (5.5) are available, the analysis proceeds exactly as in the earlier Gaussian case. The most informative values of $q$ (resp. $x^2$) are at

- if $r < \frac{1}{4}$, then $q = 4r$ or $x^2 \approx 4\delta_n$,

- if $r \geq \frac{1}{4}$, then $q = 1$ or $x^2 \approx 2\log(n)$.

correspondingly.
5.2 Generalized Gaussian (Subbotin) Distribution

The Generalized Gaussian (Subbotin) distribution [21] $GN_\gamma(\mu)$ has density function

$$\frac{1}{C_\gamma} \exp\left(-\frac{|x-\mu|^\gamma}{\gamma}\right), \quad C_\gamma = 2\Gamma\left(\frac{1}{\gamma}\right)^{\frac{1}{\gamma}-1}.$$

This class of density is introduced by M. T. Subbotin 1923 [21], and has been discussed in [11] page 195. The Gaussian is of course the special case $\gamma = 2$. The case $\gamma = 1$ corresponds to the Double Exponential (Laplace) distribution, which is a well-understood and widely-used distribution. The case $\gamma < 1$ is of interest in image analysis of natural scenes, where it has been found that wavelet coefficients at a single scale can be modelled as following a Subbotin distribution with $\gamma \approx 0.7$ [20]. This suggests that various problems of image detection, such as in watermarking and steganography, could reasonably use the model above.

A natural generalization of (1.3)-(1.4) is the following:

$$H_0: X_i \overset{i.i.d.}{\sim} GN_\gamma(0), \quad 1 \leq i \leq n, \quad (5.6)$$

$$H_1^{(n)}: X_i \overset{i.i.d.}{\sim} (1-\varepsilon)GN_\gamma(0) + \varepsilon GN_\gamma(\mu), \quad 1 \leq i \leq n. \quad (5.7)$$

Here we choose the calibrations

$$\varepsilon_n = n^{-\beta}, \quad \mu = \mu_{\gamma,n} = (\gamma \tau \log(n))^{\frac{1}{\gamma}}, \quad \frac{1}{2} < \beta < 1, \quad 0 < \tau < 1.$$

First, we will discuss the case $\gamma > 1$. In this range, the number of nonzero means is too small to be noticeable in any sense which is of expectation of order $n$; if $\tau$ is not large, we can’t expect a visible effect on the upper extreme. In short, this detection problem will be a difficult problem.

For $\gamma > 1$ we showed in [17] that the detection boundary is defined as:

$$\rho_\gamma^*(\beta) = \left\{ \begin{array}{l}
(2 \gamma^{-\frac{1}{\gamma}} - 1)^{\gamma-1}(\beta - \frac{1}{2}), \quad \frac{1}{2} < \beta \leq 1 - 2^{-\frac{1}{\gamma}}, \\
(1 - (1 - \beta)^{\frac{1}{\gamma}})^\gamma, \quad 1 - 2^{-\frac{1}{\gamma}} \leq \beta < 1.
\end{array} \right. \quad (5.8)$$

Now we discuss the case $\gamma \leq 1$. [17] shows that there is a threshold effect for the likelihood ratio test, but the detection boundary is quite different, and surprisingly it can be described in terms of $(\tau, \beta)$ independent of $\gamma$:

$$\rho_\gamma^*(\beta) = 2(\beta - \frac{1}{2}), \quad \frac{1}{2} < \beta < 1. \quad (5.9)$$

We have the following:

**Theorem 5.1** Consider applying Higher Criticism to the p-values $p_i = P\{GN_\gamma(0) > X_i\}$, $i = 1, \ldots, n$ in the setting just described. Then the detection boundary $\rho_{HC,\gamma}$ for this procedure is the same as the efficient detection boundary:

$$\rho_{HC,\gamma}(\beta) = \rho_\gamma^*(\beta), \quad 1/2 < \beta < 1.$$

The basic phenomena are depicted in the Figure 4.

The analysis can be made very similar to the Normal case. Specifically, introduce the notation $z_{\gamma,n}(q) = (\gamma q \log(n))^{\frac{1}{\gamma}}$, and let $\pi_{\gamma,n,q} = P\{GN_\gamma(0) > z_{\gamma,n}(q)\}$, $0 < q < 1$. Note that when $\gamma = 2$, $z_{\gamma,n}(q) \equiv z_\gamma(q)$ and $\pi_{\gamma,n,q} \equiv p_{n,q}$. We have

$$\max_{0 < q < \frac{1}{2}} W_n(t) = \max_{0 < q < \infty} W_n(p_{\gamma,n,q}).$$

Similarly, let $N_{\gamma,n}(q)$ count the observations exceeding $z_{\gamma,n}(q)$:

$$N_{\gamma,n}(q) = \#\{i : X_i \geq z_{\gamma,n}(q)\},$$
Figure 4: Left: Detection boundaries of the $\beta - \tau$ plane for Model (5.6) – (5.7) for $\gamma \leq 1$, and $\gamma = 1.5$, 2, 3 from top to bottom. Red parts are line segments, green parts are curves. The detection boundary for $\gamma \leq 1$ is all in red. Right: The blue region on the bottom illustrates the region of $(\beta, \tau)$ that no statistics can separate $H_1^{(n)}$ from $H_0$, and the detection boundary of $HC^*$ for all $\gamma \leq 1$ (the upper boundary of the blue region). Three lines from top to bottom correspond to the detection boundaries of Bonferroni method with $\gamma = \frac{1}{2}$, $\gamma = \frac{3}{4}$, and $\gamma = 1$.

and also

$$V_{\gamma,n}(q) = \frac{N_{\gamma,n}(q) - n\pi_{\gamma,n,q}}{\sqrt{n\pi_{\gamma,n,q}(1 - \pi_{\gamma,n,q})}}.$$ 

By arguments which are obvious at this point,

$$P(GN_\gamma(\mu_{\gamma,n}) > z_{\gamma,n}(q)) = L_n \cdot n^{-(q^{\frac{1}{2}} - r^{\frac{1}{2}})^\gamma}, \quad P(GN_\gamma(0) > z_{\gamma,n}(q)) = L_n \cdot n^{-q}, \quad r < q \leq 1.$$ 

It follows that under the alternative $H_1^{(n)}$, we have

$$EV_{\gamma,n}(q) = L_n \cdot \frac{n^{1-\beta} \cdot n^{-(q^{\frac{1}{2}} - r^{\frac{1}{2}})^\gamma}}{\sqrt{n \cdot n^{-q}}} = L_n \cdot n^{(1 + \frac{1}{2} - \beta - (q^{\frac{1}{2}} - r^{\frac{1}{2}})^\gamma)},$$

while, under the null, $EV_{\gamma,n}(q) = 0$. The most informative value of $q$ will optimize the growth rate of $n^{(1 + \frac{1}{2} - \beta - (q^{\frac{1}{2}} - r^{\frac{1}{2}})^\gamma))}$ to $\infty$.

For the case $\gamma > 1$, define

$$r_\gamma = (1 - 2^{-\frac{1}{\gamma-1}})^\gamma.$$ 

The most informative value of $q$ satisfies

- if $r < r_\gamma$, then $q = \frac{r}{r_\gamma}$ and $EV_n(q) = L_n \cdot n^{((2^{\frac{1}{\gamma-1}} - 1)^{1-\gamma} - (\beta - \frac{1}{2}))},$

- if $r \geq r_\gamma$, then $q = 1$ and $EV_n(q) = L_n \cdot n^{(1-\beta - (1-r^{\frac{1}{2}})^\gamma)}$.

For the case $0 < \gamma \leq 1$, the story is quite different, and the main reason is that

$$\frac{1 + q}{2} - \beta - (q^{\frac{1}{2}} - r^{\frac{1}{2}})^\gamma,$$

as a function of $q$, is strictly decreasing for any fixed $0 < \gamma \leq 1$, so the most informative place to look at is

$q = r$ or equivalently at $x \approx \mu$.

Notice that under $H_0$, $HC^*$ behaves the same as in the Normal case. Under $H_1^{(n)}$ the above analysis shows the behavior at the most informative place. We can argue exactly as in the proof.
of Theorem 1.2. The the growth of $EV_{\gamma,n}(q)$ easily surpasses the $\sqrt{4\log\log(n)}$ threshold, and the result follows.

There are some interesting points here.

First, the detection boundary for all the cases where $\gamma \leq 1$ looks like the limit of the boundaries for $\gamma > 1$ as $\gamma \to 1$. Second, the most informative place to look at, for the case $\gamma > 1$, is at:

$$x \approx \frac{1}{1 - 2^{-\frac{1}{\gamma - 1}}} \mu,$$

the coefficient $\frac{1}{1 - 2^{-\frac{1}{\gamma - 1}}}$ goes to 1 as $\gamma \to 1$; in comparison, for case $\gamma \leq 1$, the most informative place to look at is $x = \mu$.

Third, it is interesting to notice that for case $\gamma \leq 1$, the best that either Maximum or FDR-controlling methods can obtain is:

$$r > (1 - (1 - \beta)^{1/\gamma}).$$

(5.10)

This is strictly above the detection boundary as defined in (5.9) for any $\frac{1}{2} < \beta < 1$, while in comparison, Higher Criticism can obtain the full interior of the region of detectability, for all $\gamma$. Fourth, notice that the performance of Maximum or FDR-controlling methods worsens compared to $HC^*$ when $\gamma \to 0$; the best that the Maximum or FDR-controlling methods can do when $\gamma \approx 0$ is to detect for $r > 1$, while Higher Criticism is able to detect for $r > 2\beta - 1$, $\frac{1}{2} < \beta < 1$ independent of $\gamma$; the superiority of $HC^*$ can be seen most prominently for the case $\beta \approx \frac{1}{2}$, $\gamma \approx 0$, in which $HC^*$ is able to detect for $r > 2\beta - 1 \approx 0$, while Maximum or FDR-controlling methods are able to detect only for $r > 1$.

6 Appendix

6.1 Proof of Theorem 1.3

Lemma 6.1 If $z_n \sim \text{Binomial}(n, \pi_n)$ and $\pi_n \to 0$, $n\pi_n \to \infty$, then $P\{z_n \geq 1\} \to 1$.

Proof of Lemma 6.1:

$$P\{z_n = 0\} = (1 - \pi_n)^n = e^{-n\log(1 - \pi_n)} \to 0.$$ 

Proof of Theorem 1.3. When $r > \rho^+(\beta)$ or equivalently $1 - \beta > (1 - \sqrt{r})^2$, we can pick a constant $c > 0$ depending only on $(r, \beta)$ such that

$$1 - \beta > (\sqrt{1 + c} - \sqrt{r})^2.$$

To prove Theorem 1.3, it's sufficient to prove

$$P_{H_1^{(n)}}\{M_n \geq \sqrt{2(1 + c)\log(n)}\} \to 1, \text{ as } n \to \infty.$$ 

(6.1)

Let

$$N(c) = \#\{i : X_i \geq \sqrt{2(1 + c)\log(n)}\},$$

then under $H_1^{(n)}$, $N(c) \sim \text{Binomial}(n, q_{n,c})$ where

$$q_{n,c} = P\{(1 - \epsilon_n)N(0,1) + \epsilon_nN(\mu_n,1) \geq \sqrt{2(1 + c)\log(n)}\} = L_n \cdot n^{-\beta - (\sqrt{1 + c} - \sqrt{r})^2}.$$

Notice as $n \to \infty$, $q_{n,c} \to 0$, $nq_{n,c} \to \infty$. So let $\pi_n = q_{n,c}$ in Lemma 6.1 we have:

$$P_{H_1^{(n)}}\{N(c) \geq 1\} \to 1, \text{ as } n \to \infty.$$

Finally the desired result (6.1) follows from:

$$P_{H_1^{(n)}}\{M_n \geq \sqrt{2(1 + c)\log(n)}\} = P_{H_1^{(n)}}\{N(c) \geq 1\}.$$

\qed
6.2 Proof of Theorem 1.4

Lemma 6.2 For constants $\frac{1}{2} < \delta < 1$, $0 < a < 1$, there exists constant $c(a)$ depending only on $a$ such that

$$\sup_{\{t: t > n^{-\delta}\}} P\{\text{Binomial } (n, t) \geq n \cdot t/a\} \leq 2e^{-c(a)n^{1-\delta}}, \quad n \to \infty.$$ 

Proof of Lemma 6.2. Let $z_i^0 \overset{i.i.d.}{\sim} \text{Bernoulli } (n, t), 1 \leq i \leq n$, and let $z_i = z_i^0 - t$. Notice that $Ez_i = 0$; $E|z_i|^k \leq t, \forall k \geq 2$.

By using Bernstein’s Inequality ([19] Appendix A),

$$P\{\text{Binomial } (n, t) \geq n \cdot t/a\} = P\left\{\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} z_i \geq \left(\frac{1}{a} - 1\right)\sqrt{n} \cdot t\right\} \leq e^{-c(a)n^{-\delta}},$$

where $c(a) > 0$ is a constant which only depends on $a$. Lemma 6.2 follows:

$$\sup_{\{t > n^{-\delta}\}} P\{\text{Binomial } (n, t) \geq n \cdot t/a\} \leq 2e^{-c(a)n^{1-\delta}}.$$ 

Now for any $\frac{1}{2} < \delta < 1$, introduce statistics:

$$F_1^\delta = \min_{\{t: p_t > n^{-\delta}\}} \frac{P_t}{i/n}, \quad F_2^\delta = \min_{\{t: p_t \leq n^{-\delta}\}} \frac{P_t}{i/n}.$$ 

Lemma 6.3 For any constants $\frac{1}{2} < \delta < 1$, $0 < a < 1$, if $r < \rho_{FDR}(\beta)$ then as $n \to \infty$,

$$P_{H_0} \{F_1^\delta \leq a\} \to 0, \quad P_{H_1^{(n)}} \{F_1^\delta \leq a\} \to 0. \quad (6.2)$$

Proof of Lemma 6.3. Recall that $n \cdot F_n(t) = \sum_{\{p_t \leq t\}} 1\{p_t \leq t\}, 0 < t < 1$. We have $n \cdot F_n(t) \sim \text{Binomial } (n, t)$ under $H_0$ and $n \cdot F_n(t) \sim \text{Binomial } (n, \pi(n, t))$ under $H_1^{(n)}$, where $\pi(n, t) = P_{H_1^{(n)}}\{p_t \leq t\} \geq t$; since $r < \rho_{FDR}(\beta)$ we also have:

$$a_n \triangleq \sup_{\{t > n^{-\delta}\}} \pi(n, t)/t \to 1.$$ 

Observe that $i/n = F_n(p_t)$ so

$$\frac{P_t}{i/n} \leq a \iff F_n(p_t) \geq p_t/a;$$

by Lemma 6.2, the desired result in (6.2) follows from

$$P_{H_0} \{F_1^\delta \leq a\} \leq n \cdot \sup_{\{t: t > n^{-\delta}\}} P\{\text{Binomial } (n, t) \geq n \cdot t/a\} \to 0,$$

$$P_{H_1^{(n)}} \{F_1^\delta \leq a\} \leq n \cdot \sup_{\{t: t > n^{-\delta}\}} P\{\text{Binomial } (n, \pi(n, t)) \geq n \cdot t/a\} \to 0.$$

Proof of Theorem 1.4. For the FDR-controlling procedure:

Reject if and only if: $\min_{1 \leq i \leq n} \frac{P_i}{i/n} \leq h(n, \alpha_n)$,

where $h(n, \alpha_n) \leq a < 1$ is any given critical value. Since the attainability of FDR-controlling procedure is as good as Maximum or Bonferroni method, so all we need to prove is that for $(r, \beta)$ in the region $\rho^*(\beta) < r < (1 - \sqrt{1 - \beta})^2$, FDR-controlling method totally fails or:

$$P_{H_0}\{\text{Reject } H_0\} + P_{H_1^{(n)}}\{\text{Accept } H_0\} \to 1, \quad \text{as } n \to \infty. \quad (6.3)$$
Now under $H_1^{(n)}$ we break $\{1, 2, \ldots, n\}$ into two sets $A_1^{(n)}$ and $A_2^{(n)}$ where

\[ i \in A_1^{(n)}, \text{ if } X_i \text{ is sampled from } N(0, 1), \]
\[ i \in A_2^{(n)}, \text{ if } X_i \text{ is sampled from } N(\mu_n, 1). \]

Introduce an event:

\[ E_n^{\delta_0} = \{ p(i) \leq n^{-\delta_0} \text{ for some } i \in A_2^{(n)} \}. \]

Since $r < (1 - \sqrt{1 - \beta})^2$, or $\sqrt{r} + \sqrt{1 - \beta} < 1$, we can choose $\delta_0$ to be close enough to 1 such that $P_{H_1^{(n)}}(E_n^{\delta_0}) \rightarrow 0$. Notice $\{ F_2^{\delta_0} \mid H_0 \} = \{ F_2^{\delta_0} \mid (H_1^{(n)}, (E_n^{\delta_0})^c) \}$ and

\[ P_{H_1^{(n)}} \{ F_2^{\delta_0} < h(n, \alpha_n) \} = (1 - P_{H_1^{(n)}}(E_n^{\delta_0})) P_{H_1^{(n)}} \{ F_2^{\delta_0} < h(n, \alpha_n) \} (E_n^{\delta_0})^c + P_{H_1^{(n)}}(E_n^{\delta_0}) P_{H_1^{(n)}} \{ F_2^{\delta_0} < h(n, \alpha_n) \} |E_n^{\delta_0}|, \]

so $P_{H_0} \{ F_2^{\delta_0} < h(n, \alpha_n) \} - P_{H_1^{(n)}} \{ F_2^{\delta_0} < h(n, \alpha_n) \} \rightarrow 0$.

Finally by Lemma 6.3

\[ |P_{H_0} \{ \text{Reject} \} - P_{H_1^{(n)}} \{ \text{Reject} \} | \leq |P_{H_0} \{ F_2^{\delta_0} < h(n, \alpha_n) \} - P_{H_1^{(n)}} \{ F_2^{\delta_0} < h(n, \alpha_n) \} | \]
\[ + |P_{H_0} \{ F_1^{\delta_0} < h(n, \alpha_n) \} | + |P_{H_1^{(n)}} \{ F_1^{\delta_0} < h(n, \alpha_n) \} | \rightarrow 0, \]

the desired result in (6.3) follows. \hfill \square

### 6.3 Proof of Theorem 1.5

**Proof of Theorem 1.5.** Under $H_0$, $F_n \sim \chi_n^2$, and $EF_n = 2n$, $\text{Var}(F_n) = 4n$. Under $H_1^{(n)}$ with $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2n \log(n)}$, $\frac{1}{2} < \beta < 1$, $0 < r < 1$, direct calculations show that $2n \leq EF_n = 2n[1 + O(\epsilon \cdot L_n)]$, $\text{Var}(F_n) = 4n[1 + O(\epsilon \cdot L_n)]$; since $\beta > \frac{1}{2}$, the conclusion follows. \hfill \square

### 6.4 Proof of Theorem 1.6

To prove Theorem 1.6 we need the following lemma.

**Lemma 6.4** For $0 < x < t < 1$, then when $t \rightarrow 0$,

\[ K^+(t, x) = \begin{cases} \frac{1}{2} (t-x)^2 (1 + O(t + \frac{1}{x} - 1)), & \text{if } \frac{t}{x} \rightarrow 1, \\ t \log \frac{1}{(1+o(1))}, & \text{if } \frac{t}{x} \rightarrow \infty. \end{cases} \]

**Proof of Lemma 6.4.** The second conclusion is obvious, for the first conclusion notice $0 < x < t < 1$ and $\frac{(t-x)^2}{x^2} = o\left(\frac{(t-x)^2}{x^2}\right)$, so

\[ K^+(t, x) = t \log \frac{t}{x} + (1-t) \log \frac{1-t}{1-x} \]
\[ = t \left( \frac{t}{x} - 1 \right) - \frac{1}{2} \left( \frac{t}{x} - 1 \right)^2 \right) + (1-t) \left[ \frac{1}{2} \left( \frac{1-t}{1-x} - 1 \right)^2 \right] \right) + O\left( \frac{(t-x)^2}{x^2} \right) \]
\[ = \frac{1}{2} \left( \frac{t}{x} - 1 \right)^2 \left( 1 + O(t + \frac{t}{x} - 1) \right). \]

**Proof of Theorem 1.6.** The behavior of $BJ_n^+$ under $H_0$ is discussed in [5]:

\[ BJ_n^+ = \frac{1}{2} \left( HC^* \right)^2 \left[ 1 + O_p \left(n^{-\frac{1}{5}}\right) \right]. \]

20
Now we consider the behavior of $BJ^+_n$ under $H_1^{(n)}$. We examine the cases $r < \frac{\theta}{3}$ and $r > (1 - \sqrt{1 - \beta})^2$ separately; these two cases are actually overlapping, but the conclusion in each case is the same.

First for the case $r < \frac{\theta}{3}$, notice that $r < \frac{1}{4}$. Take $r_0$ such that $0 < r_0 < r$, similar to Lemma 6.3, it’s easy to prove that under $H_0^{(n)}$:

$$\max_{\{i: n^{-4} < r(i) < n^{-4 \rho_0}\}} \left| \frac{P(i)}{i/n} - 1 \right| \to 0 \quad \text{in probability.}$$

Introduce the following statistic:

$$HC_{r, r_0}^* = \max_{\{i: n^{-4} < r(i) < n^{-4 \rho_0}\}} \left\{ \sqrt{n} \frac{i/n - P(i)}{\sqrt{P(i)(1 - P(i))}} \right\}.$$

Now from Lemma 6.4,

$$n \cdot K^+(i/n, P(i)) \sim \frac{1}{2} \left( \max \left\{ \sqrt{n} \frac{i/n - P(i)}{\sqrt{P(i)(1 - P(i))}}, 0 \right\} \right)^2, \quad \text{if } \frac{i/n}{P(i)} \sim 1;$$

so it’s easy to prove:

$$BJ^+_n \geq \max_{\{i: n^{-4} < r(i) < n^{-4 \rho_0}\}} n \cdot K^+(i/n, P(i)) = \frac{1}{2} [HC_{r, r_0}^*]^2 (1 + o(1)), \quad (6.4)$$

thus in this case, $BJ^+_n$ is able to separate $H_0$ and $H_1^{(n)}$.

For the second case, notice $\frac{r + \beta}{2 \sqrt{r}} < 1$, pick constant $q$ such that $\max \left\{ \frac{r + \beta}{2 \sqrt{r}}, \sqrt{q} \right\} < \sqrt{q} < 1$.

Observe that under $H_1^{(n)}$:

$$\# \{i: p_i \leq n^{-q} \} \sim \text{Binomial} \left( n, L_n \cdot n^{-[\beta + (\sqrt{q} - \sqrt{q})^2]} \right), \quad L_n \cdot n^{-[\beta + (\sqrt{q} - \sqrt{q})^2]} \gg n^{-q}.$$

This implies that, under $H_1^{(n)}$, for those $p$-values $p(i) \sim n^{-q}$, $\frac{i/n}{P(i)} \gg 1$; so from Lemma 6.4,

$$n \cdot K^+(P(i), n^{-q}) \sim L_n \cdot \text{Binomial} \left( n, L_n \cdot n^{-[\beta + (\sqrt{q} - \sqrt{q})^2]} \right).$$

Notice $1 - \beta - (\sqrt{q} - \sqrt{q})^2 > 0$, it’s easy to see that $BJ^+_n$ is able to separate $H_1^{(n)}$ and $H_0$.

**Remark.** For the second case, $HC^*$ is more powerful than $BJ^+_n$. In fact, the best $BJ^+_n$ can do is to choose $i$ as large as possible while keeping $\frac{i/n}{P(i)} \gg 1$. This is roughly equivalent to $i \sim n^{1-q}$ with $q$ satisfying

$$n^{1-\beta} n^{-(\sqrt{q} - \sqrt{q})^2} \gg n^{1-q} \iff \sqrt{q} > (r + \beta)/(2 \sqrt{r}).$$

As a result, $BJ^+_n \approx L_n \cdot n^{1-(r+\beta)^2/(4r)}$. To see the main idea, take the region $\frac{1}{2} < \beta < \frac{3}{4}$, $\rho^*(\beta) < r < 1 - \beta$ for comparison. For $(r, \beta)$ in this range, $(HC^*)^2 \approx L_n \cdot n^{2r - \beta + 1}$. Since $1 - (r + \beta)^2/(4r) < 2r - \beta + 1$, $HC^*$ has better performance than $BJ^+_n$.

### 6.5 Proof of Lemma 5.1

With $z_t \overset{i.i.d.}{\sim} N(0, 1)$, $\delta_n = \mu^2_n = 2r \log(n)$, $\chi^2(\delta_n) = D \cdot z_t^2 + z_2^2 + \ldots + (z_r + \mu_n)^2$, so

$$P\{\chi^2(\delta_n) \geq 2q \log(n)\} = P\{z_t^2 + z_2^2 + \ldots + (z_r + \mu_n)^2 \geq 2q \log(n)\}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{A(\theta_1, \phi)} \cos^{\nu-2} \theta_1 \rho^{\nu-1} e^{-\frac{\xi^2}{2}} d\theta_1 d\rho \int_{-\frac{\xi}{2}}^{\frac{\xi}{2}} \cos^{\nu-3} \theta_2 d\theta_2 \ldots \int_0^{2\pi} d\theta_{v-1}$$

$$= \frac{2^{-\nu+1}}{\sqrt{\pi} \Gamma\left(\frac{n-1}{2}\right)} \int_{|\theta_1| \leq \frac{\xi}{2}} \cos^{\nu-2} \theta_1 d\theta_1 \int_{\sqrt{-\xi^2 \sin \theta_1 + \sqrt{2\xi \cos \theta_1}} \log(n)}^{\infty} \rho^{\nu-1} e^{-\frac{\xi^2}{2}} d\rho,$$
where \( A(\theta_1, \rho) = \{ (\theta_1, \rho) : |\theta_1| \leq \frac{\pi}{\nu}, \rho^2 + 2\rho \mu_n \sin \theta_1 \geq 2(q - r) \log(n) \} \).

The case \( \nu = 1 \) is obvious, while for \( \nu = 2 \) we have:

\[
P(\chi_2^2(\delta_n) \geq 2q \log(n)) = \frac{1}{\pi} \int_{|\theta_1| \leq \frac{\pi}{2}} n^{-|\sqrt{q-r} \cos \theta_1 - \sqrt{r} \sin \theta_1|^2} d\theta_1
\]

\[
= \frac{1}{\pi \sqrt{\log(n)}} n^{-|\sqrt{q-r}|^2} \int_0^\infty e^{-\sqrt{r}(\sqrt{q-r})^2x^2} dx(1 + o(1))
\]

\[
= \frac{1}{\sqrt{2\pi \log(n)}} \frac{1}{\sqrt{2q - \sqrt{2r}}} (\frac{r}{q})^{\frac{1}{4}} \cdot n^{-|\sqrt{q-r}|^2}(1 + o(1)).
\]

Now for the case \( \nu \geq 3 \). Notice for fixed \( r \) and \( q \) but large \( n \), \( n^{-|\sqrt{r} \sin \theta_1 + \sqrt{q-r} \cos \theta_1|^2} \) obtains its maximum rate of growth at \( \theta_1 = \frac{\pi}{2} \), and for \( \theta_1 \approx \frac{\pi}{2} \),

\[
[-\sqrt{r} \sin \theta_1 + \sqrt{q-r} \cos \theta_1]^2 \sim [\sqrt{q-r}]^2 + \frac{\sqrt{r}}{\sqrt{q}}(\sqrt{q-r})^2(\theta_1 - \frac{\pi}{2})^2;
\]

moreover notice for any \( y \to \infty \), \( \int_y^\infty \rho^{-\nu - 1} e^{-\frac{\rho^2}{2}} d\rho = y^{\nu - 2} e^{-\frac{y^2}{2}} (1 + o(1)) \), these enable us to write:

\[
\int_{|\theta_1| \leq \frac{\pi}{2}} \cos^{-2} \theta_1 d\theta_1 \int_0^\infty n^{-|\sqrt{q-r} \sin \theta_1 + \sqrt{q-r} \cos \theta_1|^2} \log(n) \rho^{-\nu - 1} e^{-\frac{\rho^2}{2}} d\rho
\]

\[
= (\sqrt{2q - \sqrt{2r}} \log(n))^{\nu - 2} \int_{|\theta_1| \leq \frac{\pi}{2}} \cos^{-2} \theta_1 \cdot n^{-|\sqrt{q-r} \sin \theta_1 + \sqrt{q-r} \cos \theta_1|^2} d\theta_1(1 + o(1))
\]

\[
= (\sqrt{2q - \sqrt{2r}} \log(n))^{\nu - 2} \int_0^1 (1 - x^2)^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}x(1+x)|^2} dx(1 + o(1))
\]

\[
= (\sqrt{2q - \sqrt{2r}} \log(n))^{\nu - 2} \int_0^1 (1 + x)^{\frac{\nu+3}{2}} (1 - x)^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}x(1+x)|^2} dx(1 + o(1))
\]

\[
= 2^{\frac{\nu+3}{2}} (\sqrt{2q - \sqrt{2r}} \log(n))^{\nu - 2} \int_0^1 (1 - x)^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}x(1+x)|^2} dx(1 + o(1)).
\]

To evaluate this integration, notice that

\[
\int_0^1 (1 - x)^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}x(1+x)|^2} dx
\]

\[
\int_0^1 x^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}x(1+x)|^2} dx
\]

\[
= n^{-|\sqrt{q-r}|^2} \int_0^1 x^{\frac{\nu+3}{2}} n^{-2\sqrt{r}(\sqrt{q-r})^2x^2} dx(1 + o(1))
\]

\[
= \frac{1}{\log(n)} n^{-|\sqrt{q-r}|^2} \int_{\log(n)}^{\infty} \left( \frac{x}{\log(n)} \right)^{\frac{\nu+3}{2}} e^{-\sqrt{r}(\sqrt{q-r})^2x^2} dx(1 + o(1))
\]

\[
= (\log(n))^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}|^2} \int_0^\infty x^{\frac{\nu+3}{2}} e^{-\sqrt{r}x(\sqrt{q-r})^2} dx(1 + o(1))
\]

\[
= \Gamma(\frac{\nu+1}{2}) (\frac{r}{q}) (\sqrt{2q - \sqrt{2r}} \log(n))^{\frac{\nu+3}{2}} n^{-|\sqrt{q-r}|^2}(1 + o(1)),
\]

so finally we have:

\[
P(\chi_2^2(\delta_n) \geq 2q \log(n)) = \frac{1}{\sqrt{2\pi \log(n)}} \frac{(\frac{r}{q})^{\frac{\nu+3}{4}}}{\sqrt{2q - \sqrt{2r}}} \frac{1}{\sqrt{2q - \sqrt{2r}}} n^{-|\sqrt{q-r}|^2}(1 + o(1)).
\]
References


