The Annals of Statistics
2020, Vol. 48, No. 4, 2230–2252
https://doi.org/10.1214/19-AOS1885

BEYOND HC: MORE SENSITIVE TESTS FOR RARE/WEAK ALTERNATIVES

BY THOMAS PORTER[1] AND MICHAEL STEWART[2]

In memory of Peter Hall

[1]*School of Mathematics and Statistics, University of Melbourne,* thomas.porter@outlook.com.au
[2]*School of Mathematics and Statistics F07, University of Sydney,* michael.stewart@sydney.edu.au

Higher criticism (HC) is a popular method for large-scale inference problems based on identifying unusually high proportions of small p-values. It has been shown to enjoy a lower-order optimality property in a simple normal location mixture model which is shared by the 'tailor-made' parametric generalised likelihood ratio test (GLRT) for the same model; however, HC has also been shown to perform well outside this 'narrow' model.

We develop a higher-order framework for analysing the power of these and similar procedures, which reveals the perhaps unsurprising fact that the GLRT enjoys an edge in power over HC for the normal location mixture model. We also identify a similar parametric mixture model to which HC is similarly 'tailor-made' and show that the situation is (at least partly) reversed there. We also show that in the normal location mixture model a procedure based on the empirical moment-generating function enjoys the same local power properties as the GLRT and may be recommended as an easy to implement (and interpret), complementary procedure to HC. Some other practical advice regarding the implementation of these procedures is provided. Finally, we provide some simulation results to help interpret our theoretical findings.

1. Introduction. With the 'data flood' of recent times, methods to handle high-dimensional data have become increasingly important. Higher criticism, introduced in Donoho and Jin (2004) has become a widely used method for multiple testing and variable selection. It was motivated originally as an alternative to parametric methods for a simple sparse normal location mixture detection problem and involves null-standardising the empirical cumulative distribution function (CDF) and rejecting for large values of the supremum of this one-dimensional empirical process. It has since been extended and generalised to many settings including sparse covariance matrix estimation; see the survey papers Donoho and Jin (2015), Jin and Ke (2016). In particular, it has been shown to perform well in contexts far beyond the simple mixture model it was originally motivated by; see, for example, Cai, Jeng and Jin (2011) and Cai and Wu (2014). One of its stated advantages at its inception was that it was not 'tied to the narrowly specified model' in the way the corresponding parametric methods were suggested to be.

The purpose of this article is to make the case that we should also consider the parametric methods referred to above, including the generalised likelihood ratio test (GLRT) and a related method involving the empirical moment-generating function (EMGF), as complementary methods to HC and that indeed they are closer to HC in nature than might appear at first glance. In particular, we show that there is another similar 'narrow' model to which HC is 'tied' in precisely the same way as the GLRT and EMGF are 'tied' to the normal location

Received November 2016; revised July 2019.
MSC2020 subject classifications. Primary 62F03; secondary 62G30, 62G32, 62F05.
Key words and phrases. Higher criticism, sparse normal means, phi-divergence, multiple comparisons, mixture model.

mixture model. Our intention is not to discredit HC; it is rather to point out that this notion of being ‘tied to a narrowly specified model’ is a misplaced criticism for statistics of this type.

The framework under which the theoretical properties of HC were originally developed was not detailed enough to discern any difference in performance between HC and the GLRT under the normal location mixture model. The main technical contribution of this article is to provide a framework for higher-order power comparisons between these and related statistics which reveals that each statistic has an edge in power under the model to which it is ‘tied’, which is perhaps unsurprising. What is useful is to see the degree to which this occurs and also to see the scales balanced in a certain sense: each statistic has benefits and drawbacks in different situations, none is ‘uniformly better’ across all scenarios.

Our technical results only hold in the restricted setting of simple mixture models. However, there is nothing to suggest that the EMGF and GLRT procedures cannot at least partly match HC’s broader usefulness. While the GLRT is more complicated to implement than HC, the EMGF is much simpler than and potentially as useful as HC itself. We do not explore the case of correlated test statistics here, although a modification analogous to the innovated HC of Hall and Jin (2010) represents a very interesting avenue of further research. A similar comment holds for extensions to variable selection settings as in Donoho and Jin (2008); see also the discussion in Donoho (2017). Adapting either the GLRT or EMGF as tools for identifying as well as detecting significant effects ought to be possible, indeed estimating underlying optimal thresholds may be facilitated by plugging parameter estimates provided by, for example, the GLRT into certain theoretical expressions for the thresholds.

The remainder of the article is organised as follows: Section 2 introduces our contamination model, explains how it relates to multiple testing and defines the various procedures being compared. Section 3 presents the two main examples we wish to compare and contrast as well as our main theoretical results. Section 4 gives detailed technical arguments that are used to prove the main results in Section 3. Section 5 presents a summary of some simulation experiments used to illustrate the theoretical results. Section 6 concludes the paper with a brief discussion. Further technical details and a more complete set of simulation results are provided in the Supplementary Material (Porter and Stewart (2020)).

2. Contamination models and procedures for multiple testing.

2.1. *Generalisation of Donoho and Jin’s contamination model.* We present a generalisation of the contamination model used in Donoho and Jin (2004) for modelling test statistics in a simple multiple testing framework. We have n independent and identically distributed (IID) random variables $X_1, \dots, X_n$ and interpret X_i as the test statistic for the ith ‘sub-hypothesis’. The X_i’s have common cumulative distribution function (CDF)

$$P(X_1 \le x) = (1-p)F_0(x) + pF_\theta(x). \tag{1}$$

Here, F_0 represents the common null distribution of the test statistics and is embedded in a 1-parameter family $\{F_\theta\}$ of CDFs, each of which possess a density with respect to Lebesgue measure and is absolutely continuous with respect to F_0. The mixing proportion p represents the proportion of false null sub-hypotheses and is typically ‘small’, reflecting ‘sparseness’. The connection to sparse regression models is that each X_i could be a test statistic for assessing the significance of the ith regression coefficient in a regression model with a large number of independent, random predictors as is assumed, for example, in so-called naïve Bayes classification (see, e.g., Bickel and Levina (2004)). We are interested in power under sparse local alternatives, so that $(p, \theta) = (p_n, \theta_n)$ depend on the sample size n, in particular, we define $p_n = p_n(\beta) = n^{-\beta}$ for fixed $\beta \in (\frac{1}{2}, 1)$. We then focus on which sequences $\{\theta_n\}$ are or are not detectable for each β.

We suppose for simplicity that large X_i provides evidence against the ith null *sub*-hypothesis. In this case it also makes sense to only consider families $\{F_\theta\}$ within which each F_θ is stochastically larger than F_0, so that

$$F_\theta(x) \le F_0(x) \quad \text{for each } x \text{ and each } \theta. \tag{2}$$

It also makes sense to then define p-values via $V_i = 1 - F_0(X_i)$. The model (1) for the X_i's induces a similar model on the V_i's where F_0 is replaced by the $U(0,1)$ CDF: for $0 < v < 1$,

$$P(V_1 \le v) = (1-p)v + pG_\theta(v), \tag{3}$$

where

$$G_\theta(v) = 1 - F_\theta\big[F_0^{-1}(1-v)\big] \tag{4}$$

represents the assumed common distribution of the p-values corresponding to false null sub-hypotheses. While (3) resembles (1), there is a different stochastic ordering: due to (2),

$$G_\theta(v) \ge v \quad \text{for each } 0 < v < 1 \text{ and each } \theta. \tag{5}$$

2.2. *The global hypothesis test.* The first stage in a multiple testing procedure is to test whether any null sub-hypotheses are false. If not then, the X_i's are IID F_0. The 'global' hypothesis test we are interested in is

$$H_0\colon\ P(X_1 \le x) = F_0(x) \tag{6}$$

for all x. We shall first identify two test statistics in cases where the parameter θ is known, and then consider generalisations of these to the case where θ is unknown and follow these with some examples.

2.3. *Test statistics for known θ.* It is useful to first consider standard test statistics for the restrictive case where the distribution of each test statistic under its alternative sub-hypothesis is F_θ for some known $\theta \ne 0$ which we can interpret as a 'common effect size' across those tests where the null sub-hypothesis is false. The global hypothesis (6) then reduces directly to $H_0\colon\ p = 0$.

2.3.1. (*Generalised*) *log-likelihood ratio.* The (generalised) log-likelihood ratio test (GLRT) statistic is given by

$$L_n = L_n(\theta) = \sup_{0 \le p \le 1} \sum_{i=1}^{n} \log\left\{1 + p\left[\frac{dF_\theta}{dF_0}(X_i) - 1\right]\right\}. \tag{7}$$

We use the qualifier 'generalised' to indicate that we have maximised over the parameter p. Some authors reserve the term '(log-)likelihood ratio test' for the Neyman–Pearson (NP) test of simple null hypothesis versus simple alternative, a distinction which proves convenient below.

2.3.2. *Rao score statistic.* The Rao score statistic is the standardised gradient of the log-likelihood ratio at $p = 0$:

$$U_n = U_n(\theta) = (n v_\theta)^{-1/2} \sum_{i=1}^{n} \left[\frac{dF_\theta}{dF_0}(X_i) - 1\right], \tag{8}$$

where

$$v_\theta = \int \left(\frac{dF_\theta}{dF_0}\right)^2 dF_0 - 1, \tag{9}$$

assumed to be finite (for each θ).

The hypothesised value $p = 0$ lies on the boundary of the parameter space, a violation of the classical regularity condition for asymptotic likelihood theory. The first asymptotic analysis for problems of this type is Chernoff (1954). An elementary Taylor series argument can be used to show that under suitable regularity conditions,

$$2L_n = \max(0, U_n)^2 + o_p(1). \tag{10}$$

The $\max(0, \cdot)$ appears because if the gradient of the log-likelihood ratio at $p = 0$ is negative, due to concavity the maximum over p in (7) occurs on the boundary giving $L_n = 0$. Importantly, *only large positive values of U_n provide evidence against* H_0, not large absolute values as in the regular (nonboundary) case.

2.4. *Test statistics for unknown θ.* Now we relax the restriction that the distribution under each alternative sub-hypothesis is known, letting θ now denote an *unknown* (but still common) effect size across those tests where the null sub-hypothesis is false. However, since F_0 is included in the family $\{F_\theta\}$ we have an identifiability issue: the global null hypothesis

$$H_0 \colon P(X_1 \le x) = F_0(x)$$

now corresponds parametrically to either $p = 0$ *or* $\theta = 0$ (or both). There is no unique pair of parameter values (p, θ) corresponding to H_0.

2.4.1. *GLRT.* For θ unknown, the quantity $L_n(\theta)$ becomes a *profile* log-likelihood ratio and a natural statistic is the *full GLRT statistic*

$$\sup_\theta L_n(\theta). \tag{11}$$

2.4.2. *Maximal score.* The generalisation of the Rao score statistic is not straightforward due to the nonidentifiability problem: at which 'true' parameter value do we evaluate the gradient of the log-likelihood?

Instead, we can appeal to the approximation (10) and, as with (11) above, simply maximise over θ, giving the *maximal score statistic*:

$$\sup_\theta U_n(\theta). \tag{12}$$

It is important to realise that the $o_p(1)$ term in (10) is *not necessarily uniform in θ*, so an analogous approximation linking (11) and (12) is not available immediately. We refer to $\{U_n(\theta)\}$ as the *score process*.

2.5. *Other statistics.* We introduce two other statistics at this point; others will be introduced later in Section 4.1.

2.5.1. *Higher criticism.* As stated in the Introduction, the higher criticism statistic is

$$HC_n = HC_n(I_n) = \sup_{x \in J_n} n^{-1/2} \sum_{i=1}^{n} \frac{1\{X_i \le x\} - F_0(x)}{\sqrt{F_0(x)[1 - F_0(x)]}}, \tag{13}$$

where $J_n = \{x \colon 1 - F_0(x) \in I_n\}$. If $I_n = (a_n, b_n)$, $[a_n, b_n)$, etc. we write $HC_n(a_n, b_n)$, $HC_n[a_n, b_n)$, etc. The interval I_n can be chosen in various ways, and leads to various versions each of which has its own strengths and weaknesses. Henceforth we denote the unrestricted version, where $I_n = (0, 1)$, as HC_n unless stated otherwise; other variants are introduced and discussed below in Section 4.1.1.

There are alternate ways this statistic can be expressed. In particular, writing $\mathbb{F}_n(v)$ for the empirical CDF of the p-values $V_i = 1 - F_0(X_i)$, we have

$$HC_n(I_n) = \sup_{v \in I_n} \mathbb{Z}_n(v), \tag{14}$$

where

$$\mathbb{Z}_n(v) = \frac{n^{1/2}[\mathbb{F}_n(v) - v]}{\sqrt{v(1-v)}}. \tag{15}$$

Since the supremum is attained at one of the jump points of $\mathbb{F}_n(\cdot)$, that is, (15) with $v = V_{(j)}$ for some j, we may further restrict the maximisation:

$$HC_n(I_n) = \max_{\{j\,:\,V_{(j)} \in I_n\}} \frac{n^{-1/2}[\frac{j}{n} - V_{(j)}]}{\sqrt{V_{(j)}(1 - V_{(j)})}}. \tag{16}$$

2.5.2. *Berk–Jones statistic.* Another related statistic we consider is one of those originally proposed in Berk and Jones (1979), motivated by large deviation theory:

$$R_n^+ = \sup_{\substack{t \in (0,1) \\ \mathbb{F}_n(t) \geq t}} \left\{ [1 - \mathbb{F}_n(t)] \log\left(\frac{1 - \mathbb{F}_n(t)}{1-t}\right) + \mathbb{F}_n(t) \log\left(\frac{\mathbb{F}_n(t)}{t}\right) \right\}. \tag{17}$$

We discuss R_n^+ and other related statistics in Section 4.1.2.

3. Examples and main results.

In this section, we present our main technical results on power under sparse local alternatives under the two important examples of the contamination models from the previous section. Both have $F_0 = \Phi$ as the standard normal CDF but each has it embedded within a different 1-parameter family $\{F_\theta\}$. The first is the model originally studied in Hartigan (1985) and also used to motivate HC in Donoho and Jin (2004).

3.1. *Normal location mixture model.* Let $F_\theta(x) = \Phi(x - \theta)$ be the $N(\theta, 1)$ CDF. Since large X_i's are considered significant we restrict the class $\{F_\theta : \theta \geq 0\}$. Then

$$\frac{dF_\theta}{dF_0}(x) = e^{\theta x - \theta^2/2} \quad \text{and} \quad \text{the variance at (9) becomes } v_\theta = e^{\theta^2} - 1.$$

Denote the full GLRT statistic for this model as

$$\Lambda_n = \sup_{\substack{0 \leq p \leq 1 \\ \theta \geq 0}} \sum_{i=1}^n \log\{1 + p[e^{\theta X_i - \theta^2/2} - 1]\}.$$

The score process is given by

$$U_n(\theta) = n^{-1/2} \sum_{i=1}^n \frac{e^{\theta X_i - \theta^2/2} - 1}{\sqrt{e^{\theta^2} - 1}};$$

write the maximal score statistic for this model as

$$T_n = \sup_{\theta \geq 0} U_n(\theta).$$

Note that the score process $\{U_n(\theta)\}$ in this example is the null-standardised EMGF of the X_i's. The two statistics Λ_n and T_n are, in the sense of Donoho and Jin (2004), 'tied' to this normal location mixture model and might be expected to perform better than other statistics like HC_n and R_n^+.

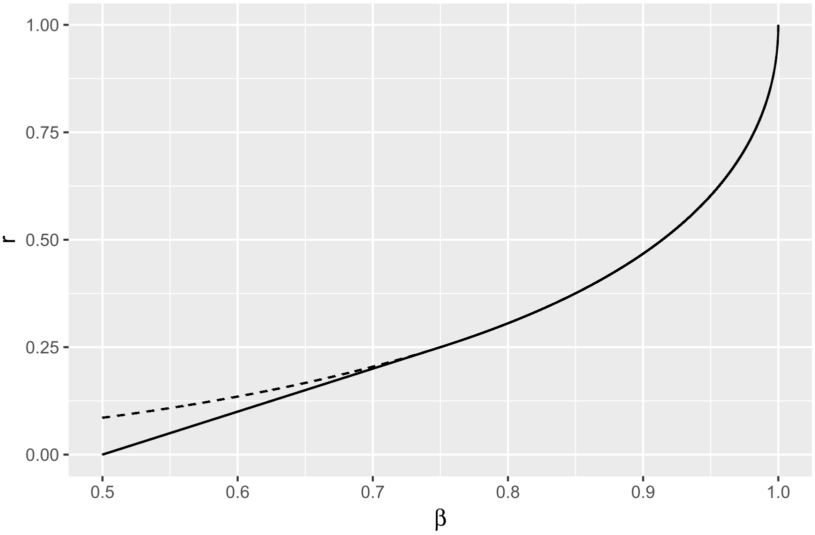

FIG. 1. *The detection boundary phase diagram with r on the y-axis, β on the x-axis. The detection boundary for the NP test $\rho^*_{\mathrm{NP}}(\beta)$ (see (18)) is the solid black line, and the detection boundary for the Bonferroni test, $(1-\sqrt{1-\beta})^2$ is the dashed black line. The Berk and Jones (1979) and HC statistics share the NP detection boundary, whereas the Bonferroni test does not.*

Adapting the work of Ingster (1997, 2001, 2002), they introduced the *detection boundary* (see Figure 1):

$$\rho^*_{\mathrm{NP}}(\beta) = \begin{cases} \beta - 1/2 & \text{if } \beta \in (1/2, 3/4), \\ (1-\sqrt{1-\beta})^2 & \text{if } \beta \in [3/4, 1). \end{cases} \tag{18}$$

In Donoho and Jin (2004) for $\beta \in (\frac{1}{2}, 1)$, the parametrisation $\theta_n = \sqrt{2r \log n}$ was used. The first proposition below contains results from various papers already cited.

PROPOSITION 3.1. *Under the local alternative* $(p_n = n^{-\beta}, \theta_n = \sqrt{2r \log n})$:

1. *if* $r < \rho^*_{\mathrm{NP}}(\beta)$, *all level-*$\alpha$ *tests have limiting power* α (*Ingster* (1997));
2. *if* $r > \rho^*_{\mathrm{NP}}(\beta)$, *level-*$\alpha$ *tests based on* HC_n (*or any of its variants described in Section* 4.1.1) *or* R^+_n *have limiting power* 1 (*Donoho and Jin* (2004)).

Thus in a sense HC_n is optimal in that it 'attains' this detection boundary. Furthermore, Donoho and Jin ((2004), page 965) states '*it is not clear that [the test based on* Λ_n *above] can be relied on to detect subtle departures from* H_0'. These, together with the fact that HC_n is not 'tied' to the model gives an impression that HC_n is 'uniformly better' than Λ_n: it has this particular optimality property but is perhaps less sensitive to model assumptions. Analogous remarks apply when replacing HC_n with R^+_n. Our second proposition provides a more detailed and balanced picture of the situation.

PROPOSITION 3.2. *Under the local alternative*,

$$\left(p_n = n^{-\beta}, \theta_n = \sqrt{2\rho^*_{\mathrm{NP}}(\beta) \log n + \varepsilon_n}\right),$$

there exist a constant C_β *and sequences* $\varepsilon_{n0} \le \varepsilon_{n1}, \varepsilon_{n2} \le \varepsilon_{n3} \le \varepsilon_{n4}$ *satisfying*:

- $\varepsilon_{nj} \sim \log\log\log n$ *for* $j = 0, 1$;
- $\varepsilon_{nj} \sim C_\beta \log\log n$ *for* $j = 2, 3, 4$

such that:

1. *for* $\beta \in (1/2, 3/4)$ *and* $\varepsilon_n = \varepsilon_{n0}$, *the limiting power of all level-*α *tests is* α;
2. *for* $\beta \in (1/2, 3/4)$ *and* $\varepsilon_n = \varepsilon_{n1}$, *the limiting power of the level-*α *test based on* Λ_n *tends to* 1;
3. *for* $\beta \in (3/4, 1)$ *and* $\varepsilon_n = \varepsilon_{n2}$ *the limiting power of all level-*α *tests is* α;
4. *for* $\beta \in (3/4, 1)$ *and* $\varepsilon_n = \varepsilon_{n3}$, *the limiting power of the level-*α *test based on* Λ_n *tends to* 1;
5. *for* $\beta \in (1/2, 1)$ *and*:

(a) $\varepsilon_n = \varepsilon_{n3}$ *the limiting power of the level-*α *test based on* HC_n *tends to* α;
(b) $\varepsilon_n = \varepsilon_{n4}$ *the limiting power of the level-*α *test based on* HC_n *tends to* 1.

Strictly speaking, we should say ‘all *adaptive* level-α tests’ in statements 1 and 3; power could be improved by incorporating knowledge of the alternative parameter values. However, the parameters p and θ are unknown so all tests under consideration are ‘adaptive’ in the sense of Ingster (2001, 2002); see Section 4.3 below for further explanation.

Including the higher-order term ε_n has allowed us to see a difference in performance between Λ_n and HC_n here in the model to which it is ‘tied’. Statements 1 and 3 follow from Ingster (2001), the remaining statements follow from results in Section 4.3. Note that Λ_n achieves the best possible rates in both cases. There is a gap in rates for $\beta \in (\frac{1}{2}, \frac{3}{4})$ between Λ_n and HC_n but not for $\beta \in (\frac{3}{4}, 1)$; however, note also that statements 4 and 5(a) are under the same sequence ε_{n3} so we can still distinguish between the two tests in this case. We see in Section 4.3 that T_n also achieves the same rates as Λ_n.

Thus Λ_n has a *slight* edge in performance in this, the model for which we would expect it to do better. We see an interesting reversal in the next example.

3.1.1. *Implementation and minor modifications.* We can modify T_n, while retaining all of the asymptotic and empirical results in this paper, by maximising over a sieve $\Theta_n \subseteq \Theta$ that is growing at a sufficient rate. One such example is $\Theta_n \equiv [0, X_{(n)}/2]$, or alternatively we can replace $X_{(n)}/2$ with its expectation under H_0 given by $0.5\Phi^{-1}(1 - \gamma/n)$, where γ is the Euler–Mascheroni constant and Φ^{-1} is the standard normal quantile function. We justify this by considering the following approximation to $U_n(\theta)$ provided by

$$\begin{aligned}(1 - e^{-\theta^2})^{1/2} U_n(\theta) &= \frac{e^{-\frac{\theta^2}{2}}}{\sqrt{n}} \sum_{i=1}^{n} (e^{\theta X_i - \theta^2/2} - 1) \\ &= U_n(\theta)\{1 + o[(\log\log n)^{-1}]\}\end{aligned}$$

uniformly in $\theta \geq \log\log\log n$. Bickel and Chernoff (1993) determined that the limiting distribution for the supremum of the approximating process is the same as that for T_n. Moreover, Bickel and Chernoff (1993) showed that the maximiser of the approximating process occurs in the range $\theta \in [0, X_{(n)}/2]$ with probability tending to one. There is no asymptotic loss of power (to the degree required for our results to be unchanged) by maximising $U_n(\theta)$ over $\Theta_n \equiv [0, X_{(n)}/2]$.

A further simplification is to restrict the maximisation over only values of θ equal to X_i-values. That is, writing $U_n(\theta) = U_n(\theta; X_1, \ldots, X_n)$, consider the statistic

$$\widetilde{T}_n = \max_{i=1,\ldots,n} U_n(X_i; X_1, \ldots, X_n).$$

While avoiding the need to use numerical optimisation, this can be slower to evaluate that T_n, but also still possesses all the relevant theoretical properties enjoyed by T_n. First, the behaviour under the null hypothesis is controlled, as the Bickel and Chernoff (1993) sequence of critical values is an upper bound on the sequence of critical values for $\tilde{T}_n$. Second, it suffices for the higher-order power analysis to show that there exists a sequence d_n diverging to infinity slowly enough—of order $\log\log\log(n)$ or $\log\log(n)$ in this paper—so that the arguments that maximise T_n and $\tilde{T}_n$, respectively, differ by $o_p(d_n)$.

3.2. *Truncated normal mixture model.*

Let

$$F_\theta(x) = (1-\theta)^{-1}\Phi(x)1\{x \geq \Phi^{-1}(\theta)\}$$

denote the standard normal CDF truncated below the θ-quantile (see Figure 2 for an example of the general mixture model (1) for this choice of F_θ).

Then

$$\frac{dF_\theta}{dF_0}(x) = (1-\theta)^{-1}1\{x \geq \Phi^{-1}(\theta)\} = (1-\theta)^{-1}1\{\Phi(x) \geq \theta\},$$

the variance at (9) becomes

$$v_\theta = (1-\theta)^{-1}\theta,$$

the full GLRT statistic takes the form

$$\sup_{\theta, 0\leq p\leq 1} \sum_{i=1}^{n} \log\{1 + p[(1-\theta)^{-1}1\{\Phi(X_i) \geq \theta\} - 1]\}$$

and the score process is given by

$$U_n(\theta) = n^{-1/2}\sum_{i=1}^{n} \frac{1\{\Phi(X_i) \geq \theta\} - (1-\theta)}{\sqrt{\theta(1-\theta)}}.$$

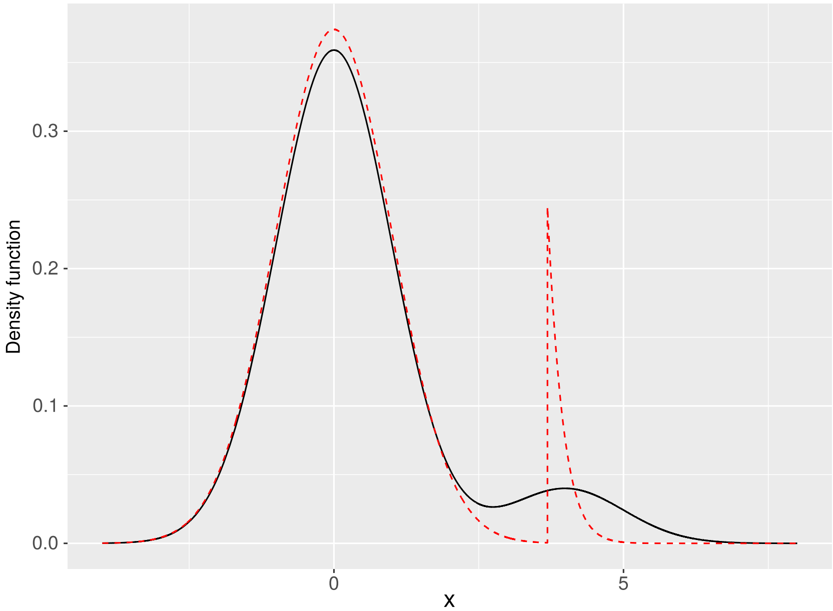

FIG. 2. *The solid curve is the normal location mixture density with parameters $p = 0.1$ and $\theta = 4$. The dashed curve is the truncated normal mixture density with parameters $p = 0.0623$ and $\theta = 3.6869$ and is the closest such density to the curve (in Kullback–Liebler divergence).*

Note that $\{U_n(\theta)\}$ is the null-standardised empirical CDF of the p-values evaluated at $1-\theta$. In particular, for this model the maximal score statistic is the HC statistic

$$\sup_\theta U_n(\theta) = HC_n.$$

Furthermore, we show below in Theorem 4.14 that the full GLRT statistic here is nR_n^+, n times the Berk–Jones statistic.

Thus the statistics R_n^+ and HC_n are 'tied' to the truncated normal mixture model *in exactly the same way* that Λ_n and T_n are tied to the normal location mixture model. Our observation is not a criticism of any of these statistics; instead, it is more a statement that one should be wary of finding any fo these statistics less desirable by association to a parametric model.

Note, too, that expressed in terms of the p-values, the general model (3) reduces to

$$P(V_1 \le v) = (1-p)v + pv(1-\theta)^{-1}1\{0 < v \le 1-\theta\}, \tag{19}$$

that is a *uniform scale mixture*; the distribution G_θ in (4) of p-values corresponding to false hypotheses is $U(0, 1-\theta)$.

We have an analogous proposition outlining the detection boundary for this model. We express it in terms of all 4 statistics: HC_n, R_n^+ above, as well as the statistics Λ_n and T_n motivated by the previous example.

PROPOSITION 3.3. *Under the local alternative* $(p_n = n^{-\beta}, \theta_n = 1 - n^{-r})$, *for some* $(r, \beta) \in (0,1) \times (\frac{1}{2}, 1)$:

1. *any level-*α *test has limiting power* α *if* $r < 2\beta - 1$;
2. *level-*α *tests based on any of the statistics* HC_n, R_n^+, Λ_n *and* T_n *all have limiting power* 1 *if* $r > 2\beta - 1$.

Again, this detection boundary result does not provide enough detail to distinguish between these tests. We might suspect that since HC_n and R_n^+ are 'tied' to this model, they might do better than Λ_n and T_n. This is suggested by the next proposition.

PROPOSITION 3.4. *Let* $p = p_n = o(n^{-1/2})$ *and* $np \to \infty$. *There is a sequence* $c_n \sim \sqrt{2\log\log n}$ *and a constant* $K > 0$ *such that if*:

1. $\sqrt{n}p\sqrt{(1-\theta)^{-1}-1} - c_n \to \infty$, *the limiting power of the level-*α *test based on* HC_n *is* 1;
2. $\sqrt{n}p\sqrt{(1-\theta)^{-1}-1} - c_n = o(1/c_n)$, *the limiting power of the level-*α *test based on* HC_n *is* α;
3. *if* $\sqrt{n}p\sqrt{(1-\theta)^{-1}-1} \ge K(\log n)^{1/4}\sqrt{\log\log n}$, *the limiting powers of the level-*α *tests based on* Λ_n *and* T_n *are both* 1.

Statements 1 and 2 follow from Proposition 4.18 (with c_n given by c_n^{HC} there) and statement 3 follows from Remark 4.19.

4. Detailed technical arguments. In this section, we provide a series of theoretical results which include derivations of the main results in Section 3. We present additional statistics, derive all limiting null distributions and then provide results on power under the two mixture alternatives introduced in Section 3.

The proofs of these results appear in the Supplementary Material. Some key features of the methods used are very delicate approximation of the Gaussian tail using the Birnbaum–Sampford inequality, careful extreme-value expansions of upper order statistics and delicate asymptotic analysis of the mean and variance functions of certain empirical processes.

4.1. *Other statistics.*

4.1.1. *Variants of higher criticism.* According to Donoho and Jin (2004, 2015), Tukey (1976, 1989, 1994) suggested that a 'second-level significance' test based on standardising $\mathbb{F}_n(\alpha)$ (the proportion of p-values below α) and comparing to a standard normal quantile is one way to assess the significance of a large body of independent p-values, and can be used to test our global hypothesis. Donoho and Jin (2004) extended this idea by regarding α as a parameter, considering the whole empirical process one obtains by allowing α to vary and then taking the supremum as in (14).

As foreshadowed above in Section 2.5.1, there is a whole class of higher criticism statistics, based on restricting the interval I_n over which the supremum is taken in (13), (14) or (16). While the primary focus of our theoretical analysis is $HC_n = HC_n(0, 1)$, we also describe two other variants (introduced in Donoho and Jin (2004)) which we consider in our simulations in Section 5. Our theoretical results for HC_n extend to these variants with little difficulty.

Following Donoho and Jin (2004), for a fixed $0 < \alpha_0 < 1$ we define the two variants

$$HC_n^*(\alpha_0) = HC_n(0, \alpha_0)$$

and

$$HC_n^+(\alpha_0) = HC_n(n^{-1}, \alpha_0).$$

REMARK 4.1 (Choice of α_0). We note that Donoho and Jin (2004, 2015) set α_0 to $1/2$. There are good reasons for doing so (see Section 5.2.5), as there is the potential for a loss of power if the true alternative is a normal location under the dense β regime ($\beta < 1/2$) for $\alpha_0 \ll 1/2$.

We henceforth use HC_n^* and HC_n^+ to denote these variants with $\alpha_0 = \frac{1}{2}$.

4.1.2. *Jager–Wellner statistics.* Donoho and Jin (2015) refer to several 'statistics with higher criticism-like construction' that also attain the detection boundary of Proposition 3.1. We will examine two such statistics in more detail, namely the Berk and Jones (1979) statistic R_n^+ (see (17)) and the Jager and Wellner (2007) family of supremum-type φ-divergence statistics $S_n^+(s)$ for some $s \in [-1, 2]$. The latter family of statistics includes (upon re-scaling) R_n^+ and the HC statistics (see Jager and Wellner (2007)).

The maximisation in (17) is restricted to $\mathbb{F}_n(t) > t$ for obvious reasons: we are only interested in one-sided alternatives, where we observe more p-values than expected under H_0. It is perhaps not surprising (given its similarity in construction to HC_n) that R_n^+ also attains the NP detection boundary under the normal location mixture alternative (see Donoho and Jin (2004)).

There are otherways to 'standardise' $\mathbb{F}_n(\alpha)$. We can observe that R_n^+ is derived from a KL divergence from one Bernoulli random variable to another. Similarly HC_n is derived from a divergence between two Bernoulli random variables, as we can interpret $\mathbb{Z}_n^2(\alpha)$ as a re-scaled χ^2 divergence from one Bernoulli to another. Jager and Wellner (2007) recognised this, and introduced a family of statistics constructed with different divergences from a Bernoulli($\mathbb{F}_n(\alpha)$) random variable to a Bernoulli(α) random variable.

DEFINITION 4.2 (Jager and Wellner (2007)). The $S_n^+(s)$ statistic is given by

$$S_n^+(s) = \begin{cases} \sup_{t \in [0,1]} K_s^+(\mathbb{F}_n(t), t) & \text{if } s \in [1, 2], \\ \sup_{t \in [V_{(1)}, V_{(n)}]} K_s^+(\mathbb{F}_n(t), t) & \text{if } s \in [-1, 1), \end{cases}$$

where $K_s^+(u,v) = v\varphi_s(u/v) + (1-v)\varphi_s((1-u)/(1-v))$ when $u \geq v$, but is zero otherwise, and

$$\varphi_s(x) = \begin{cases} (1-s+sx-x^s)/[s(1-s)] & \text{if } s \neq 0, 1, \\ x[\log(x)-1]+1 & \text{if } s = 1, \\ \log(1/x)+x-1 & \text{if } s = 0. \end{cases}$$

Note that φ_s is not the Gaussian density function ϕ.

REMARK 4.3 (Relation to HC_n and R_n^+). The statistics HC_n and R_n^+ are recovered from $S_n^+(s)$ by the transformations $HC_n = \sqrt{2nS_n^+(2)}$ and $R_n^+ = S_n^+(1)$.

This family of statistics is intimately tied to HC_n. We note that Jager and Wellner (2007) used several connections to HC_n to derive the limiting behaviour of $S_n(s)$ (which is a two-sided variant of $S_n^+(s)$ without the restriction $u \geq v$ on $K_s^+(u,v)$ in Definition 4.2) under H_0 and its respective power, where they show that it also attains the detection boundary of Proposition 3.1 under the normal location mixture alternative. We will ultimately use this link to derive our results for $S_n^+(s)$.

It is worth mentioning that there are several other statistics that also attain the detection boundary of Proposition 3.1. They include: the other Berk and Jones (1979) statistic M_n^+ in Gontscharuk and Finner (2017) and Moscovich, Nadler and Spiegelman (2016); the Csörgő et al. (1986) standardisation of HC_n in Stepanova and Pavlenko (2018); the average likelihood ratio test in Walther (2013), which is a compromise between HC_n and R_n^+; the cumulative sum test in Arias-Castro and Wang (2017); and the order statistic test in Laurent, Marteau and Maugis-Rabusseau (2016). Determining higher-order behaviour for these statistics in a similar vein to our work is an interesting avenue of future research, and is beyond the scope of this paper.

4.2. *Limiting null distributions.* Under both examples in Section 3, the null distribution is $N(0,1)$. The nonregular limiting null behaviour of Λ_n was first pointed out in Hartigan (1985) where an approximation like (10) was established between the profile log likelihood $L_n(\theta)$ and a self-normalised version of the process $\{U_n(\theta)\}$ implying that for any finite set $\{\theta_1, \dots, \theta_k\}$,

$$\max_{1 \leq j \leq k} U_n(\theta_j)$$

is an asymptotic lower bound for $\sqrt{2\Lambda_n}$. Furthermore, he conjectured that $\Lambda_n = O_p(\log\log n)$. The results in Bickel and Chernoff (1993) imply that under the null hypothesis the limiting distribution of $T_n = \sup_\theta U_n(\theta)$ is of Gumbel type:

$$\lim_{n\to\infty} P\left(T_n \leq \sqrt{\log\log(n)} + \frac{x - \log(\sqrt{2}\pi)}{\sqrt{\log\log(n)}}\right) = \exp(-e^{-x}).$$

Liu and Shao (2004) went on to show that the approximation (10) is suitably accurate where both the score process $\{U_n(\theta)\}$ and profile log-likelihood $L_n(\theta)$ are maximised, so that $\Lambda_n = T_n^2/2 + o_p(1)$, implying that

$$\lim_{n\to\infty} P\left(\Lambda_n \leq \frac{\log\log(n)}{2} + x - \log(\sqrt{2}\pi)\right) = \exp(-e^{-x}),$$

which confirmed Hartigan's conjecture.

4.2.1. *Higher criticism.* Jaeschke (1979) showed that the limiting null distribution of HC_n is

$$\lim_{n\to\infty} P\left(HC_n \le \sqrt{2\log\log(n)} + \frac{2x + \log\log\log(n) - \log(4\pi) + o(1)}{\sqrt{8\log\log(n)}}\right) = \exp(-e^{-x}),$$

which is of Gumbel type. Its derivation involves several key steps which we outline in Section 4.2.3 below.

REMARK 4.4 (The HC_n variants). The HC_n^+ and HC_n^* variants both share the same distribution as HC_n. (see Shorack and Wellner (1986), Donoho and Jin (2004), Stepanova and Pavlenko (2018)).

4.2.2. *The other statistics.* Berk and Jones (1979) argued that a two-sided variant of nR_n^+, where the restriction $\mathbb{F}_n(t) \ge t$ is dropped, is well approximated (see Wellner and Koltchinskii (2003), who correct an error in the Berk and Jones (1979) proof) by $(HC_n^{ts})^2/2$, where $HC_n^{ts} = \sup_t |\mathbb{Z}_n(t)|$ (see (15)) is a two-sided variant of higher criticism. The limiting distribution (also of Gumbel type) would then follow from results in Jaeschke (1979). One benefit of R_n^+ over HC_n is better finite-sample properties (Walther (2013), Li and Siegmund (2015)).

REMARK 4.5 (Limiting distribution of HC_n^{ts}, see Jaeschke (1979)). The limiting distribution of HC_n^{ts} is given by

$$\lim_{n\to\infty} P\left(HC_n^{ts} \le \sqrt{2\log\log(n)} + \frac{2x + \log\log\log(n) - \log(\pi) + o(1)}{\sqrt{8\log\log(n)}}\right) = \exp(-e^{-x}).$$

This result was generalised by Jager and Wellner (2007), who argued that a two-sided variant of $nS_n^+(s)$, where the restriction $K_s^+(u, v) = 0$ for $u < v$ in Definition 4.2 is dropped, shares the same limiting distribution as $(HC_n^{ts})^2/2$.

4.2.3. *Slow convergence of HC in finite samples.* The slow convergence of HC_n in finite-samples is frequently highlighted (see Donoho and Jin (2004), Jager and Wellner (2007), Walther (2013), Gontscharuk, Landwehr and Finner (2015), Li and Siegmund (2015)).

It is helpful to explain some theoretical reasons for why this is the case, which involves consideration of the key steps required to derive the limiting distribution. First, we approximate $\mathbb{Z}_n(t)$ (a standardised empirical process) with a Gaussian process for $t \in [d_n/n, 1 - d_n/n]$, where d_n diverges to infinity at a sufficiently fast rate. The maximum of this Gaussian process is known to have a limiting distribution of Gumbel type (see Darling and Erdös (1956) or Leadbetter and Rootzén (1988)). Jaeschke (1979) showed that asymptotically the maximum of this Gaussian process dominates the maximiser of $\mathbb{Z}_n(t)$ for $t \notin [d_n/n, 1 - d_n/n]$, so HC_n shares the same limiting distribution.

The derivation involves several approximations and asymptotic results. First, the rate of convergence for the maximum of the approximating Gaussian process to converge to its limiting distribution is no faster than $\log\log(n)^{-1}$ (see Hall (1991), Theorem 2.1). The second asymptotic result, where the maximiser for $t \in [d_n/n, 1 - d_n/n]$ dominates that for $t \notin [d_n/n, 1 - d_n/n]$, is more problematic. We borrow an argument from Donoho and Jin (2004) to show this. Let $V_{(1)}$ denote as the smallest p-value, then we can show that

$$P(\mathbb{Z}_n(V_{(1)}) > x) \xrightarrow{d} 1 - \exp(-(\sqrt{4 + x^2} - x)^2/4),$$

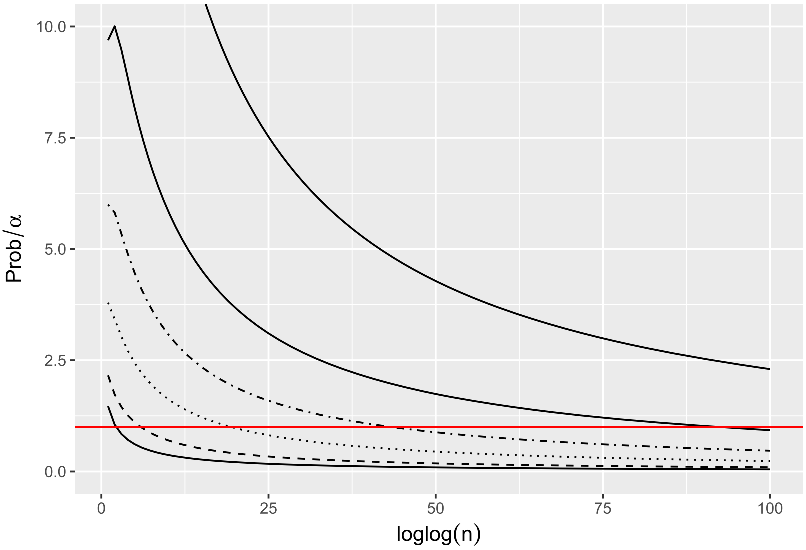

FIG. 3. *The ratio $P(\mathbb{Z}_n(V_{(1)}) > c_n^{HC}(\alpha))/\alpha$, where $c_n^{HC}(\alpha)$ is the Jaeschke (1979) sequence of α-level critical values, is plotted against $\log\log(n)$. The line represents where the ratio is one. The curves from top to bottom correspond to $\alpha = 0.002, 0.005, 0.01, 0.05$ and 0.10, respectively. The $\alpha = 0.05$ curve crosses the curve when $\log\log(n) \approx 5.9$, so $n \approx 3.4 \times 10^{158}$ (this is a very large number).*

which behaves like $x^{-2}[1 + o(1)]$ for large x. The effect of this algebraic tail can be observed in Figure 3, where we plot the probability that $Z_n(V_{(1)})$ exceeds the Jaeschke (1979) α-level sequence of critical values. We can see that the probability exceeds α in a more dramatic fashion as α gets smaller. This means that the behaviour of $\mathbb{Z}_n(t)$ for $t \notin [d_n/n, 1 - d_n/n]$ is still a prominent factor in determining finite-sample Monte Carlo critical values for HC_n.

The motivation behind the HC_n^+ variant is clear; it seeks to avoid such problems for small t under H_0. However, as we shall see in Section 5, there is a trade-off in power when the now 'ignored' $V_{(1)}$ largely determines the power.

4.3. *Power under the normal location mixture alternative.* As part of a wider study on certain signal detection problems Ingster (1997, 2001, 2002) studied a 'symmetrised' version of our normal location mixture model:

$$(1 - p)F_0(x) + \frac{p}{2}F_\theta(x) + \frac{p}{2}F_{-\theta}(x),$$

for $0 \le p \le 1$, $\theta \ge 0$ and F_θ the $N(\theta, 1)$ CDF. All relevant results in this work can be reformulated with minor changes to apply to our 'un-symmetrised' normal location mixture model given by

$$(1 - p)F_0(x) + pF_\theta(x),$$

where $F_\theta(x) = \Phi(x - \theta)$, $\theta \ge 0$, which we assume to be the true distribution of the X_i's for the rest of this section.

Ingster (1997) studied the large-sample properties of the *nonadaptive* Neyman–Pearson (NP) test under sparse local alternatives formulated as in Proposition 3.1, whence we obtain Statement 1. Statement 2 is taken directly from Donoho and Jin (2004).

In Ingster (2001) adaptive tests, applicable when p and θ are both unknown, were considered and an upper bound to power was derived providing statements 1 and 3 of Proposition 3.2 (see Ingster (2001), Theorem 3.1). The following theorem provides Statements 2 and 4 of Proposition 3.2.

THEOREM 4.6. *There exists a sequence u_n that satisfies*

$$u_n \sim \begin{cases} \log\log\log(n) & \textit{if } \beta \in (1/2, 3/4), \\ \dfrac{(1-\sqrt{1-\beta})^2}{\sqrt{1-\beta}} \log\log(n) & \textit{if } \beta \in [3/4, 1) \end{cases}$$

so that the limiting power of the test based on Λ_n is one when $\varepsilon_n \geq u_n$.

The next theorem provides analogous results for the maximal score statistic T_n.

THEOREM 4.7. *There exists a sequence u_n that satisfies*

$$u_n \sim \begin{cases} \log\log\log(n) & \textit{if } \beta \in (1/2, 3/4), \\ 2\dfrac{(1-\sqrt{1-\beta})^2}{\sqrt{1-\beta}} \log\log(n) & \textit{if } \beta \in [3/4, 1) \end{cases}$$

so that the limiting power of the test based on T_n is one when $\varepsilon_n \geq u_n$.

The finite-sample simulations (to appear in Section 5) lead us to conjecture that the constant 2 under $\beta \in [3/4, 1)$ in Theorem 4.7 can be relaxed to match Theorem 4.6.

The following two theorems give us statements 5(a) and 5(b) of Proposition 3.2, respectively.

THEOREM 4.8. *There exists a sequence q_n that satisfies*

$$q_n \sim \max\left(\frac{(1-\sqrt{1-\beta})^2}{\sqrt{1-\beta}}; \frac{1}{2}\right) \log\log(n)$$

so that the limiting power of HC_n tends to α if $\varepsilon_n \leq q_n$.

THEOREM 4.9. *There exists a sequence u_n that satisfies*

$$u_n \sim \max\left(\frac{(1-\sqrt{1-\beta})^2}{\sqrt{1-\beta}}; \frac{1}{2}\right) \log\log(n)$$

so that the limiting power of HC_n tends to one if $\varepsilon_n \geq u_n$.

COROLLARY 4.10 (Corollary to Theorems 4.8 and 4.9). *The results in Theorems 4.8 and 4.9 also apply to HC_n^+ and HC_n^* for any fixed α_0.*

REMARK 4.11. Note that $1/2 \geq (1-\sqrt{1-\beta})^2/\sqrt{1-\beta}$ if and only if $\beta \leq 3/4$, which corresponds to $\rho^*_{\mathrm{NP}}(\beta) = \beta - 1/2$.

An interesting question is whether any member of the φ-divergence statistics can improve on HC_n when $\beta \in (1/2, 3/4)$, as it is stated in Jager and Wellner ((2007), pages 2030–2031) that *'the different Poisson boundary behaviours for $s < 0$ and $s \geq 1$ suggest that the [two-sided variant of $S_n^+(s)$] with $s \geq 1$ are geared toward heavy tails, while the statistics with $s \leq 0$ are geared more toward light tails'*. There are also simulation studies (see Walther (2013), Li and Siegmund (2015) and Moscovich, Nadler and Spiegelman (2016)) that suggest statistics such as R_n^+, which recall is a particular member of $S_n^+(s)$, has greater power than HC_n when $\beta < 3/4$.

The following theorem shows that this is not possible, and despite the additional flexibility in selecting the s, all of the $S_n^+(s)$ statistics suffer a loss of power relative to the GLRT and score test when $\beta \in (1/2, 3/4)$.

THEOREM 4.12. *There exists a sequence q_n that satisfies*

$$q_n \sim \max\left(\frac{(1-\sqrt{1-\beta})^2}{\sqrt{1-\beta}}; \frac{1}{2}\right)\log\log(n)$$

so that the limiting power of a modified version of $S_n^+(s)$, where the underlying process is maximised with the added restriction $\mathbb{F}_n(t) \geq 1/2$, is α if $\varepsilon_n \leq q_n$.

The power deficiency of HC_n is shared by $S_n^+(s)$ at the level of detail that we have analysed, and can be summarised by the following corollary.

COROLLARY 4.13 (Corollary to Theorem 4.12). *Suppose that $\beta \in (1/2, 3/4)$ and $\alpha = \alpha_n$ tends to zero slowly enough. There are choices of θ_n so that*:

1. *the power of the homoscedastic normal GLRT and maximal score test tend to one*;
2. *the power of the modified $S_n^+(s)$ with the restriction detailed in Theorem 4.12 tends to zero for each $s \in [-1, 2]$.*

4.4. *Power under the truncated normal mixture alternative.* We now provide analogous theoretical results which imply Propositions 3.3 and 3.4. It is convenient to change notation slightly. We assume for the rest of this section that the p-values V_i have common uniform scale mixture distribution

$$(1-q)U(0,1) + qU(0,\nu) \tag{20}$$

for $0 < q, \nu < 1$. Thus we replace p and θ at (19) in Section 3.2 with q and $1-\nu$, respectively. This will facilitate our theoretical developments below. Our first theorem below is proved in the Supplementary Material.

THEOREM 4.14. *The full GLRT statistic for testing $H_0 : V_i \overset{d}{=} U(0,1)$ against $H_1 : V_i \overset{d}{=} (1-q)U(0,1) + qU(0,\nu)$ where $(q,\nu) \in (0,1)^2$ is the Berk and Jones (1979) statistic nR_n^+.*

Our next theorem relates to the NP test between the $U(0,1)$ global null hypothesis and the uniform scale mixture alternative (20) above. It requires knowledge of the alternative parameters but is the most powerful test. It provides analogues of the key results from Ingster (1997) for the normal location mixture.

THEOREM 4.15 (NP test). *Suppose that $q = o(n^{-1/2})$ and $q = o(\nu)$, then the NP test has limiting power α if $nq^2\nu^{-1} = o(1)$ and limiting power one if $nq^2\nu^{-1}$ diverges to ∞.*

COROLLARY 4.16 (The NP detection boundary). *Let $q = n^{-\beta}$ and $\nu = n^{-r}$ for some $(r,\beta) \in (0,1) \times (1/2,1)$. The NP test has limiting power α if r is a fixed point satisfying $r < 2\beta - 1$, and has limiting power one if r is a fixed point satisfying $r > 2\beta - 1$.*

We provide the following result analogous to Donoho and Jin ((2004), Theorem 1.4) which we use to illustrate an interesting property of higher criticism in this context.

PROPOSITION 4.17. *Let $q = o(n^{-1/2})$ and $nq \to \infty$. If $q\nu^{-1} = o(1)$, then the limiting powers of the minimum p-value test* (*Bonferroni test*) *and the Benjamini and Hochberg* (1995) *FDR procedure are both α. If $q\nu^{-1}$ diverges to ∞, then the limiting powers of the minimum p-value test* (*Bonferroni test*) *and the Benjamini and Hochberg* (1995) *procedure are both one.*

PROPOSITION 4.18. *Let $q = o(n^{-1/2})$ and $nq \to \infty$. The limiting power of HC_n is one if $\sqrt{nq}\sqrt{\nu^{-1}-1} - c_n^{HC}$ diverges, and is α if $\sqrt{nq}\sqrt{\nu^{-1}-1} - c_n^{HC} = o(1/c_n^{HC})$, where c_n^{HC} is a sequence of critical values for HC_n.*

It follows from Proposition 4.17 that a Donoho and Jin (2004)-style detection boundary for the Bonferroni test is β, so that the limiting power is one (α, resp.) if $r > \beta$ ($r < \beta$, resp.) for any fixed r. The Bonferroni (and FDR procedures) do not 'attain' the NP detection boundary *for any* $\beta \in (1/2, 1)$; for each $\beta \in (1/2, 1)$ we can find $r \in (2\beta - 1, \beta)$ such that the limiting power of the NP test is one, whereas the limiting power of the Bonferroni test is α.

The same is not true for HC, which does 'attain' the NP detection boundary (see Proposition 4.18). The implication of both Propositions 4.17 and 4.18 is that power is not largely determined by the minimum p-value $V_{(1)}$. Finally, the following remark gives us Statement 3 in Proposition 3.4.

REMARK 4.19. The (normal location mixture) maximal score test based on T_n and full GLRT based on Λ_n share the NP detection boundary. It is possible to modify the proof of Theorem 4.7 to show that a sufficient condition for the limiting power of both tests to be one is

$$\sqrt{nq}\sqrt{\frac{1}{\nu} - 1} \geq K \log(n)^{1/4}\sqrt{\log\log(n)},$$

for some $K > 0$. A proof of this appears in the Supplementary Material.

5. Simulations.

Outline. We explore finite-sample empirical behaviour of Λ_n, T_n, R_n^+, HC_n^* (with $\alpha_0 = 1/2$) and HC_n^+ (with $\alpha_0 = 1/2$), but note that HC_n has almost identical behaviour to HC_n^* in all cases.

In the following sections, the test statistics are each sampled 10^4 times for each sample size (10^2, 10^3, 10^4 and 10^5) under the null hypothesis in Section 5.1, normal location mixture alternative in Section 5.2.1, and truncated normal mixture alternative in Section 5.2.2. We calculate the empirical power of a test as the proportion of times the sampled statistics exceeded their respective Monte Carlo calibrated critical value under the null.

We represent the results graphically using receiver operator characteristic (ROC) curves, which plot the power as a function the level, but we restrict the level to the range $(0, 0.1)$. Below we highlight the main patterns using a small selection of ROC curves; the full set can be found in the Supplementary Material.

5.1. *Behaviour under the null hypothesis.* Consider the following Monte Carlo calibrated critical values for each statistic in Table 1 at the 5% and 1% significance levels.

REMARK 5.1 (Decreasing critical values for T_n). The asymptotic critical values for T_n appear to be decreasing as n increases at the 1% significance level, which despite its paradoxical nature is not a mistake. The asymptotic critical values can be plotted against n to observe that the asymptotic critical values only start to increase for some $n \gg 10^9$.

Main observation. At least for the sample sizes that we analysed, it seems that the asymptotic critical values *can* generally be used instead of the Monte Carlo critical values for Λ_n. The resultant procedure is conservative, so that if H_0 is rejected at the α-level using the

TABLE 1

The Monte Carlo critical values with the asymptotic Gumbel distribution critical values in parentheses. The asymptotic results for HC_n^+ are excluded, which are the same as that for HC_n^*

Sample size	R_n^+	HC_n^*	HC_n^+	Λ_n	T_n
			5% significance level		
100,000	5.65 (4.59)	4.44 (3.18)	3.29	2.23 (2.70)	2.19 (2.51)
10,000	5.29 (4.32)	4.69 (3.11)	3.20	2.09 (2.59)	2.16 (2.48)
1000	5.03 (3.97)	4.92 (3.00)	3.17	2.08 (2.45)	2.20 (2.45)
100	4.61 (3.44)	4.77 (2.84)	3.01	1.95 (2.24)	2.22 (2.43)
			1% significance level		
100,000	7.33 (6.22)	9.13 (3.92)	3.93	3.79 (4.33)	3.17 (3.55)
10,000	7.01 (5.95)	9.53 (3.88)	3.93	3.62 (4.22)	3.28 (3.58)
1000	6.86 (5.60)	10.44 (3.83)	3.97	3.57 (4.08)	3.39 (3.63)
100	6.25 (5.07)	10.48 (3.78)	3.86	3.42 (3.87)	3.55 (3.75)

asymptotic critical values, then it will also be rejected at the α-level if the Monte Carlo critical values were used instead. If Λ_n with the α-level asymptotic critical values fail to reject H_0, then the Monte Carlo critical values are perhaps needed to verify this.

This is a major benefit over the HC and Berk and Jones (1979) statistics, which are typically anti-conservative. Our results reaffirm existing results (see Donoho and Jin (2004), Walther (2013) and Li and Siegmund (2015)) on the empirical heavy tail of HC_n^*. We described this theoretically in Section 4.2.3; however, the influence of the heavy tail is unmistakable for the smaller significance levels, where the critical values of HC_n^* are disproportionately larger than say that of HC_n^+, which possesses attenuated tails.

REMARK 5.2 (A similar observation for T_n). We can also use the asymptotic critical values instead of the Monte Carlo ones for T_n with the caveat that the resulting procedure may be anti-conservative for small α-levels. In particular, when α is smaller than 0.0072 when $n = 10^2$ or $n = 10^3$, and 0.005 when $n = 10^4$ or $n = 10^5$. This can be seen in the Supplementary Material. However, a larger simulation study is required to assess this properly.

5.2. *Behaviour under mixture alternatives.* We simulated statistics for three alternatives 10^4 times for each different parametrisation.

5.2.1. *The normal location mixture.* We simulated the normal location mixture alternative over a grid $\beta \in \{0.55, 0.60, \dots, 0.95\}$,

$$\begin{aligned}\theta^2 = \theta_C^2 &= 2\rho_{\mathrm{NP}}^*(\beta)\log(n) + \varepsilon_n \\ &\text{with } \varepsilon_n = C\max\bigl(1/2; \rho_{\mathrm{NP}}^*(\beta)(1-\beta)^{-1/2}\bigr)\log\log(n)\end{aligned} \tag{21}$$

and $n = 100, 1000, 10^4, 10^5$. We set $C = 2$ to provide HC with near 'ideal' asymptotic power (cf. Theorems 4.8 and 4.9, where $C = 1 + o(1)$). See Figure 4 for ROC curves for $\beta = 0.6, 0.7, 0.8, 0.9$ and $n = 10^5$; the remaining curves can be found in the Supplementary Material.

5.2.2. *The uniform scale mixture.* We simulated the uniform scale mixture alternative over a grid of $q = n^{-\beta}$ for $\beta \in \{0.55, 0.60, \dots, 0.95\}$ with $\nu = n^{1-2\beta}/(8\log\log(n))$ and $n = 100$, 1000, 10^4, 10^5. We transformed the statistics from the p-value scale to the z-score scale using $\Phi^{-1}(1 - \cdot)$ (giving the truncated normal mixture) to simulate Λ_n and T_n. See Figure 5 for ROC curves for $\beta = 0.6, 0.7, 0.8, 0.9$ and $n = 10^5$; the remaining curves can be found in the Supplementary Material.

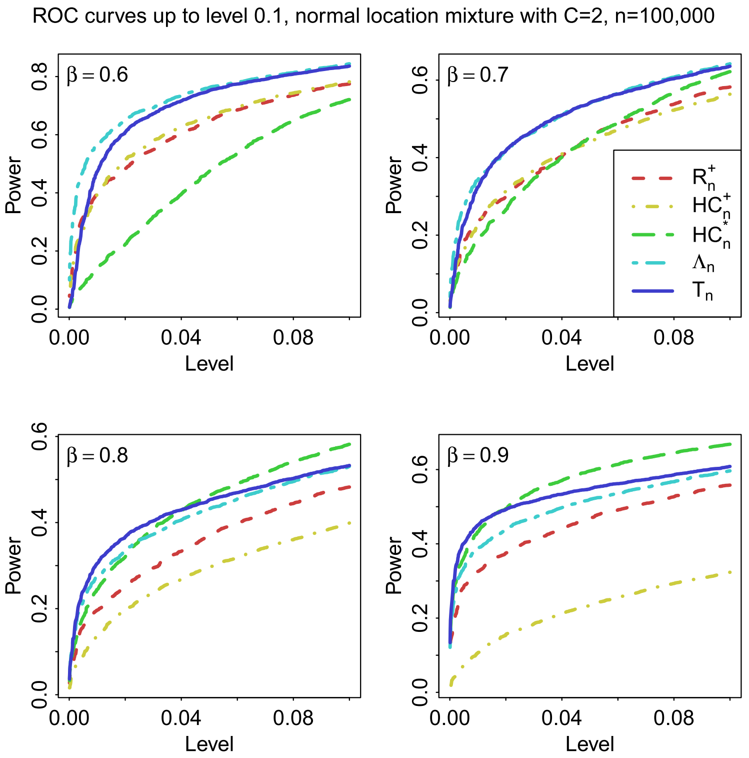
ROC curves up to level 0.1, normal location mixture with C=2, n=100,000

FIG. 4. *The graphs above plot power against level, for levels* $0 < \alpha < 0.1$ *for each test, with data simulated from a normal location mixture alternative. The solid curve is the maximal EMGF statistic* T_n, *the cyan double-dashed is the full GLRT statistic for the normal location mixture model* Λ_n, *the shorter-dashed is the Berk–Jones statistic* R_n^+, *the long-dashed is the unrestricted HC statistic* HC_n^* *and the dot-dashed is the restricted HC statistic* HC_n^+.

5.2.3. *The uniform-prior normal location mixture.* Following a suggestion of a referee we also examined a third mixture alternative scenario where the contaminating normal location shift changes from observation to observation. Thus $\theta_1, \dots, \theta_n$ are i.i.d. $U[\theta_C, \theta_D]$ distribution, where θ_C^2 is as defined at (21) in Section 5.2.1 above; $I_1, \dots, I_n$ are iid $B(1, n^{-\beta})$; conditionally, $X_i | (\theta_i, I_i) \sim N(I_i \theta_i, 1)$. Unconditionally, the X_i's are i.i.d. with density

$$(1 - n^{-\beta})\phi(x) + \frac{n^{-\beta}}{\theta_D - \theta_C} \int_{\theta_C}^{\theta_D} \phi(x - \theta)\, d\theta.$$

We set $C = 2$ and $D = 3$ so that the contaminating means were at least as large as under the normal location mixture alternatives in Section 5.2.1. The general patterns of behaviour were mostly the same as in that case. See Figure 6.

REMARK 5.3. We observed that a test that rejects H_0 when the sample maximum $X_{(n)}$ is large (Bonferroni test) has near identical power to HC_n^*, and is excluded from the plot for clarity.

5.2.4. *Observations.* The empirical performance of HC_n^+ and HC_n^* is contingent on the influence of the maximal extreme-order statistics on the power. The maximal extreme-order statistics are highly influential when $\beta > 3/4$, but less so when $\beta < 3/4$. The attenuation of these extreme-order statistics in the definition of HC_n^+, also attenuates the power of HC_n^+ when these extreme-order statistics are influential; for example, when $\beta > 3/4$ under the normal location mixture alternative.

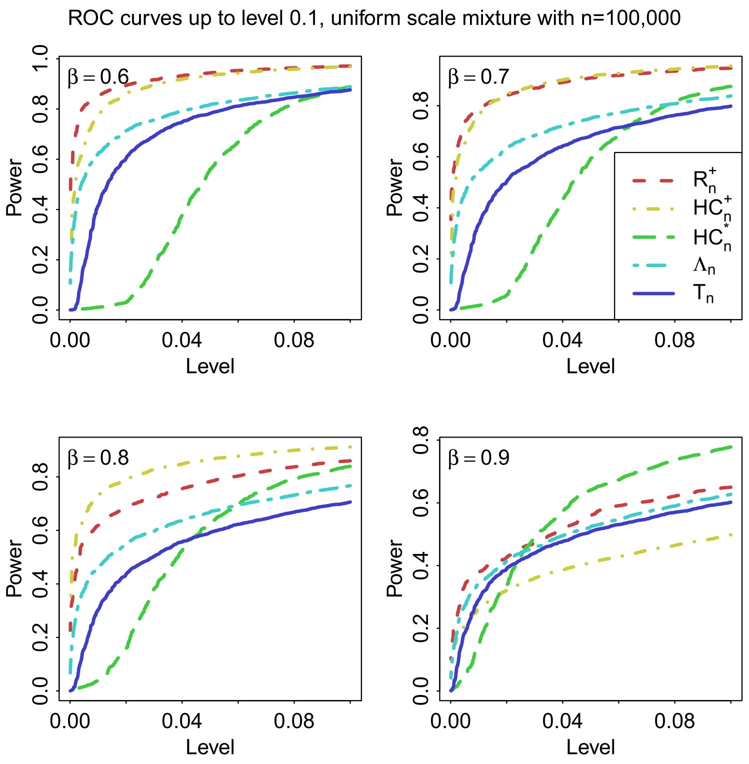

FIG. 5. *The graphs above are the same as for Figure 4 except the data are simulated from a uniform scale/truncated normal mixture alternative.*

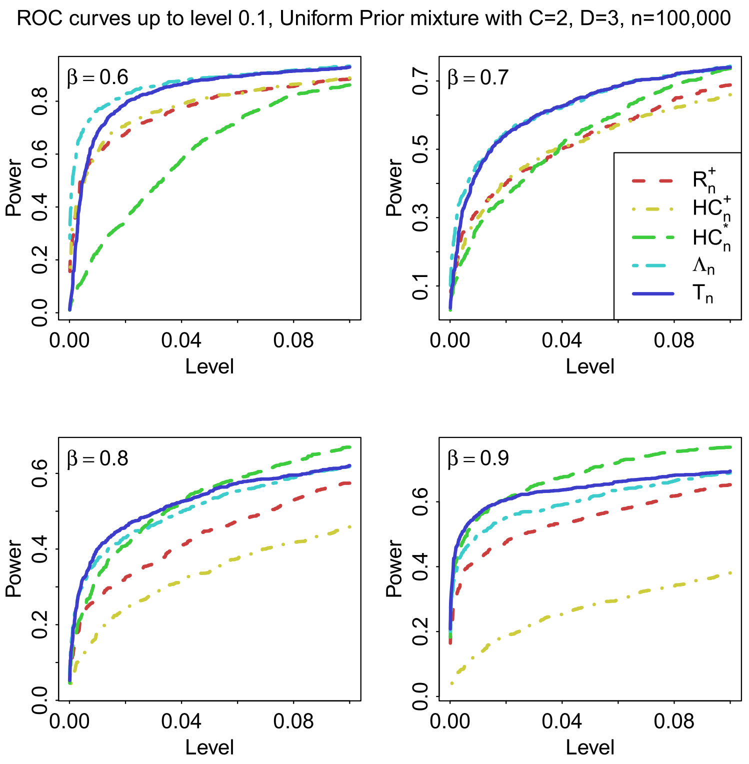

FIG. 6. *The graphs above are the same as for Figure 4 except the data are simulated from a uniform scale/truncated normal mixture alternative.*

However, the attenuation of the same extreme-order statistics can improve the power if these statistics do not largely determine the power. This is the case for all $\beta \in (1/2, 1)$ under the uniform scale mixture alternative, and for all $\beta \in (1/2, 3/4)$ under the normal location mixture alternative. Under these conditions with small α-levels, the extreme-order statistics largely determine the behaviour of HC_n^*. The asymptotics for HC have not come into effect, and we found that it is not clear that HC_n^* is empirically better than the Bonferroni test.

The implication of the differing powers of HC_n^+ and HC_n^* presents a dilemma. We do not know β and do not know whether the extreme-order statistics determine the power, but if we use the wrong variant of HC then the power can be very poor compared to other methods.

Key messages. The culmination of our simulation studies leads us to the following recommendations for using HC:

1. If you decide to use HC, then you should use the attenuated variant HC_n^+ (or something similar).
2. If the underlying process for HC_n^+ is maximised near $1/n$, then other more sophisticated methods, such as the normal location mixture score and GLRTs, should be considered instead.

In the case of the second point above, there has to be some care taken when using multiple procedures as the α-levels for the procedures will need to be adjusted.

5.2.5. *The dense regime.* A higher-order analysis of HC when $\beta < 1/2$ is currently unknown, but we hope to provide some insight. The variance of $\mathbb{Z}_n(t)$ is $1 + o_p(1)$, and so its expected value largely determines the power, which is maximised in some neighbourhood of $t = 1/2$. Therefore, the power of HC_n is approximately

$$P\left(n\mathbb{F}_n(1/2) > \frac{n}{2} + \sqrt{\frac{n \log\log(n)[1 + o(1)]}{2}}\right),$$

which tends to α (or one) when $\theta_n \overset{(>)}{<} \sqrt{\pi}\sqrt{\log\log(n)}n^{\beta-\frac{1}{2}}$. There would be a loss of power, which would also apply to R_n^+ and $S_n^+(s)$ by an extension of Theorem 4.12, when compared to the correctly specified parametric tests.

The definitions of HC_n^+ and HC_n^* require a choice of α_0. We refer the reader to Remark 4.1 and note that unlike the sparse regime with $\beta > 1/2$, the choice of α_0 is important under the dense regime. The HC process seems to be maximised in some neighbourhood of $t = \alpha_0$ if $\alpha_0 \ll 1/2$. This suggests that there may be a potential loss of power if α_0 is too small, which perhaps offsets a potential gain in power under the sparse regime for small α_0.

6. Discussion.

HC has been recommended and used in a large variety of situations, almost all motivated by an underlying need to detect a sparse normal mixture.

In the original paper (Donoho and Jin (2004)), it was suggested that HC was preferable to the parametric ‘tailor-made’ procedure (normal location mixture GLRT) for three reasons, namely:

1. it was not clear that the GLRT works well;
2. the procedure in Ingster (2002) has good power, but is delicate and complex; and
3. HC is simple, intuitive, and not tied to a ‘narrow’ model.

Our results show that the GLRT possesses many of the desirable properties of Ingster’s delicate complex procedure, but is simpler to implement. Ingster’s optimal procedure involves constructing a Bonferroni test using approximate p-values based on values of a score process evaluated over a progressively finer grid as n diverges to infinity. This construction suggests also considering the maximiser of the score process T_n as a test statistic, which shares many of the desirable properties of the GLRT; the version $\widetilde{T}_n$ which restricts the maximisation to X_i-values seems particularly appealing due to its ease of implementation. These statistics are natural parametric procedures under the normal location mixture, so it is not surprising that they do well.

The HC and Berk and Jones (1979) statistics (HC_n and R_n^+, resp.) may also be interpreted as natural parametric procedures for a *different* mixture model. In this sense, they may be considered as 'tied' to a 'narrow' uniform scale mixture model in the same way that the normal Λ_n and T_n are tied to the normal location mixture model. In light of this, it is perhaps not surprising that each pair of statistics may have superior performance under the model for which they are 'tailor-made'.

Other statistics such as the Jager and Wellner (2007) supremum-type φ-divergence statistics, the Berk and Jones (1979) statistic M_n^+, the average likelihood ratio test in Walther (2013) and the Csörgő et al. (1986) standardisation of HC_n in Stepanova and Pavlenko (2018) may all be viewed as re-weighted likelihood ratio tests for the uniform-scale mixture model. In this sense, they are all tied to the uniform scale mixture model. We showed that the φ-divergence statistics suffered a theoretical loss of power under the normal location mixture model, and it would not be surprising if the others also do.

Our simulations may have several practical implications in finite-sample applications. We found that the asymptotic critical values for Λ_n and T_n can be used instead of the Monte Carlo ones to produce conservative procedures—we remark that it is unknown whether this is true for T_n if α is small. However, the same is not true for HC and R_n^+.

There are practical problems with HC due to the heavy tail of the underlying process. The choice between HC_n^* and HC_n^+ depends on the nature of the alternative, which is not ideal. We agree with the recommendation (see Donoho and Jin (2004)) to use HC_n^+ over HC_n^*, but we also highlighted that HC_n^+ should be viewed more cautiously when the underlying process is maximised near $1/n$. We find that Λ_n and T_n can address several of these problems with greater power under the normal location mixture alternative. This suggests that T_n is worth considering in practice.

We finish by remarking that although both the normal location and uniform scale mixture models are *toy models*, this should not prohibit their application to more complex problems. We need only consider the wide-range of problems that HC has had some success to illustrate that *simple models often work*. And in cases where they do not, it begs the question, 'why not use a different simple model instead'.

Acknowledgements. We would like to thank the Associate Editor and three referees for their constructive feedback, which has helped to vastly improve the presentation and several thematic elements of this paper.

SUPPLEMENTARY MATERIAL

Supplement to 'Beyond HC: More sensitive tests for rare/weak alternatives' (DOI: 10.1214/19-AOS1885SUPP; .pdf). We prove the main theorems, and provide additional supporting plots that show performance of the maximal score test in several examples.

REFERENCES

ARIAS-CASTRO, E. and WANG, M. (2017). Distribution-free tests for sparse heterogeneous mixtures. *TEST* **26** 71–94. MR3613606 https://doi.org/10.1007/s11749-016-0499-x

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47–59. MR0521531 https://doi.org/10.1007/BF00533250

BICKEL, P. and CHERNOFF, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability*: *A Raghu Raj Bahadur Festschrift* 83–96. Wiley, New York.

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040 https://doi.org/10.3150/bj/1106314847

CAI, T. T., JENG, X. J. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 629–662. MR2867452 https://doi.org/10.1111/j.1467-9868.2011.00778.x

CAI, T. T. and WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory* **60** 2217–2232. MR3181520 https://doi.org/10.1109/TIT.2014.2304295

CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Stat.* **25** 573–578. MR0065087 https://doi.org/10.1214/aoms/1177728725

CSÖRGŐ, M., CSÖRGŐ, S., HORVÁTH, L. and MASON, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85. MR0815960

DARLING, D. A. and ERDÖS, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J.* **23** 143–155. MR0074712

DONOHO, D. (2017). 50 years of data science. *J. Comput. Graph. Statist.* **26** 745–766. MR3765335 https://doi.org/10.1080/10618600.2017.1384734

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506

GONTSCHARUK, V. and FINNER, H. (2017). Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Comm. Statist. Theory Methods* **46** 2332–2342. MR3576717 https://doi.org/10.1080/03610926.2015.1041985

GONTSCHARUK, V., LANDWEHR, S. and FINNER, H. (2015). The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biom. J.* **57** 159–180. MR3298224 https://doi.org/10.1002/bimj.201300255

HALL, P. (1991). On convergence rates of suprema. *Probab. Theory Related Fields* **89** 447–455. MR1118558 https://doi.org/10.1007/BF01199788

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764

HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, *Vol. II* (*Berkeley*, *Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 807–810. Wadsworth, Belmont, CA. MR0822066

INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

INGSTER, Y. I. (2001). Adaptive detection of a signal of growing dimension. I. *Math. Methods Statist.* **10** 395–421. MR1887340

INGSTER, Y. I. (2002). Adaptive detection of a signal of growing dimension. II. *Math. Methods Statist.* **11** 37–68. MR1900973

JAESCHKE, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.* **7** 108–115. MR0515687

JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962 https://doi.org/10.1214/0009053607000000244

JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. MR3468343

LAURENT, B., MARTEAU, C. and MAUGIS-RABUSSEAU, C. (2016). Non-asymptotic detection of two-component mixtures with unknown means. *Bernoulli* **22** 242–274. MR3449782 https://doi.org/10.3150/14-BEJ657

LEADBETTER, M. R. and ROOTZÉN, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.* **16** 431–478. MR0929071

LI, J. and SIEGMUND, D. (2015). Higher criticism: p-values and criticism. *Ann. Statist.* **43** 1323–1350. MR3346705 https://doi.org/10.1214/15-AOS1312

LIU, X. and SHAO, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference* **123** 61–81. MR2058122 https://doi.org/10.1016/S0378-3758(03)00138-1

MOSCOVICH, A., NADLER, B. and SPIEGELMAN, C. (2016). On the exact Berk–Jones statistics and their p-value calculation. *Electron. J. Stat.* **10** 2329–2354. MR3544289 https://doi.org/10.1214/16-EJS1172

PORTER, T. and STEWART, M. (2020). Supplement to "Beyond HC: More sensitive tests for rare/weak alternatives." https://doi.org/10.1214/19-AOS1885SUPP.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0838963

STEPANOVA, N. and PAVLENKO, T. (2018). Goodness-of-fit tests based on sup-functionals of weighted empirical processes. *Teor. Veroyatn. Primen.* **63** 358–388. MR3796493 https://doi.org/10.4213/tvp5160

TUKEY, J. W. (1976). T13 N: The higher criticism. Technical report.

TUKEY, J. W. (1989). Higher criticism for individual significances in several tables or parts of tables. Technical report.

TUKEY, J. W. (1994). The problem of multiple comparisons. In *The Collected Works of John W. Tukey. Vol. VIII* lxii+475+i10. CRC Press, New York.

WALTHER, G. (2013). The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. In *From Probability to Statistics and Back*: *High-Dimensional Models and Processes*. *Inst. Math. Stat.* (*IMS*) *Collect.* **9** 317–326. IMS, Beachwood, OH. MR3202643 https://doi.org/10.1214/12-IMSCOLL923

WELLNER, J. A. and KOLTCHINSKII, V. (2003). A note on the asymptotic distribution of Berk–Jones type statistics under the null hypothesis. In *High Dimensional Probability*, *III* (*Sandjberg*, 2002). *Progress in Probability* **55** 321–332. Birkhäuser, Basel. MR2033896