Electronic Journal of Statistics
Vol. 13 (2019) 208–230
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1512

# Detection of sparse mixtures: higher criticism and scan statistic*

**Ery Arias-Castro**

*e-mail:* earisca@ucsd.edu

**Andrew Ying**

*e-mail:* anying@ucsd.edu

**Abstract:** We consider the problem of detecting a sparse mixture as studied by Ingster (1997) and Donoho and Jin (2004). We consider a wide array of base distributions. In particular, we study the situation when the base distribution has polynomial tails, a situation that has not received much attention in the literature. Perhaps surprisingly, we find that in the context of such a power-law distribution, the higher criticism does not achieve the detection boundary. However, the scan statistic does.

**Keywords and phrases:** Sparse mixtures, contamination model, rare effects, normal means model, higher criticism, scan statistic.

Received February 2018.

## Contents



*Both authors are with the Department of Mathematics, University of California, San Diego, USA. Contact information is available here and here. We are grateful to a reviewer for suggesting references and the part of the discussion on multiple testing (which happens to be related to some separate work of ours). This work was partially supported by a grant from the US Office of Naval Research (N00014-13-1-0257).

## 1. Introduction

We consider the problem of detecting a sparse mixture. A simple variant of the problem can be formulated as follows. Let $F$ be a continuous distribution function on the real line, and $\varepsilon \in (0, 1/2]$ and $\mu > 0$. The problem is to test

$$\mathcal{H}_0^n : X_1, \ldots, X_n \overset{\text{iid}}{\sim} F, \tag{1}$$

versus

$$\mathcal{H}_1^n : X_1, \ldots, X_n \overset{\text{iid}}{\sim} (1-\varepsilon)F(\cdot) + \varepsilon F(\cdot - \mu). \tag{2}$$

Mixtures models such as in (2) have been considered for quite some time, particularly in the context of robust statistics, where they are known as contamination models (Huber and Ronchetti, 2009, Eq 1.22).

Rather, our contribution is in line with the testing problems studied by Ingster (1997) in the context of the normal sequence model, where $F$ above corresponds to the standard normal distribution. In that setting, Ingster considered the following parameterization

$$\varepsilon = \varepsilon_n = n^{-\beta}, \qquad \mu = \mu_n = \sqrt{2r \log n}, \tag{3}$$

for some $\beta > 0$ and $r > 0$. The advantage of this parameterization is that, holding $\beta$ and $r$ fixed, the situation admits a relatively simple description. Indeed, since both the null and the alternative hypotheses are simple, by the Neyman-Pearson Lemma, the likelihood ratio test (set at level $\alpha$) is most powerful. Ingster studied the large-sample behavior of this test procedure and discovered that, in the case where $\beta > 1/2$, when $r < \rho(\beta)$, the test is powerless in the sense of achieving power $\alpha$, while when $r > \rho(\beta)$, the test was fully powerful in the sense of achieving power 1, where the function $\rho$ is given by

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

Thus the existence of a detection boundary in the $(\beta, r)$ plane given by $r = \rho(\beta)$. In such a situation, we will say that a test procedure 'achieves the detection boundary', or is 'first-order optimal' (or simply 'optimal'), if it is fully powerful when $r > \rho(\beta)$.

Such detection boundaries where derived for other models, for example, in (Cai and Wu, 2014; Cai, Jeng and Jin, 2011; Donoho and Jin, 2004). We also mention that the situation where $\beta \le 1/2$ is also well-understood, but quite different, and will not be considered here. Most of the literature has focused on the more interesting setting where $\beta > 1/2$ and we do the same here.

### 1.1. Threshold tests

After determining what one can hope for, it becomes of interest to understand what one can achieve with less information. Indeed, the likelihood ratio test requires knowledge of all the quantities and objects defining the testing problem, in this case $(F, \varepsilon, \mu)$, and even in the present stylized setting we might want to know what can be done when some of this information is missing, in particular what defines the alternative, namely $(\varepsilon, \mu)$. (The case where $F$ is also unknown has attracted much less attention. We discuss it in Section 5.)

When $F$ is known, the problem is that of goodness-of-fit testing, albeit with alternatives of the form (2) in mind. Donoho and Jin (2004) opened this investigation with the analysis of various tests, including the max test based on $\max_i X_i$ and a variant of the Anderson-Darling test (Anderson and Darling, 1952). Seeing as a problem of multiple testing based on p-values defined as $U_i = 1 - F(X_i)$, the max test coincides with the Tippett-Šidák test combination test, while the Anderson-Darling test coincides with a proposal by Tukey called the higher criticism (HC). More recently, Moscovich, Nadler and Spiegelman (2016) analyzed a goodness-of-fit (BJ) test proposed by Berk and Jones (1979) in the same setting. For $t \in \mathbb{R}$, define

$$N_n(t) = \#\{i \in [n] : X_i \geq t\}.$$

We note that, under the null hypothesis, $N_n(t)$ is binomial with parameters $(n, 1 - F(t))$, which motivates the test that rejects for large values of

$$\sup_{t:F(t)\geq 1/2} \frac{N_n(t) - n(1 - F(t))}{\sqrt{nF(t)(1 - F(t)) + 1}}.$$

This is one of many possible variants of HC.[1]

Let $U_{(1)} \leq \cdots \leq U_{(n)}$ denote the ordered $U_i$'s. We note that, under the null hypothesis, $U_{(i)}$ has the beta distribution with parameters $(i, n - i + 1)$, which motivates the definition of BJ, rejecting for small values of

$$\min_{i\in[n]} P_i, \tag{4}$$

where $P_i := \mathsf{B}(U_{(i)}; i, n - i + 1)$ and $\mathsf{B}(\cdot; a, b)$ denotes the distribution function of the beta distribution with parameters $(a, b)$.

The verdict is the following. In the normal setting, HC and BJ achieve the detection boundary in the full range $\beta > 1/2$, while the max test is only able to achieve the detection in the upper half of the range $\beta > 3/4$. The same extends to other models, in particular to generalized Gaussian models where $F$ has density proportional to $\exp(-|x|^a/a)$ for some $a > 1$. (The case $a \leq 1$ is

[1] The constraint '$F(t) \leq 1/2$ can be replaced by $F(t) \leq \gamma$, where $\gamma$ can be taken to be smaller, say $\gamma = 0.05$. The '+1' in the denominator is roughly equivalent to adding the constraint $F(t) \geq 1/n$, which Donoho and Jin (2004) recommend for reasons of stability. In any case, this variant performs as well (to first order) as any other variant of HC considered in the literature, at least in all the regimes commonly considered.

qualitatively different. HC and BJ are still first-order optimal while the max test is suboptimal everywhere.)

These tests are all threshold tests, where we define a threshold test as any test with a rejection region of the form $\bigcup_{t\in\mathcal{T}}\{N_n(t) \geq c_t\}$, for some subset $\mathcal{T} \subset \mathbb{R}$ and some critical values $c_t > 0$. More broadly, any combination test that we know of that is discussed in the multiple-testing literature is a threshold test. (This includes the tests proposed by Fisher, Lipták-Stouffer, Tippett-Šidák, Simes, and more.) Thus it might be of interest to understand what can be achieved with a threshold test. In this regard, it is useful to examine how one would optimize such an approach if one had perfect knowledge of the model. Let $\phi_t$ denote the test with rejection region $\{N_n(t) \geq c_t\}$, where

$$c_t := \min\big\{c \geq 0 : \mathbb{P}_0(N_n(t) \geq c) \leq \alpha\big\}.$$

We define the oracle threshold test as the test $\phi_{t*}$, where

$$t_* := \arg\max_{t\in\mathbb{R}} \mathbb{P}_1(N_n(t) \geq c_t), \tag{5}$$

with $\mathbb{P}_0$ denoting the distribution under the null (1) and $\mathbb{P}_1$ that under the alternative (2). (Here and elsewhere, $\alpha$ denotes the desired significance level.) Note that computing $c_t$ only requires knowledge of $F$, while computing $t_*$ requires knowledge of the entire model, namely $(F, \varepsilon, \mu)$. Thus the construction of the test $\phi_{t*}$ relies on the oracle knowledge of $(\varepsilon, \mu)$.

### *1.2. Scan tests*

Detection problems arise in a variety of contexts and in very many applications. An important example is in spatial statistics (itself a rather wide area), where the detection of 'hot spots', meaning areas of unusually high concentration, has been considered for quite some time (Kulldorff, 1997). An early contribution to this literature is that of Naus (1965), who considered the distribution of the maximum number of points in an interval of given length (say $\ell$) when the points are drawn iid from the uniform distribution on $[0, 1]$. This would nowadays be referred to as the scan statistic and arises when testing the null that the points are uniformly distributed in $[0, 1]$ against the (composite) alternative that there is an sub-interval of length $\ell$ with higher intensity. Settings where sub-interval length is unknown have been considered (Arias-Castro, Donoho and Huo, 2005).

For $s \leq t$, define $N_n[s, t] = \#\{i \in [n] : X_i \in [s, t]\}$ and $F[s, t] = F(t) - F(s)$. We note that, under the null hypothesis, $N_n[s, t]$ is binomial with parameters $(n, F[s, t])$, which motivates the test that rejects for large values of

$$\sup_{(s,t):F[s,t]\leq 1/2} \frac{N_n[s, t] - nF[s, t]}{\sqrt{nF[s, t](1 - F[s, t]) + 1}}. \tag{6}$$

Although there are many possible variants, this is the one we will be working with.

We note that, under the null hypothesis, for any pair of indices $i < j$, $U_{(j)} - U_{(i)}$ has the beta distribution with parameters $(j-i, n-j+i+1)$ — see (Gibbons and Chakraborti, 2011, Th 11.1). This motivates the definition of the scan test which rejects for small values of

$$\min_{0 \le i < j \le n+1} P_{i,j}, \tag{7}$$

where $P_{i,j} := \mathsf{B}(U_{(j)} - U_{(i)}; j-i, n-j+i+1)$, $U_{(0)} := 0$, $U_{(n+1)} := 1$, $P_{0,n+1} := 1$.

In general, we define a scan test as any test with region rejection of the form $\bigcup_{(s,t)\in\mathcal{K}} \{N_n[s,t] \ge c_{s,t}\}$, where $\mathcal{K}$ is a subset of $\{(s,t) : s < t\}$ and $c_{s,t} \ge 0$ are critical values. Let $\phi_{s,t}$ denote the test with rejection region $\{N_n[s,t] \ge c_{s,t}\}$, where

$$c_{s,t} := \min \big\{ c \ge 0 : \mathbb{P}_0(N_n[s,t] \ge c) \le \alpha \big\}.$$

We define the oracle scan test as the test $\phi_{s_\bullet,t_\bullet}$, where

$$(s_\bullet, t_\bullet) := \arg\max_{s<t} \mathbb{P}_1(N_n[s,t] \ge c_{s,t}).$$

Indeed, $\phi_{s_\bullet,t_\bullet}$ relies on oracle knowledge of $(\varepsilon, \mu)$.

To the best of our knowledge, this is the first time that such tests are considered in the line of work that concerns us here with roots in the work of Ingster (1997) and Donoho and Jin (2004) — although a similar procedure is used in Cai, Jin and Low (2007) to estimate the contamination proportion $\varepsilon$. The main reason for considering these tests in the present context is that they happen to be first-order optimal, not only in the models considered in the literature (such as generalized Gaussian), but also in power-law models where $F$ has fat tails (e.g., t distribution, Cauchy or Pareto), whereas threshold tests fail are suboptimal for such models. We observe that power-law models are mostly absent from this literature, although they are mentioned in Jin et al. (2005) in the context of an application in cosmology.

### 1.3. Content

For simplicity and the sake of clarity, we will focus on oracle-type, rather than likelihood ratio, performance bounds. The former are indeed more transparent and can be obtained under more generality and with simpler arguments. Also our main intention here is to compare what can be achieved with threshold tests compared to the more general scan tests, defined next, and comparing the corresponding oracle tests seems more appropriate.

In Section 2, we study the oracle threshold test and the oracle scan test. We then consider a number of models. In Section 3, we consider the two scan tests described above and compare them to the oracle scan test. In Section 4, we present the result of some numerical experiments that illustrate our theory. We briefly discuss the performance of the likelihood ratio test and that of nonparametric approaches in Section 5.

## 2. Oracle threshold test and oracle scan test

In this section we state and prove some basic results for the oracle threshold and oracle scan tests.

### *2.1. Power monotonicity*

It is natural to guess that the testing (1) versus (2) becomes easier as the shift $\mu$ increases. This is indeed the case, at least from the point of view of both oracle tests.

**Proposition 1.** *The oracle threshold test has monotonic power in the shift.*

*Proof.* We assume that $\varepsilon > 0$ is fixed and let $\mathbb{P}_\mu$ denote the data distribution under the alternative (2). Take $\mu_1 \le \mu_2$ and let $t_k$ denote the oracle threshold (5) for $\mu_k$, so that the oracle test for $\mu_k$, meaning $\phi_{t_k}$, has rejection region $\{N_n(t_k) \ge c_{t_k}\}$ and power $\pi_k := \mathbb{P}_{\mu_k}(N_n(t_k) \ge c_{t_k})$. Thus we need to show that $\pi_1 \le \pi_2$. This is so because of the fact that, for any $t$, $N_n(t)$ is stochastically non-decreasing in $\mu$, leading to

$$\pi_1 = \mathbb{P}_{\mu_1}(N_n(t_1) \ge c_{t_1}) \le \mathbb{P}_{\mu_2}(N_n(t_1) \ge c_{t_1}) \le \mathbb{P}_{\mu_2}(N_n(t_2) \ge c_{t_2}) = \pi_2,$$

where the last inequality is by construction of $t_2$ and $c_2$. □

Clearly, the oracle scan test has at least as much power as the oracle threshold test. Interestingly, it does not have monotonic power in general, although it does under some natural assumptions on the base distribution.

**Proposition 2.** *Assume that $F$, as a distribution, is unimodal. Then the oracle scan test has monotonic power in the shift.*

*Proof.* We stay with the setting and notation introduced in the proof of Proposition 1. Let $d \ge 0$ be smallest such that

$$F[s_1 + d, t_1 + \mu_2 - \mu_1] = F[s_1, t_1].$$

The fact that $F$, as a distribution, is unimodal implies that $d \le \mu_2 - \mu_1$. Now, under the null, by construction,

$$\mathbb{P}_0(N_n[s_1 + d, t_1 + \mu_2 - \mu_1] \ge c_{s_1,t_1}) = \mathbb{P}_0(N_n[s_1, t_1] \ge c_{s_1,t_1}) \le \alpha.$$

On the other hand, under $\mathbb{P}_{\mu_1}$, $N_n[s_1, t_1]$ is binomial with parameters $n$ and $q_1 := (1-\varepsilon)F[s_1, t_1] + \varepsilon F[s_1 - \mu_1, t_1 - \mu_1]$, while under $\mathbb{P}_{\mu_2}$, $N_n[s_1 + d, t_1 + \mu_2 - \mu_1]$ is binomial with parameters $n$ and

$$\begin{aligned} q_2 &:= (1-\varepsilon)F[s_1 + d, t_1 + \mu_2 - \mu_1] + \varepsilon F[s_1 + d - \mu_2, t_1 + \mu_2 - \mu_1 - \mu_2] \\ &= (1-\varepsilon)F[s_1, t_1] + \varepsilon F[s_1 + d - \mu_2, t_1 - \mu_1] \\ &\ge q_1, \end{aligned}$$

using the fact that $d \leq \mu_2 - \mu_1$. This explains the first inequality in the following derivation

$$\begin{aligned}\pi_1 &= \mathbb{P}_{\mu_1}(N_n[s_1, t_1] \geq c_{s_1,t_1}) \\ &\leq \mathbb{P}_{\mu_2}(N_n[s_1 + d, t_1 + \mu_2 - \mu_1] \geq c_{s_1,t_1}) \\ &\leq \mathbb{P}_{\mu_2}(N_n[s_2, t_2] \geq c_{s_2,t_2}) = \pi_2,\end{aligned}$$

and the second inequality is by definition of $(s_2, t_2)$. □

### *2.2. Performance bounds*

We now provide necessary and sufficient conditions for the the oracle threshold test and the oracle scan test to be fully powerful in the large-sample limit ($n \to \infty$). We focus on the case where

$$n\varepsilon_n \to \infty, \qquad \sqrt{n}\varepsilon_n \to 0,$$

where the first condition implies that, under the alternative, the sample is indeed contaminated with probability tending to 1, while the second condition puts us in the regime corresponding to $\beta > 1/2$ under Ingster's parameterization (3).

Our analysis below is based on the following simple result, which is an immediate consequence of Chebyshev's inequality and the central limit theorem.

**Lemma 1.** *Suppose that we are testing $N \sim Bin(n, p_n)$ versus $N \sim Bin(n, q_n)$ where $p_n \leq 1/2$ and $p_n \leq q_n$, and consider the test at level $\alpha$ that rejects for large values of $N$ — which is the most powerful test. It is asymptotically powerful if $n(q_n - p_n)^2/q_n \to \infty$, while it is asymptotically powerless if $n(q_n - p_n)^2/p_n \to 0$.*

Using Lemma 1, we easily obtain performance guarantees for the oracle threshold test and the oracle scan test.

**Proposition 3.** *The oracle threshold test is powerful if there is a sequence of thresholds $(t_n)$ such that*

$$\begin{aligned} n\varepsilon_n \bar{F}(t_n - \mu_n) \to \infty, \quad \textit{and} \\ n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) \to \infty. \end{aligned} \tag{8}$$

*It is powerless if for any sequence of thresholds $(t_n)$*

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) \to 0. \tag{9}$$

*Proof.* Let $(t_n)$ denote a sequence of thresholds satisfying (8), and define $p_n = \bar{F}(t_n)$ and $q_n = (1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)$. We know that $N_n(t_n) \sim \mathrm{Bin}(n, p_n)$ under the null and $N_n(t_n) \sim \mathrm{Bin}(n, q_n)$ under the alternative, with

$$n(q_n - p_n)^2/q_n = \frac{n\varepsilon_n^2(\bar{F}(t_n - \mu_n) - \bar{F}(t_n))^2}{(1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)}.$$

If the second part of (8) holds, then necessarily $\bar{F}(t_n - \mu_n) \gg \bar{F}(t_n)$, since

$$\begin{aligned} n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2 / \bar{F}(t_n) &= \big[n\varepsilon_n^2 \bar{F}(t_n)\big]\big[\bar{F}(t_n - \mu_n)/\bar{F}(t_n)\big]^2 \\ &\le (n\varepsilon_n^2)\big[\bar{F}(t_n - \mu_n)/\bar{F}(t_n)\big]^2, \end{aligned}$$

with $n\varepsilon_n^2 = o(1)$ by assumption. Hence,

$$\begin{aligned} n(q_n - p_n)^2/q_n &\sim \frac{n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2}{(1-\varepsilon_n)\bar{F}(t_n) + \varepsilon_n \bar{F}(t_n - \mu_n)} \\ &\asymp n\varepsilon_n \bar{F}(t_n - \mu_n) \bigwedge n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n). \end{aligned}$$

Therefore, by Lemma 1, the sequence of tests $(\phi_{t_n})$ has full power in the limit when (8) holds.

Now let $(t_n)$ be any sequence of thresholds and consider the sequence of tests $(\phi_{t_n})$. By Lemma 1, it has power $\alpha$ in the limit since

$$n(q_n - p_n)^2/p_n \le n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/(1-\varepsilon_n)\bar{F}(t_n) \to 0,$$

where the convergence to 0 comes from (9). □

*Remark* 1. Note that the first part of (8) may be replaced by

$$n\bar{F}(t_n) \to \infty.$$

This is because this and $n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \to \infty$ implies $n\varepsilon_n \bar{F}(t_n - \mu_n) \to \infty$.

**Proposition 4.** *The oracle scan test is powerful if there is a sequence of intervals* $([s_n, t_n])$ *such that*

$$\begin{aligned} &n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] \to \infty, \quad \textit{and} \\ &n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2/F[s_n, t_n] \to \infty. \end{aligned} \tag{10}$$

*It is powerless if for any sequence of intervals* $([s_n, t_n])$

$$n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2/F[s_n, t_n] \to 0.$$

The proof is completely parallel to that of Proposition 3 and is omitted.

### *2.3. Examples: generalized Gaussian models and more*

We look at a number of models and in each case derive the performance of the oracle threshold and oracle scan tests, and compare that with the performance of the likelihood ratio test.

To place the results in line with the literature on the topic, we adopt Ingster's parameterization (3) for $\varepsilon_n$, in fact a softer version of that

$$\varepsilon = \varepsilon_n \sim n^{-\beta}, \tag{11}$$

for some fixed $\beta$. The parameterization of $\mu = \mu_n$ will depend on on the model.

To further simplify matters, we assume throughout that

$$\log \bar{F}(x) \sim -\varphi(x),$$

where $\varphi(x)$ is continuous and strictly increasing for $x$ large enough. In that case, in view of Remark 1, we note that (8) is satisfied when

$$\begin{aligned} \log n - \varphi(t_n) &\to \infty, \\ (1-2\beta)\log n + \varphi(t_n) - 2\varphi(t_n - \mu_n) &\to \infty. \end{aligned} \tag{12}$$

*2.3.1. Extended generalized Gaussian*

This class of models is defined by the property that $\varphi$ satisfies[2]

$$\varphi(ut)/\varphi(t) \to u^a, \quad t \to \infty, \quad \forall u \geq 0. \tag{13}$$

Here $a > 0$ parameterizes this class of models. This covers the generalized Gaussian models, which are often used as benchmarks in this line of work. It also covers the case where $\varphi(t) \sim t^a (\log t)^b$ where $b \in \mathbb{R}$ is arbitrary.

For $a > 1$, define

$$\rho_a(\beta) = \begin{cases} (2^{1/(a-1)} - 1)^{a-1}(\beta - 1/2), & 1/2 < \beta < 1 - 2^{-a/(a-1)}, \\ (1 - (1-\beta)^{1/a})^a, & 1 - 2^{-a/(a-1)} \leq \beta < 1. \end{cases}$$

For $a \leq 1$, define

$$\rho_a(\beta) = 2(\beta - 1/2).$$

In addition to (11), assume that

$$\mu = \mu_n \text{ satisfies } \varphi(\mu_n) \sim r \log n, \text{ with } r \geq 0 \text{ fixed.} \tag{14}$$

**Proposition 5.** *The curve $r = \rho_a(\beta)$ in the $(\beta, r)$ plane is the detection boundary that the oracle threshold test achieves.*

*Proof.* We focus on proving that the oracle threshold test achieves that boundary. A simple inspection of the arguments reveal that they are tight, so that this is the precise detection boundary that the test achieves. (See the proof of Proposition 8 for an example.)

We divide the proof into several cases.

*Case 1:* $a > 1$. Define $b = 2^{-1/(a-1)}$ and note that $0 < b < 1$.

*Case 1.1:* $1/2 < \beta < 1 - b^a$ *and* $r > \rho_a(\beta)$. Under these conditions, $\beta < 1/2 + r(1/b - 1)^{-(a-1)}$, and in particular there is $\eta > 0$ such that

$$1 - 2\beta \geq -2r(1/b - 1)^{-(a-1)} + \eta. \tag{15}$$

[2] It is tempting to consider a more general condition where there is a function $\omega$ on $\mathbb{R}_+$ such that $\lim_{t\to\infty} \varphi(ut)/\varphi(t) \to \omega(u)$ for all $u \geq 0$. However, as long as $\omega$ is not constant (equal to zero in that case), it can easily be shown that $\omega(u) = u^a$ for some $a > 0$.

Setting $t_n = (1-b)^{-1}\mu_n$, by (13) and (14), we have the following

$$\varphi(t_n - \mu_n) = \big(rb^a/(1-b)^a + o(1)\big)\log n,$$
$$\varphi(t_n) = \big(r/(1-b)^a + o(1)\big)\log n.$$

By Proposition 1 we may focus on $r$ small enough that $r/(1-b)^a < 1$. This is possible because $\rho_a(\beta) < (1-b)^a$ when $\beta < 1-b^a$, which we assume here. (This can be easily verified using the definition of $b$.) Assuming that $r$ is as such, the first part of (12) is satisfied. For the second part, with (15), we have

$$\begin{aligned}&(1-2\beta)\log n - 2\varphi(t_n-\mu_n) + \varphi(t_n)\\ &\geq \big[-2r(1/b-1)^{-(a-1)} + \eta - 2rb^a/(1-b)^a + r/(1-b)^a + o(1)\big]\log n\\ &= [\eta + o(1)]\log n \to \infty,\end{aligned}$$

using the definition of $b$ and simplifying. Thus the second part of (12) is also satisfied and the oracle threshold test is powerful.

*Case 1.2:* $1-b^a \leq \beta < 1$ *and* $r > \rho_a(\beta)$. Under these conditions, we have $1-\beta > (1-r^{1/a})^a$, and in particular there is $\eta > 0$ such that

$$1-\beta-\eta \geq (1-r^{1/a})^a \geq ((1-\eta)^{1/a} - r^{1/a})^a.$$

Set $t_n = (\frac{1}{r}(1-\eta))^{1/a}\mu_n$, we have the following

$$\varphi(t_n-\mu_n) = \big((1-\eta)^{1/a} - r^{1/a})^a + o(1)\big)\log n,$$
$$\varphi(t_n) = (1-\eta+o(1))\log n.$$

By looking at the speed of $\varphi(t_n)$, the first part of (12) is satisfied immediately. For the second part, with (15), we have

$$\begin{aligned}&(1-2\beta)\log n - 2\varphi(t_n-\mu_n) + \varphi(t_n)\\ &= (1-2\beta)\log n - 2\big((1-\eta)^{1/a} - r^{1/a})^a + o(1)\big)\log n + (1-\eta+o(1))\log n\\ &= 2\big[1-\beta-\eta/2 - ((1-\eta)^{1/a} - r^{1/a})^a + o(1)\big]\log n \to \infty.\end{aligned}$$

Thus the second part of (12) is also satisfied and the oracle threshold test is powerful.

*Case 2:* $a \leq 1$. By Proposition 1 we may restrict attention to the case where $2\beta - 1 < r < 1$. Here we set $t_n = \mu_n$. Then the first part in (12) is clearly satisfied. For the second part, notice that

$$\begin{aligned}&(1-2\beta)\log n - 2\varphi(t_n-\mu_n) + \varphi(t_n)\\ &= (1-2\beta)\log n + (r+o(1))\log n\\ &= [1-2\beta+r+o(1)]\log n \to \infty.\end{aligned}$$

This completes the proof. □

Thus, although the conditions are much more general here, the detection boundary is the same as in the corresponding generalized Gaussian model and, moreover, the oracle threshold test achieves that boundary.

*Remark* 2 (max test). In this class of models, it can be shown that the max test achieves the detection boundary over the upper range, meaning when $\beta \geq 1-2^{-a/(a-1)}$. In fact, $\rho^{\max}(\beta) := (1-(1-\beta)^{1/a})^a$ defines the detection boundary for the max test.

### *2.3.2. Other models*

In the next few classes of models, $\varphi$ satisfies

$$\frac{\varphi^{-1}(t) - \varphi^{-1}(vt)}{\lambda(t)} \to \omega(v), \quad t \to \infty, \quad \forall v \in (0,1]. \tag{16}$$

for some functions $\lambda$ and $\omega$, with the latter being non-increasing, continuous, and such that $\omega(1) = 0$. This is actually also the case when $\varphi(t) \sim t^a(\log t)^b$ with $a > 0$ and $b \in \mathbb{R}$, with $\lambda(t) = t^{1/a}(\log t)^{-b/a}$ and $\omega(v) = (1 - v^{1/a})/a^{b/a}$.

Define

$$\rho(\beta) = \inf_{0<h<1-\beta} \big[\omega(h) - \omega(2\beta - 1 + 2h)\big]. \tag{17}$$

In addition to (11), assume that

$$\mu = \mu_n \sim r\lambda(\log n), \quad r \geq 0 \text{ fixed}.$$

**Proposition 6.** *The curve $r = \rho(\beta)$ in the $(\beta, r)$ plane is the detection boundary that the oracle threshold test achieves.*

*Proof.* We focus on proving that the oracle threshold test achieves that boundary.

Since $\omega(v)$ is continuous, we may define

$$h^* = \operatorname*{arg\,min}_{0\leq h\leq 1-\beta} \big[\omega(h) - \omega(2\beta - 1 + 2h)\big].$$

We focus on the case where $h^* < 1 - \beta$. In the case where $h^* = 1 - \beta$, the max test is powerful (Remark 3), and therefore so is the oracle threshold test. By Proposition 1 we may focus on the case where $r < \omega(h^*)$. With these assumptions and the fact that $\omega(h^*) - \omega(2\beta - 1 + 2h^*) = \rho(\beta) < r$, there is $\eta > 0$ be such that

$$2\beta - 1 + 2h^* + 2\eta < 1, \tag{18}$$

and

$$\omega(h^*) - \omega(2\beta - 1 + 2h^* + \eta) < r < \omega(h^*) - \omega(2\beta - 1 + 2h^* + 2\eta).$$

Define $t_n := \mu_n + \varphi^{-1}(h^* \log n)$. Using (16) multiple times, for $n$ sufficiently large, we have the following

$$\begin{aligned}
\mu_n &= (r + o(1))\lambda(\log n) \\
&\le [\omega(h^*) - \omega(2\beta - 1 + 2h^* + 2\eta)]\lambda(\log n) \\
&= \varphi^{-1}(\log n) - \varphi^{-1}(h^* \log n) - \varphi^{-1}(\log n) + \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n) \\
&= \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n) - \varphi^{-1}(h^* \log n).
\end{aligned}$$

Hence, eventually, $t_n \le \varphi^{-1}((2\beta - 1 + 2h^* + 2\eta)\log n)$, implying that

$$\begin{aligned}
\log n - \varphi(t_n) &= \log n - (2\beta - 1 + 2h^* + 2\eta)\log n \\
&= [1 - (2\beta - 2 + 2h^* + 2\eta)]\log n \to \infty,
\end{aligned}$$

using (18). Thus the first part of (12) is satisfied.

Similarly, for $n$ sufficiently large,

$$\begin{aligned}
\mu_n &= (r + o(1))\lambda(\log n) \\
&\ge [\omega(h^*) - \omega(2\beta - 1 + 2h^* + \eta)]\lambda(\log n) \\
&= \varphi^{-1}((2\beta - 1 + 2h^* + \eta)\log n) - \varphi^{-1}(h^* \log n),
\end{aligned}$$

so that, eventually, $t_n \ge \varphi^{-1}((2\beta - 1 + 2h^* + \eta)\log n)$, implying that

$$\begin{aligned}
&(1 - 2\beta)\log n - 2\varphi(t_n - \mu_n) + \varphi(t_n) \\
&\ge (1 - 2\beta)\log n - 2h^* \log n + (2\beta - 1 + 2h^* + \eta)\log n \\
&= \eta \log n \to \infty.
\end{aligned}$$

Thus the second part of (12) is satisfied. □

*Remark* 3 (max test). In the present situation, it can be shown that $\rho^{\max}(\beta) := \omega(1 - \beta)$ defines the detection boundary for the max test.

### *2.3.3. Extended generalized Gumbel*

This class of models is defined by $\varphi(t) = \exp(t^a)$ for some $a > 0$, which satisfies (16) with $\lambda(t) = \frac{1}{a}(\log t)^{1/a-1}$ and $\omega(v) = \log(1/v)$. In this case,

$$\mu = \mu_n \sim \frac{r}{a}(\log\log n)^{1/a-1},$$

and the detection boundary is given by $r = -\log(1 - \beta)$. Note that, at the detection boundary, $\mu_n \to \infty$ when $a > 1$; that $\mu_n \asymp 1$ when $a = 1$; and $\mu_n \to 0$ when $a < 1$.

*2.3.4. Extended generalized Gumbel*

This class of models is defined by $\varphi(t) = \exp((\log t)^a)$ for some $a > 1$, which satisfies (16) with $\lambda(t) = \frac{1}{a}(\log t)^{1/a-1}\exp((\log t)^{1/a})$ and $\omega(v) = \log(1/v)$. In this case,

$$\mu = \mu_n \sim \frac{r}{a}(\log\log n)^{1/a-1}\exp((\log\log n)^{1/a}),$$

and the detection boundary is given by $r = -\log(1-\beta)$ as in the previous class of models (since $\omega$ is the same).

*Remark* 4 (max test). Based on Remark 3, in the last two classes of models, the max test achieves the detection boundary over the whole $\beta$ range. The same is true, more generally, when the infimum in (17) is at $h = 1-\beta$.

### *2.4. Examples: power-law models and more*

In the next few classes of models, $F$ satisfies

$$\log(F(t+v) - F(t)) \sim -\lambda(t), \quad t \to \infty, \quad \forall v \geq 0, \tag{19}$$

for some function $\lambda$ which is increasing eventually and such that $\lambda(t) \to \infty$ as $t \to \infty$. This includes models where

$$\bar{F}(t) \propto t^{-a}(\log t)^b(1 + o(1/t)), \quad t \to \infty, \tag{20}$$

with $a > 0$ and $b \in \mathbb{R}$, in which case (19) holds with $\lambda(t) = (a+1)\log t$. It also includes models where $\bar{F}(t) \propto (\log t)^{-a}(1 + o(1/t\log t))$, with $a > 0$, in which case (19) holds with $\lambda(t) = \log t$, as well as other distribution with even slower decay.

In addition to (11), assume that

$$\mu = \mu_n \quad \text{satisfies} \quad \lambda(\mu_n) \sim r\log n, \quad r \geq 0 \text{ fixed}. \tag{21}$$

**Proposition 7.** *The curve $r = \rho(\beta) := 2\beta - 1$ in the $(\beta, r)$ plane is the detection boundary that the oracle scan test achieves.*

*Proof.* We focus on proving that the oracle scan test achieves that boundary.

Fix $r$ such that $r > 2\beta - 1$. Consider the interval $[s_n, t_n]$ with $s_n := \mu_n$ and $t_n := \mu_n + v$, where $v > 0$ is such that $F[0, v] > 0$. We need to verify that (10) holds. On the one hand, we have

$$n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] = n^{1-\beta}F[0, v] \to \infty,$$

because $\beta < 1$ by assumption. So the first part of (10) holds. On the other hand,

$$\begin{aligned} n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2/F[s_n, t_n] &= n^{1-2\beta}F[0, v]^2/n^{r+o(1)} \\ &= n^{r+1-2\beta+o(1)} \to \infty, \end{aligned}$$

since $r > 2\beta - 1$. So the second part of (10) holds. □

We now show that threshold tests are suboptimal in the main class of models satisfying (19), namely (20). (The same happens to be true in other models with fat tails satisfying (19).) This is the main motivation for considering scan tests.

**Proposition 8.** *In a model satisfying* (20)*, and with the same parameterization* (21)*, the curve* $r = (1+1/a)(2\beta-1)$ *in the* $(\beta, r)$ *plane is the detection boundary that the oracle threshold test achieves.*

*Proof.* We first prove that the oracle threshold test achieves this detection boundary. By Proposition 1 we may assume that $r < 1 + 1/a$. Therefore, fix $r$ such that $(1 + 1/a)(2\beta - 1) < r < 1 + 1/a$. Set the threshold $t_n = \mu_n + v$, where $v$ is such that $\bar{F}(v) > 0$. We need to verify that (8) holds, and we do so via Remark 1. Note that $t_n \sim \mu_n = n^{r/(a+1)+o(1)}$. In particular,

$$n\bar{F}(t_n) \sim n\mu_n^{-a}(\log \mu_n)^b = n^{1-ar/(a+1)+o(1)} \to \infty,$$

and, by the same token,

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \sim n^{1-2\beta} n^{-ar/(a+1)+o(1)} = n^{1-2\beta-ar/(a+1)+o(1)} \to \infty.$$

We now turn to proving that this is the statement boundary is the best that the oracle threshold test can hope for. For this, fix $r < (1+1/a)(2\beta-1)$. We need to verify (9). Suppose for contradiction that there is a sequence of thresholds, $(t_n)$, such that (9) does not hold. By extracting a subsequence if needed, we may assume that

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \to \lambda \in (0, \infty]. \tag{22}$$

First, suppose that $\liminf t_n/\mu_n < \infty$. Extracting a subsequence if needed, we may assume that $t_n = O(\mu_n)$. In that case, we have

$$\begin{aligned} n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) &\le n\varepsilon_n^2/\bar{F}(t_n) \\ &\le n^{1-2\beta+o(1)}\mu_n^{-a+o(1)} \\ &= n^{1-2\beta-ar/(a+1)+o(1)} \to 0. \end{aligned}$$

Since this contradicts (22), we must have $\liminf t_n/\mu_n = \infty$, meaning that $t_n \gg \mu_n$. In that case, we have $\bar{F}(t_n - \mu_n) \sim \bar{F}(t_n)$, implying that

$$n\varepsilon_n^2 \bar{F}(t_n - \mu_n)^2/\bar{F}(t_n) \sim n\varepsilon_n^2 \bar{F}(t_n) \le n\varepsilon_n^2 \to 0.$$

This also contradicts (22). Since there is no other option, it must be that (22) cannot hold. We conclude that, indeed, (9) holds for any sequence of thresholds. □

## 3. Scan tests

In this section, we study the scan tests (6) and (7), and show that both of them do as well as the oracle scan test, at least to first-order in the asymptote where $n \to \infty$ and under the various parameterizations used in the previous section. We refer to (6) as the Stouffer scan test, as it is constructed as Stouffer's combination test (Stouffer et al., 1949); while we refer to (7) as the Tippett scan test, for similar reasons (Tippett, 1931).

### 3.1. Stouffer scan test

We study the Stouffer scan test (6). The main work goes into controlling this statistic under the null hypothesis. The limiting distribution of higher criticism can be derived from (Jaeschke, 1979) and the limiting distributions of some variants of the scan statistic are known under other models (Kabluchko, 2011; Sharpnack and Arias-Castro, 2016). We will not pursue such a fine result here, but contend ourselves with a relatively rough upper bound.

**Lemma 2.** *Given observations $x_1, \dots, x_n$, the maximum in (6) is attained at some $(s,t) = (x_i, x_j)$.*

*Proof.* Define

$$R_n(s,t) := \frac{N_n[s,t] - nF[s,t]}{\sqrt{nF[s,t](1-F[s,t])+1}}. \tag{23}$$

Let $x_{(1)} \le \cdots \le x_{(n)}$ denote the ordered observations, and set $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. It suffices to show that, for any $1 \le i \le j \le n$ and any $(s,t)$ such that $x_{(i-1)} < s \le x_{(i)}$ and $x_{(j)} \le t < x_{(j+1)}$, in addition to $F[s,t] \le 1/2$, we have $R_n(x_{(i)}, x_{(j)}) \ge R_n(s,t)$. The crucial observation is that $N_n[s,t] = N_n[x_{(i)}, x_{(j)}]$ while $F[x_{(i)}, x_{(j)}] \le F[s,t]$.

It is thus enough to show that the function $p \mapsto (a-p)/(p(1-p)+b)^{1/2}$ is decreasing over $[0, 1/2]$ for any $a, b \ge 0$. This is so since this function has derivative $-(a(1-2p)+2b+p)/(p(1-p)+b)^{3/2}$. □

**Theorem 1.** *With $S_n$ defined as the statistic (6), we have*

$$\mathbb{P}_0(S_n \ge 3\log n) \to 0.$$

*Proof.* We place ourselves under the null hypothesis. Recall the definition of $R_n$ in (23). By Lemma 2 and the fact that $R_n(X_i, X_i) = 1$ for all $i$, if $S_n \ge 3\log n$ necessarily $S_n = S_n^* := \max_{i \ne j} R_n(X_i, X_j)$. For any $i \ne j$, we have

$$R_n(X_i, X_j) \le 2 + S_{i,j},$$

with

$$\begin{aligned} S_{i,j} &:= \frac{N_{i,j} - 2 - (n-2)p_{i,j}}{\sqrt{(n-2)p_{i,j}(1-p_{i,j})+1}}, \\ N_{i,j} &:= N_n[X_i, X_j], \quad p_{i,j} := F[X_i, X_j]. \end{aligned} \tag{24}$$

The point of this reorganizing is that, given $(X_i, X_j)$, $N_{i,j} - 2 \sim \mathrm{Bin}(n-2, p_{i,j})$, and an application of Bernstein's inequality gives

$$\begin{aligned} \mathbb{P}_0(S_{i,j} \ge s \mid X_i, X_j) &\le \exp\left( - \frac{s^2 b_{i,j}^2/2}{b_{i,j}^2 + s b_{i,j}/3} \right) \\ &\le \exp\left( - \frac{s^2/2}{1 + s/3} \right) \\ &\le \exp(-s), \quad \forall s \ge 6, \end{aligned}$$

because $b_{i,j} := \sqrt{(n-2)p_{i,j}(1-p_{i,j})+1} \geq 1$. Thus, with the union bound, as $n \to \infty$, we have

$$\begin{aligned}\mathbb{P}_0(S_n \geq 3\log n) &= \mathbb{P}_0(\exists i \neq j : S_{i,j} + 2 \geq 3\log n) \\ &\leq \sum_{i<j} \mathbb{P}_0(S_{i,j} \geq 3\log n - 2) \leq n^2 \exp(-3\log n + 2) \to 0,\end{aligned}$$

which proves the statement. □

With Theorem 1, one obtains the following performance bound for the Stouffer scan test.

**Corollary 1.** *The Stouffer scan test is powerful if there is a sequence of intervals* $([s_n, t_n])$ *such that*

$$\begin{aligned} n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] &\gg \log n, \quad \textit{and} \\ n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2 / F[s_n, t_n] &\gg (\log n)^2. \end{aligned} \tag{25}$$

*Proof.* By Theorem 1, the Stouffer scan test at level $\alpha$ is at least as powerful as the test $\{S_n \geq 3\log n\}$, eventually. Now, under the alternative, this test is powerful if we can prove that $p_n := F[s_n, t_n] \leq 1/2$ and $R_n(s_n, t_n) \geq 3\log n$. Define $p'_n := F[s_n - \mu_n, t_n - \mu_n]$ and $q_n := (1-\varepsilon_n)p_n + \varepsilon_n p'_n$, so that (25) can be expressed as

$$n\varepsilon_n p'_n \gg \log n \to \infty, \quad \text{and} \quad n\varepsilon_n^2 {p'_n}^2 / p_n \gg (\log n)^2 \to \infty.$$

That $p_n \leq 1/2$ is true, eventually, comes from the fact that

$$\infty \leftarrow n\varepsilon_n^2 {p'_n}^2 / p_n \leq n\varepsilon_n^2 / p_n,$$

with $n\varepsilon_n^2 \to 0$ by assumption, so that necessarily $p_n \to 0$. Note that this implies that $q_n \to 0$ also.

Given that $N_n[s_n, t_n]$ is a binomial distribution with parameters $n$ and $q_n$, with $nq_n \geq np'_n \to \infty$ by the first part of (25), we have $N_n[s_n, t_n] = nq_n + O_P(\sqrt{nq_n(1-q_n)})$, and so

$$R_n(s_n, t_n) = \frac{n\varepsilon_n(p'_n - p_n) + O_P(\sqrt{nq_n(1-q_n)})}{\sqrt{np_n(1-p_n)+1}} \sim \frac{n\varepsilon_n p'_n + O_P(\sqrt{nq_n})}{\sqrt{np_n+1}},$$

since $p'_n \gg p_n$, by the fact that

$$\infty \leftarrow n\varepsilon_n^2 {p'_n}^2 / p_n = n\varepsilon_n^2 p_n (p'_n/p_n)^2 = o(p'_n/p_n)^2.$$

In addition, the same conditions imply

$$\frac{n\varepsilon_n p'_n}{\sqrt{nq_n}} \asymp \sqrt{n\varepsilon_n^2 {p'_n}^2 / p_n} \bigvee n\varepsilon_n p'_n \to \infty,$$

so that

$$R_n(s_n, t_n) \sim_P n\varepsilon_n p'_n / \sqrt{np_n+1} \asymp_P \sqrt{n\varepsilon_n^2 {p'_n}^2 / p_n} \bigvee n\varepsilon_n p'_n \gg \log n.$$

We conclude that $R_n(s_n, t_n) \geq 3\log n$ holds with probability tending to 1. □

With this performance bound, it is straightforward to verify that the Stouffer scan test performs as well as the oracle scan test to first order, at least in the context of the parameterization used in the models studied in Section 2.3 and Section 2.4. This comes from the fact that, in context of these sections, the quantity appearing in (25) increases as a (fixed) positive power of $n$ under the alternative. We formalize this into the following statement, left without formal proof.

**Corollary 2.** *The Stouffer scan test achieves the oracle scan detection boundary in all the settings considered in Section 2.3 and Section 2.4.*

### *3.2. Tippett scan test*

We study the Tippett scan test (7), which we denote by $T_n$. We control this statistic under the null hypothesis by a simple application of the union bound. A more refined control seems possible in view of Moscovich, Nadler and Spiegelman (2016), where the limiting distribution of (4) is obtained.

**Proposition 9.** *With $T_n$ defined as the statistic (7), we have*

$$\mathbb{P}_0(T_n \le 1/n^3) \to 0.$$

*Proof.* Under the null, each $P_{i,j}$ is uniformly distributed in $[0,1]$. Thus the union bound gives

$$\mathbb{P}_0(T_n \le 1/n^3) \le n^2\, \mathbb{P}_0(P_{i,j} \le 1/n^3) = n^2/n^3 = 1/n \to 0,$$

which concludes the proof. □

Thus most of the work goes into controlling the statistic under the alternative. We do so by bounding the Tippett scan statistic by an expression that resembles that of the Stouffer scan statistic. We make use of the following simple concentration bound.[3]

**Lemma 3.** *For $k \in [n]$,*

$$\mathsf{B}(u; k, n-k+1) \le \exp\left( - \frac{(k-nu)^2/2}{nu(1-u) + (k-nu)/3} \right), \quad 0 \le u \le k/n.$$

*Proof.* Let $U_{k:n}$ denote the $k$-th order statistic of an iid sample of size $n$ from the uniform distribution on $[0,1]$. For $u \in [0,1]$ such that $nu \le k$, we have

$$\mathsf{B}(u; k, n-k+1) = \mathbb{P}(U_{k:n} \le u) = \mathbb{P}(\mathrm{Bin}(n,u) \ge k),$$

and we conclude with an application of Bernstein's inequality. □

[3] Many things are known about the beta distribution and order statistics in general, but we could not immediately find such a simple bound.

**Proposition 10.** *The Tippett scan test is powerful if there is a sequence of intervals* $([s_n, t_n])$ *such that*

$$\begin{aligned} n\varepsilon_n F[s_n - \mu_n, t_n - \mu_n] &\gg \sqrt{\log n}, \\ n\varepsilon_n^2 F[s_n - \mu_n, t_n - \mu_n]^2 / F[s_n, t_n] &\gg \log n. \end{aligned} \tag{26}$$

*Proof.* Recall that $T_n = \min_{i<j} P_{i,j}$ and the expression of $P_{i,j}$. Thus an application of Lemma 3 gives

$$T_n \le 1/n^3 \;\Leftrightarrow\; \max_{i<j} \frac{(j-i-V_{i,j})_+^2/2}{nV_{i,j}(1-V_{i,j}) + (j-i-V_{i,j})_+/3} \ge 3\log n,$$

where $V_{i,j} := U_{(j)} - U_{(i)}$, after taking a logarithm.

Moreover, $V_{i,j} = F[X_{(n-j+1)}, X_{(n-i+1)}]$ and $j - i = N_n[X_{(n-j+1)}, X_{(n-i+1)}] - 1$, yielding

$$T_n \le 1/n^3 \;\Leftrightarrow\; \max_{i \ne j} \frac{(N_{i,j} - 1 - np_{i,j})_+^2}{np_{i,j}(1-p_{i,j}) + (N_{i,j} - 1 - np_{i,j})_+} \ge 6\log n,$$

with the notation of (24). The latter inequality holds when there is $i \ne j$ such that

$$N_{i,j} - 1 - np_{i,j} \ge 12\log n \quad \text{and} \quad \frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1-p_{i,j})}} \ge \sqrt{12\log n},$$

which is the case when

$$np_{i,j} \ge \sqrt{12\log n} \quad \text{and} \quad \frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1-p_{i,j})}} \ge \sqrt{12\log n}. \tag{27}$$

Let $(s_n, t_n)$ be as in the statement and let $(i, j)$ be such that $U_{(i)} \le s < U_{(i+1)}$ and $U_{(j-1)} < t \le U_{(j)}$. By construction, $p_{i,j} \ge F[s_n, t_n]$, so that the first part of (26) implies that the first part of (27) holds eventually. We also have $N_{i,j} \ge N_n[s_n, t_n] - 2$, so that

$$\frac{N_{i,j} - 1 - np_{i,j}}{\sqrt{np_{i,j}(1-p_{i,j})}} \ge \frac{N_n[s_n, t_n] - 3 - nF[s_n, t_n]}{\sqrt{nF[s_n, t_n](1 - F[s_n, t_n])}},$$

and the quantity on the RHS is controlled using the second part (26) exactly as in the proof of Proposition 4. □

Here too, these results make it straightforward to verify that the Tippett scan test performs as well as the oracle scan test (to first order) in the models and regimes seen earlier, leading us to state the following (left without a formal proof).

**Corollary 3.** *The Tippett scan test achieves the oracle scan detection boundary in all the settings considered in Section 2.3 and Section 2.4.*

## 4. Numerical experiments

We performed small-scale numerical experiments to probe our theory. We generated Student t-distributions with varying numbers of degrees of freedom, $\mathrm{df} = 0.5, 1, 2, 5\}$. Recall that the Student t-distribution with $k$ degrees of freedom has density $\propto (1 + x/k)^{-(k+1)/2}$. We considered three different scenarios with varying sparsity exponents, $\beta = 0.6, 0.7, 0.8$. The sample size was set to $n = 30,000$. We compared the higher criticism test, the Berk-Jones test, the Stouffer scan test, and the Tippett scan test in each of these settings. We repeat each setting 200 times. See Figure 1, Figure 2, and Figure 3.

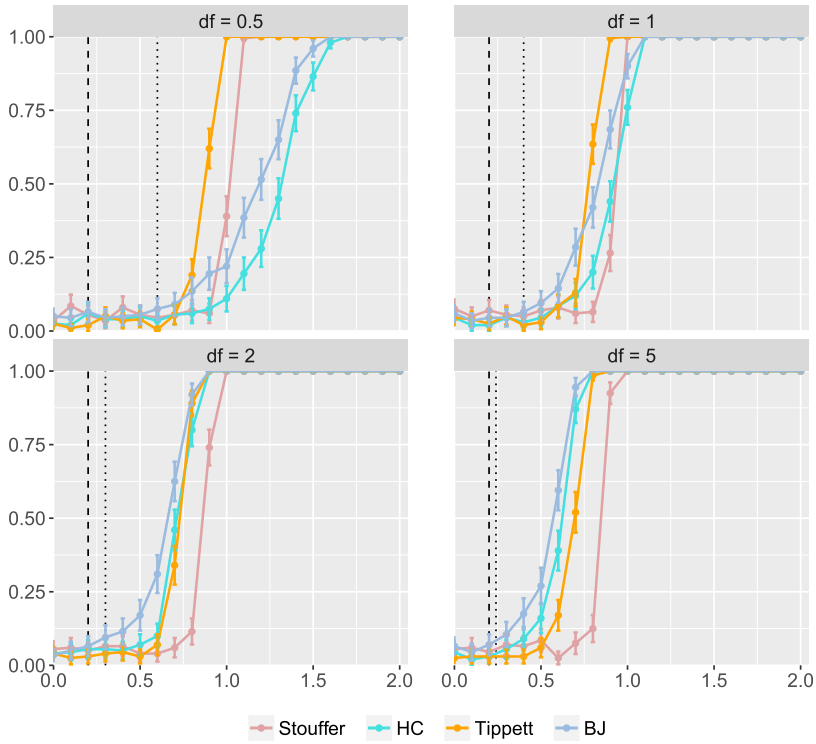


FIG 1. *Here $\beta = 0.6$, the x-axis represents $r$ in the parameterization (21), y-axis the power of the tests identified in the legend. Each subfigure corresponds to a Student t-distribution with the specified number of degrees of freedom. The black dashed vertical line corresponds to the oracle scan detection boundary established in Proposition 7, while the dotted line corresponds to the oracle threshold detection boundary established in Proposition 8.*

As the theory predicts, We can check that when the number of degrees of freedom is smaller, implying that the base distribution has fatter tails, the scan procedures dominate the threshold procedure. The threshold procedures become dominant as the tails become lighter. This is so at this particular sample size as, in principle, our theory indicates that with a larger sample size, the scan

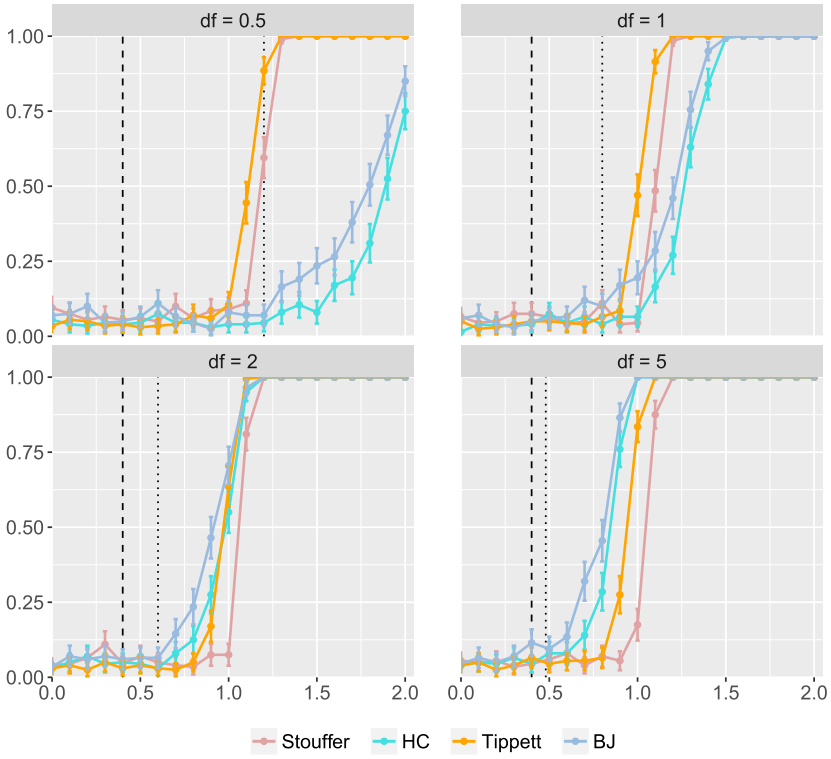


FIG 2. *Here $\beta = 0.7$, otherwise, see Figure 1 for more details.*

procedures would still dominate. The transition from powerless to powerful takes place at a larger effect size than predicted by the theory, which is also explain by the limited sample size.[4]

## 5. Discussion

While scan tests are commonly used in a number of detection problems, threshold tests are almost exclusively used in multiple testing situations. The main purpose of our work here was to reveal that scan tests can improve on threshold tests in somewhat standard multiple testing settings, particularly when the null distribution ($F$ in the paper) has heavy tails.

**Likelihood ratio performance bounds** Given our main objective, it was more natural to consider oracle-type performance bounds rather than using the likelihood ratio performance as benchmark. We can say nonetheless that, for

[4]The scan tests have computational complexity of order $O(n^2)$, which has limited the scale of our experiments.

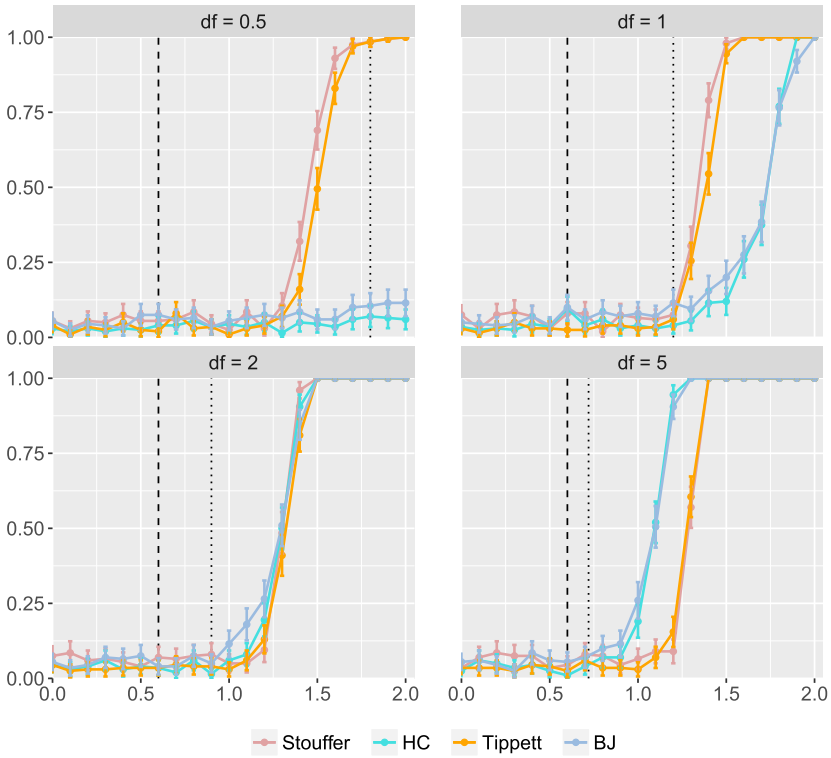


FIG 3. *Here $\beta = 0.8$, otherwise, see Figure 1 for more details.*

representative models, the oracle threshold boundaries stated in Proposition 5 and Proposition 6 match those of the likelihood ratio test — for example, this is true of generalized Gaussian models where $F$ has density of the form $f(t) \propto \exp(-|t|^a)$ for some $a > 0$. The same is true of the oracle scan boundary stated in Proposition 7 — for example, this is true of power law models where $F$ has density of the form $f(t) \propto (1 + |t|^a)^{-1}$ for some $a > 0$.

**Nonparametric approaches** Arias-Castro and Wang (2017) consider the situation where the null distribution, $F$, is symmetric about 0 but otherwise unknown. They suggest two tests for symmetry: the CUSUM sign test and the tail-run test, which are meant to be the nonparametric equivalent of the higher criticism test and the tail-run sign test, respectively. Back-of-the-envelope calculations seem to indicate that these nonparametric tests achieve the same detection boundaries as their parametric counterparts in all the settings considered here.

**Multiple testing** In separate work Chen, Ying and Arias-Castro (2018), we uncover a similar phenomenon in the context of multiple testing, where the

goal is maximizing the number of rejections while controlling the false discovery rate (FDR). Indeed, in a similar mixture model, standard in that literature at least since the work of Genovese and Wasserman (2002, 2004), we find that with heavy tail distributions, scanning can improve on thresholding (what the procedure of Benjamini and Hochberg (1995) does). This is established in the context of the asymptotic framework of Genovese and Wasserman (2002, 2004), which is different than the one considered here in that the mixture proportion, $\varepsilon$, does not converge to zero with the sample size. However, we expect this to extend to the present asymptotic model.[5]

## References


Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* 193–212. MR0050238

Arias-Castro, E. and Chen, S. (2017). Distribution-free multiple testing. *Electronic Journal of Statistics* **11** 1983–2001. MR3651021

Arias-Castro, E., Donoho, D. L. and Huo, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Transactions on Information Theory* **51** 2402–2425. MR2246369

Arias-Castro, E. and Wang, M. (2017). Distribution-free tests for sparse heterogeneous mixtures. *TEST* **26** 71–94. MR3613606

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300. MR1325392

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Probability Theory and Related Fields* **47** 47–59. MR0521531

Cai, T. T., Jeng, X. J. and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 629–662. MR2867452

Cai, T. T., Jin, J. and Low, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics* **35** 2421–2449. MR2382653

Cai, T. T. and Wu, Y. (2014). Optimal Detection of Sparse Mixtures Against a Given Null Distribution. *IEEE Transactions on Information Theory* **60** 2217–2232. MR3181520

Chen, S. and Arias-Castro, E. (2017). Sequential Multiple Testing. *arXiv preprint arXiv:1705.10190*.

Chen, S., Ying, A. and Arias-Castro, E. (2018). A Scan Procedure for Multiple Testing. *arXiv preprint arXiv:1808.00631*.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32** 962–994. MR2065195


[5] The present asymptotic model has been considered in the context of multiple testing, in particular in some of our own recent work Arias-Castro and Chen (2017); Chen and Arias-Castro (2017).

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 499–517. MR1924303

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The Annals of Statistics* 1035–1061. MR2065197

Gibbons, J. D. and Chakraborti, S. (2011). *Nonparametric Statistical Inference*. Springer. MR2681063

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons. MR2488795

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

Jaeschke, D. (1979). The Asymptotic Distribution of the Supremum of the Standardized Empirical Distribution Function on Subintervals. *The Annals of Statistics* **7** 108–115. MR0515687

Jin, J., Starck, J.-L., Donoho, D. L., Aghanim, N. and Forni, O. (2005). Cosmological non-Gaussian signature detection: Comparing performance of different statistical tests. *EURASIP Journal on Advances in Signal Processing* **2005** 297184. MR2210857

Kabluchko, Z. (2011). Extremes of the standardized Gaussian noise. *Stochastic Processes and their Applications* **121** 515–533. MR2763094

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods* **26** 1481–1496. MR1456844

Moscovich, A., Nadler, B. and Spiegelman, C. (2016). On the exact Berk-Jones statistics and their $p$-value calculation. *Electronic Journal of Statistics* **10** 2329–2354. MR3544289

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association* **60** 532–538. MR0183041

Sharpnack, J. and Arias-Castro, E. (2016). Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics* **10** 2641–2684. MR3546971

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). *The American Soldier, Vol 1: Adjustment During Army Life*. Princeton University Press.

Tippett, L. H. C. (1931). *Methods of Statistics*. Williams Norgate: London. MR0050222