Electronic Journal of Statistics
Vol. 10 (2016) 2329–2354
ISSN: 1935-7524
DOI: 10.1214/16-EJS1172

On the exact Berk-Jones statistics and their p-value calculation

Amit Moscovich, Boaz Nadler

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
e-mail: amit.moscovich@weizmann.ac.il; boaz.nadler@weizmann.ac.il

and

Clifford Spiegelman

Department of Statistics, Texas A&M University, College Station TX, USA
e-mail: cliff@stat.tamu.edu

Abstract: Continuous goodness-of-fit testing is a classical problem in statistics. Despite having low power for detecting deviations at the tail of a distribution, the most popular test is based on the Kolmogorov-Smirnov statistic. While similar variance-weighted statistics such as Anderson-Darling and the Higher Criticism statistic give more weight to tail deviations, as shown in various works, they still mishandle the extreme tails.

As a viable alternative, in this paper we study some of the statistical properties of the exact M_n statistics of Berk and Jones. In particular we show that they are consistent and asymptotically optimal for detecting a wide range of rare-weak mixture models. Additionally, we present a new computationally efficient method to calculate p-values for any supremum-based one-sided statistic, including the one-sided M_n^+, M_n^- and R_n^+, R_n^- statistics of Berk and Jones and the Higher Criticism statistic. Finally, we show that M_n compares favorably to related statistics in several finite-sample simulations.

MSC 2010 subject classifications: Primary 62G10, 62G20; secondary 62-04.
Keywords and phrases: Continuous goodness-of-fit, Hypothesis testing, p-value computation, Rare-weak model.

Received February 2016.

Contents

1. Introduction

Let $x_1, x_2, \ldots, x_n$ be a sample of n i.i.d. observations of a real-valued one-dimensional random variable X. The classical continuous goodness-of-fit (GOF) problem is to assess the validity of a null hypothesis that X follows a known (and fully specified) continuous distribution function F, against an unknown and arbitrary alternative G,

$$\mathcal{H}_0 : X \sim F \quad vs. \quad \mathcal{H}_1 : X \sim G \ \text{ with } \ G \neq F. \tag{1.1}$$

Goodness-of-fit is one of the most fundamental hypothesis testing problems (Lehmann and Romano, 2005). Most GOF tests for continuous distributions can be broadly categorized into two groups. The first comprises of tests based on some distance metric between the null distribution F and the empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_i \mathbf{1}(x_i \leq x)$. These include, among others, the tests of Kolmogorov-Smirnov (KS), Cramér-von Mises, Anderson-Darling (AD), Berk-Jones (BJ), as well as the Higher Criticism (HC) and Phi-divergence tests (Anderson and Darling, 1954; Berk and Jones, 1979; Jager and Wellner, 2007). The second group considers the first few moments of the random variable X with respect to an orthonormal basis of $L_2(\mathbb{R})$. Notable representatives are Neyman's smooth test (Neyman, 1937), and its more recent data-driven versions, where the number of moments is determined in an adaptive manner, see Ledwina (1994) and Rainer, Thas and Best (2009).

Despite the abundance of GOF tests, KS is nonetheless the most commonly used in practice. It has several desirable properties, including consistency, good power for detecting a shift in the median of the distribution (Janssen, 2000) and the availability of simple procedures to compute its p-value. However, it suffers from a well known limitation – it has little power for detecting deviations at the tails of the distribution, which is important in a variety of practical situations.

One scenario is the detection of rare contaminations, whereby only a few of the n observations are contaminated and arise from a different distribution. A specific example is the *rare-weak model* (Ingster, 1997; Donoho and Jin, 2004) and its generalization to sparse mixture models (Cai and Wu, 2014). Another example involves high dimensional variable selection or multiple hypothesis testing problems under sparsity assumptions (Walther, 2013).

Given the popularity of the KS test, a natural question is how can it be modified to have tail sensitivity, and what are the properties of the resulting test. In this paper we make several contributions regarding these questions. We start in Section 2 by viewing the KS and the variance-weighted AD and HC statistics under a common framework, as different ways to measure the deviations of order statistics from their expectations. As described in Section 3, this leads us to study a different GOF statistic, based on the following principle: Rather than looking for the largest (possibly weighted) deviation, it looks for the deviation which is *most statistically significant*. Independently of our work, equivalent GOF tests were recently suggested by several different authors, including Mary and Ferrari (2014); Gontscharuk, Landwehr and Finner (2016); Kaplan and Goldman (2014). This statistic is also closely related to the work of Aldor-Noiman et al. (2014) who instead of a GOF test, derived a method to construct confidence bands for a Normal Q-Q plot. It turns out, however, that all of these proposals are in fact *equivalent* to GOF testing based on the M_n statistic defined in a paper by Berk and Jones (1979). The M_n statistic was derived based on an earlier work by the authors on relatively optimal combinations of test statistics (Berk and Jones, 1978). The R_n statistic (often called *the* Berk-Jones statistic) was then proposed as an approximation to M_n, which is simpler to compute. However, with today's computers, this approximation is no longer necessary and the M_n statistic can be computed directly.

On the theoretical front, in Section 4 we analyze some statistical properties of the M_n^+, M_n^- and M_n. Theorem 4.1 presents the asymptotic distribution of these statistics under the null hypothesis. This theorem was recently proven by Gontscharuk and Finner (2016), based on an analysis of the HC statistic (Gontscharuk, Landwehr and Finner, 2016). In a supplementary note, we present an alternative proof, based on classical results from the theory of standardized empirical processes (Eicker, 1979; Jaeschke, 1979). Next, we use this asymptotic distribution to prove asymptotic consistency of M_n against any fixed alternative $G \neq F$, as well as against series of converging alternatives $G_n \to F$ provided that the convergence in the supremum norm $\|G_n - F\|_\infty$ is sufficiently slow. Then, following the work of Cai and Wu (2014) we show that M_n is adaptively optimal for detecting a broad family of sparse mixtures.

In a second contribution, we devise in Section 5 an $O(n^2)$ algorithm to compute p-values for *any* supremum-based one-sided test. Particular examples include HC as well as the one-sided $M_n^\pm$ and $R_n^\pm$ statistics of Berk and Jones.

Finally, in Section 6 we compare the power of M_n to other tests under the following settings: i) a change in the mean or variance of a standard Gaussian distribution; and ii) rare-weak sparse Gaussian mixtures; These results showcase scenarios where M_n has improved power compared to common tests. For other

examples involving real data and concrete applications, see Aldor-Noiman et al. (2014); Li and Siegmund (2015); Kaplan and Goldman (2014).

2. The Kolmogorov-Smirnov, Anderson-Darling and higher criticism statistics

Let us first introduce some notation. For a given sample $x_1, \ldots, x_n$, we denote by $x_{(i)}$ the i-th sorted observation (i.e. $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$), by $u_i = F(x_i)$, and by $u_{(i)} = F(x_{(i)})$ where F denotes the null distribution. Finally, we denote the empirical distribution by $\hat{F}_n(x) = \frac{1}{n}\sum_i \mathbf{1}(x_i \leq x)$.

The standard definition of the KS test statistic is based on a (two-sided) L_∞ distance over a continuous variable $x \in \mathbb{R}$,

$$K_n := \sqrt{n} \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right|. \tag{2.1}$$

Although Eq. (2.1) involves a supremum over $x \in \mathbb{R}$, in what follows we instead use an equivalent discrete formulation, whereby the two-sided KS statistic is the maximum of a pair of discrete one-sided statistics, $K_n := \max(K_n^-, K_n^+)$, where

$$K_n^- := \sqrt{n} \max_i \left(u_{(i)} - \frac{i-1}{n} \right), \quad K_n^+ := \sqrt{n} \max_i \left(\frac{i}{n} - u_{(i)} \right). \tag{2.2}$$

By the definition of $\hat{F}_n$, under the null hypothesis that all $x_i \sim F$ we have

$$n\hat{F}_n(x) \sim Binomial(n, F(x)) \qquad \forall x \in \mathbb{R}.$$

Hence, $\mathbb{E}[\hat{F}_n(x)] = F(x)$ and $Var[\hat{F}_n(x)] = \frac{1}{n}F(x)(1-F(x))$. The latter varies significantly throughout the range of x, attaining a maximum at the median of the distribution and smaller values near the tails.

Anderson and Darling (1952) were among the first to suggest different weights to deviations at different locations. Based on a weight function $\psi : [0,1] \to \mathbb{R}$. they proposed a weighted L_2 statistic

$$\mathrm{AD}_{n,\psi} = \int_{-\infty}^{+\infty} n \left(\hat{F}_n(x) - F(x) \right)^2 \psi(F(x)) f(x) dx \,, \tag{2.3}$$

and a lesser-known weighted L_∞ statistic, defined as

$$\mathrm{AD}_{n,\psi}^{sup} = \sup_{x \in \mathbb{R}} \sqrt{n} |\hat{F}_n(x) - F(x)| \sqrt{\psi(F(x))} \,. \tag{2.4}$$

Specifically, Anderson and Darling (1952) suggested to use the weight function $\psi(x) = \frac{1}{x(1-x)}$ which standardizes the variance of $\hat{F}_n(x)$.

Closely related to Eq. (2.4) is the Higher Criticism statistic, whose two variants below can be viewed as one-sided GOF test statistics,

$$\mathrm{HC}_n^{2004} := \sqrt{n} \max_{1 \leq i \leq \alpha_0 \cdot n} \frac{\frac{i}{n} - u_{(i)}}{\sqrt{u_{(i)}(1-u_{(i)})}} \qquad \text{(Donoho and Jin, 2004)}, \tag{2.5}$$

$$\mathrm{HC}_n^{2008} := \sqrt{n} \max_{1 \le i \le \alpha_0 \cdot n} \frac{\frac{i}{n} - u_{(i)}}{\sqrt{\frac{i}{n}(1-\frac{i}{n})}} \qquad \text{(Donoho and Jin, 2008).} \quad (2.6)$$

Indeed, the HC_n^{2004} test with $\alpha_0 = 1$ is equivalent to a one-sided variant of the $\mathrm{AD}_{n,\psi}^{sup}$ test with $\psi(x) = 1/x(1-x)$.

2.1. Order statistics of uniform random variables

By the *probability integral transform*, if $X \sim F$ with F a continuous cdf, then $Y = F(X)$ follows a uniform distribution $Y \sim U[0,1]$. Hence, under the null, the transformed values $u_i = F(x_i)$ are an i.i.d. sample from the $U[0,1]$ distribution and the sorted values $u_{(i)} = F(x_{(i)})$ are their *order statistics*. In particular, the distribution of the i-th order statistic, $U_{(i)}$, is given by

$$U_{(i)} \sim Beta(i, n-i+1), \quad (2.7)$$

with the following mean and variance

$$\mathbb{E}[U_{(i)}] = \frac{i}{n+1} \qquad Var(U_{(i)}) = \frac{i(n-i+1)}{(n+1)^2(n+2)}. \quad (2.8)$$

We now relate the KS and HC tests to $U[0,1]$ order statistics. Up to a small $O(1/\sqrt{n})$ correction, the one sided KS statistic of Eq. (2.2) is the maximal deviation of the n different uniform order statistics from their expectations,

$$K_n^+ = \max_i \sqrt{n} \left(\mathbb{E}[U_{(i)}] - u_{(i)} \right) + O\left(\tfrac{1}{\sqrt{n}}\right) . \quad (2.9)$$

The variance of each $U_{(i)}$ is different, with a maximum at $i = n/2$. Hence the largest deviation tends to occur near the center. Importantly, such deviations can mask small, but statistically significant, deviations at the tails, leading to poor tail sensitivity (Mason and Schuenemeyer, 1983; Calitz, 1987).

In contrast, up to a small correction term, the HC^{2008} statistic normalizes the difference $\mathbb{E}[U_{(i)}] - u_{(i)}$ by its standard deviation,

$$\mathrm{HC}_n^{2008} = \max_i \sqrt{n} \frac{i/n - u_{(i)}}{\sqrt{i/n(1-i/n)}} = \max_i \frac{\mathbb{E}[U_{(i)}] - u_{(i)}}{stdev[U_{(i)}]} \left(1 + O(1/n)\right) , \quad (2.10)$$

and the HC^{2004} / AD^{sup} statistics perform a similar normalization. Such normalizations are common when comparing Gaussian variables with different variances. Indeed, at indices $1 \ll i \ll n$, the distribution of $U_{(i)}$ is close to Gaussian. However, this is not the case when i is fixed and $n \to \infty$ (Keilson and Sumita, 1983). In particular, for any $n \ge 2$ the distribution of $U_{(1)}$ is monotone and heavily skewed towards zero. In section 6.1 we demonstrate and explain analytically why the normalization (2.10) can adversely affect the detection power of HC.

3. The exact Berk-Jones statistics

The discussion above demonstrates that both the KS and HC statistics do not uniformly calibrate the deviations of $u_{(i)}$ over the entire range $i \in \{1, \ldots, n\}$. In this paper we study the M_n, M_n^+ and M_n^- statistics, whose key underlying principle can be described as looking for the deviation $\mathbb{E}[U_{(i)}] - u_{(i)}$ which is *most statistically significant*. In details, for each transformed order statistic $u_{(i)}$, we first compute a one-sided p-value, according to its null distribution of $\text{Beta}(i, n - i + 1)$. This p-value is given by

$$p_{(i)} := \Pr\left[\text{Beta}(i, n - i + 1) < u_{(i)}\right]. \tag{3.1}$$

Then, in analogy to KS, we define the one-sided M_n^-, M_n^+ and two-sided M_n statistics by

$$M_n^+ := \min_{1 \le i \le n} p_{(i)}, \quad M_n^- := \min_{1 \le i \le n} \left(1 - p_{(i)}\right) \text{ and } M_n := \min\{M_n^+, M_n^-\}. \tag{3.2}$$

In contrast to the KS statistic, whose range is $[0, \infty)$ and for which large values lead to a rejection of the null, the M_n statistic is always in $[0, 1]$, with *small values* indicating a *bad fit* to the null hypothesis. Note that $p_{(i)} = I_{u_{(i)}}(i, n - i + 1)$, where $I_x(\alpha, \beta)$ is the regularized incomplete Beta function. This function is commonly available in standard mathematical packages, hence the numerical evaluation of the statistics M_n and $M_n^{\pm}$ is straightforward.

Independently of our work, test procedures of the form $M_n < c$ have been recently suggested in several different papers (Mary and Ferrari, 2014; Kaplan and Goldman, 2014; Gontscharuk, Landwehr and Finner, 2016). However, a close examination reveals that the definitions in Eq. (3.2) are in fact *equivalent* to those proposed by Berk and Jones (1979). In contrast to our motivation, their derivation of M_n followed a different path, building upon their earlier work on relatively optimal combinations of test statistics (Berk and Jones, 1978).

Berk and Jones (1979) also defined the R_n, R_n^+ and R_n^- statistics, as approximations to the M_n, M_n^+ and M_n^- statistics. At the time, this was necessary because computers and software to calculate the tails of a Beta distribution were not as widespread as today. As a result, the approximate statistics became known as *the* Berk-Jones statistics, whereas the exact M_n statistics seem to have received far less attention. With today's widespread availability of computers, direct calculation of the exact statistics poses no difficulty, and their approximation is no longer necessary.

In the following sections we derive the asymptotic null distribution of the M_n, M_n^+ and M_n^- statistics, present an $O(n^2)$ numerical procedure to compute exact p-values for M_n^+ and M_n^-, and empirically compare their detection power to other GOF tests in several simulations.

3.1. Confidence bands

Often, one is interested not only in the magnitude of the most statistically significant deviation from the null hypothesis, as can be measured by M_n or

other statistics, but also in gaining insight into the nature of the deviations throughout the entire range of the sample set. One common practice is to draw a Q-Q scatter plot of the points $\{(F^{-1}(\frac{i}{n+1}), x_{(i)})\}_{i=1}^{n}$. From Eq. (2.8) it follows that under the null $F(x_{(i)}) = u_{(i)} \approx \frac{i}{n+1}$, and hence the Q-Q plot should be concentrated around the $x = y$ diagonal.

Similar to Owen (1995), who constructed α-level confidence bands around the diagonal based on the R_n statistic, one can instead use the M_n statistic. Let $c_\alpha \in [0, 1]$ be the M_n threshold that corresponds to an α-level test. i.e.

$$\Pr[M_n < c_\alpha | \mathcal{H}_0] = \alpha.$$

By definition (3.2), $M_n > c_\alpha$ if and only if the transformed order statistics all satisfy $b_i < u_{(i)} < B_i$ where b_i and B_i are the c_α and $1 - c_\alpha$ quantiles of the Beta$(i, n - i + 1)$ distribution, respectively. Upon making the inverse transformation $x_{(i)} = F^{-1}(u_{(i)})$, this yields confidence bands for the entire Q-Q plot. In the Gaussian case, these confidence bands are precisely those of Aldor-Noiman et al. (2014). For a related construction of confidence bands and further discussion, see Duembgen and Wellner (2014).

4. Theoretical properties of the exact Berk-Jones statistics

The asymptotic null distribution of the M_n^+, M_n^- and M_n statistics is given by the following theorem.

Theorem 4.1. *Under the null hypothesis, for any fixed $x > 0$,*

$$\Pr\left[M_n^{\pm} < \frac{x}{2\log n \log\log n} \middle| \mathcal{H}_0\right] \xrightarrow{n\to\infty} 1 - e^{-x}, \tag{4.1}$$

$$\Pr\left[M_n < \frac{x}{2\log n \log\log n} \middle| \mathcal{H}_0\right] \xrightarrow{n\to\infty} 1 - e^{-2x}. \tag{4.2}$$

This result was recently proved by Gontscharuk and Finner (2016). Their proof is based on a detailed analysis of the local levels of the HC statistic (Gontscharuk, Landwehr and Finner, 2016). In a supplementary note we present a different proof, which approximates the distribution of each $U_{(i)}$ by a Gaussian variable and then adapts known results from the theory of standardized empirical processes. We also use this approximation technique to prove the asymptotic optimality of M_n for detecting sparse mixture models, as described below.

Using the asymptotic distribution, one can construct asymptotic α-level tests and prove the consistency of tests based on the M_n statistic. In fact, in the following subsection we show that M_n is consistent even for a series of converging alternatives $G_n \xrightarrow{n\to\infty} F$, provided that this convergence is sufficiently slow. Similar properties hold for KS and are considered desirable for any GOF statistic (Lehmann and Romano, 2005, Chapter 14). In Section 4.2 below, we show that the M_n statistic is asymptotically optimal for detecting deviations from a Gaussian distribution for a wide class of rare-weak contamination models.

4.1. Asymptotic consistency of M_n

Next, we study the asymptotics of M_n under various alternatives. First, we consider the case of a *fixed* alternative.

Theorem 4.2. *Let* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} G \neq F$. *Then, for any* $\epsilon > 0$

$$\Pr\left[M_n(F(X_1), \ldots, F(X_n)) < \frac{1+\epsilon}{4\|G-F\|_\infty^2} \cdot \frac{1}{n} \middle| \mathcal{H}_1\right] \xrightarrow{n\to\infty} 1\,. \tag{4.3}$$

Combining Theorems 4.1 and 4.2, we obtain the following key result.

Corollary 4.1. *M_n is consistent.*

In other words, as $n \to \infty$ the M_n statistic perfectly distinguishes between the null hypothesis F and any fixed alternative $G \neq F$. In fact, as the following corollary shows, M_n even distinguishes between F and a series of converging alternatives $\{G_n\}_{n=1}^\infty$ such that $G_n \to F$, provided that this convergence is sufficiently slow.

Corollary 4.2. *For any fixed* $\epsilon > 0$, *a test based on the M_n statistic is consistent over all alternatives* $\{G_n\}_{n=1}^\infty$ *satisfying*

$$\sqrt{n}\|G_n - F\|_\infty \sqrt{\log n \log\log n} \longrightarrow \infty\,. \tag{4.4}$$

We note that Berk and Jones (1979) have investigated the limiting behavior of R_n and M_n and showed that under specific conditions on the alternative distribution (Berk and Jones, 1979, Theorem 4.1) both R_n^+ and $-\frac{1}{n}\log M_n^+$ converge to a constant which depends on the alternative distribution. These results were greatly extended by Jager and Wellner (2007) for a family of GOF statistics based on phi-divergences. In contrast, our Theorem 4.2 merely gives a stochastic upper bound on M_n, but one that does not require the alternative distribution to satisfy any particular properties.

4.2. Sparse mixture detection

Motivated by the works of Donoho and Jin (2004) and Cai and Wu (2014), we now study the properties of M_n under the following class of sparse mixture models. Suppose that under the null hypothesis $X_i \overset{i.i.d.}{\sim} F$, whereas under the alternative a *small fraction* ϵ_n of the variables are contaminated and have a different distribution G_n. The corresponding hypothesis testing problem is

$$\mathcal{H}_0 : X_i \overset{i.i.d.}{\sim} F \quad vs. \quad \mathcal{H}_1 : X_i \overset{i.i.d.}{\sim} (1-\epsilon_n)F + \epsilon_n G_n\,. \tag{4.5}$$

Such models have been analyzed, among others, by Ingster (1997), Donoho and Jin (2004) and Cai and Wu (2014). Let us briefly review some results regarding these models, first for the *Gaussian mixture model*, where $F = \mathcal{N}(0,1)$ and $G_n = N(\mu_n, 1)$,

$$\mathcal{H}_0 : X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \quad vs. \quad \mathcal{H}_1 : X_i \overset{i.i.d.}{\sim} (1-\epsilon_n)\mathcal{N}(0,1) + \epsilon_n \mathcal{N}(\mu_n, 1)\,. \tag{4.6}$$

Recall that for $n \gg 1$, the maximum of n i.i.d. standard Gaussian variables is sharply concentrated around $\sqrt{2\log n}$. Thus, for any fixed $\epsilon_n = \epsilon$, as $n \to \infty$, contamination strengths $\mu_n > \sqrt{2\log n}(1+\delta)$ are perfectly detectable by the maximum statistic $\max x_i$. Similarly, for any fixed $\mu_n = \mu$, sparsity levels $\epsilon_n \gg n^{-1/2}$ visibly shift the overall mean of the samples, and hence as $n \to \infty$, can be perfectly detected by the sum statistic $\sum x_i$. These cases lead one to consider the scaling $\epsilon_n = n^{-\beta}$, $\mu_n = \sqrt{2r\log n}$ and examine the asymptotic detectability in the (r, β) plane (Ingster, 1997). Since any point (r, β) with $r > 1$ or $\beta < 0.5$ is easily detectable, the interesting region is where both $0 < r < 1$ and $0.5 < \beta < 1$.

For the model (4.6), if ϵ_n and μ_n are known, both $\mathcal{H}_0$ and $\mathcal{H}_1$ are simple hypotheses, and the optimal test is the Likelihood Ratio (LR). Its performance was studied by Ingster (1997), who found a sharp detection boundary in the (r, β) plane, given by

$$r_{min}(\beta) = \begin{cases} \beta - 0.5 & 0.5 < \beta \le 0.75, \\ (1 - \sqrt{1-\beta})^2 & 0.75 \le \beta < 1. \end{cases} \tag{4.7}$$

Namely, as $n \to \infty$, the sum of type-I and type-II error rates of the LR test tends to 0 or 1 depending on whether (r, β) lies above or below this curve.

While the LR test is optimal, it may be inapplicable as it requires precise knowledge of the model parameters μ and ϵ. Importantly, both the Higher Criticism statistic based on Eq. (2.5) and the approximate Berk-Jones test R_n were proven to achieve the optimal asymptotic detection boundary without such knowledge (Donoho and Jin, 2004, Theorems 1.2, 1.6). Thus, both statistics are *adaptively optimal* for the sparse Gaussian mixture detection problem in an asymptotic sense. In what follows, we prove that M_n is also adaptively optimal.

Recently, Cai and Wu (2014) studied more general sparse mixtures of the form (4.5) where the null distribution is Gaussian and $\epsilon_n = n^{-\beta}$, but G_n is not necessarily Gaussian. The following is a simplified version of their Theorem 1, describing the asymptotic detectability under this model.

Theorem 4.3. *Let G_n be a continuous distribution with density function g_n. If the following limit exists for all $u \in \mathbb{R}$*

$$h(u) := \lim_{n\to\infty} \frac{\log\left(\sqrt{2\pi} g_n(u\sqrt{2\log n})\right)}{\log n} \tag{4.8}$$

then the hypothesis testing problem (4.5) with $F = \mathcal{N}(0,1)$ and $\epsilon_n = n^{-\beta}$ has an asymptotic detection threshold given by

$$\beta^\star = \frac{1}{2} + \max\left(0, \sup_{u\in\mathbb{R}}\left\{h(u) + \frac{1}{2}\min(1, u^2)\right\}\right). \tag{4.9}$$

Namely, for any $\beta < \beta^\star$ the error rate of the likelihood ratio test tends to zero as $n \to \infty$.

In their paper, Cai and Wu (2014) proved that HC is adaptively optimal under the conditions of Theorem 4.3. As we now show, M_n has the same adaptive

optimality properties, in particular for the Gaussian mixture model of Eq. (4.6). We note that for finite sample sizes M_n may have considerably higher power compared to HC as we show in Section 6.

Theorem 4.4. *Let G_n be a continuous distribution satisfying Eq. (4.8) and let*

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} (1 - n^{-\beta})\mathcal{N}(0,1) + n^{-\beta} G_n.$$

If $\beta < \beta^*$ where β^* is given by Eq. (4.9), then there is some $\epsilon > 0$ such that

$$\Pr\left[M_n < \frac{1}{(\log n)^{1+\epsilon}} \right] \xrightarrow{n\to\infty} 1.$$

where $\mathcal{N}(0,1)$ is taken as the null distribution.

The proof is in the appendix. Combining this result with Theorem 4.1 gives

Corollary 4.3. *For any $\epsilon > 0$ and $\beta < \beta^*$, the test*

$$M_n < \frac{1}{\log n (\log\log n)^{1+\epsilon}}$$

perfectly separates, as $n \to \infty$, the null distribution $\mathcal{N}(0,1)$ from a sparse-mixture alternative of the form $(1 - n^{-\beta})\mathcal{N}(0,1) + n^{-\beta} G_n$. Namely, inside the asymptotic detectability region, the error rate of the test tends to zero.

5. Computing p-values

For the classical one-sided and two-sided KS statistics, there are many methods to compute the corresponding p-values, see Durbin (1973); Marsaglia, Tsang and Wang (2003); Brown and Harvey (2008a,b). Most of these methods, however, are particular to KS and inapplicable to other GOF statistics. Notable exceptions include (Noé, 1972; Friedrich and Schellhaas, 1998; Khmaladze and Shinjikashvili, 2001) whose recursion formulas can compute the p-value of any supremum-based two-sided (or one-sided) statistic using $O(n^3)$ operations and the recent algorithm of Moscovich and Nadler (2016) that runs in $O(n^2 \log n)$ steps. For another recent work with time complexity $O(n^3)$, see Barnett and Lin (2014).

In this section we present an $O(n^2)$ algorithm to compute p-values of any supremum-based *one-sided* test statistic, including $M_n^+, M_n^-, R_n^+, R_n^-$ and the Higher Criticism. Furthermore, it may be used to obtain approximations of the p-value of two-sided statistics.

To describe our approach, note that by Eq. (3.1),

$$\Pr\left[M_n^+ \geq c | \mathcal{H}_0\right] = \Pr\left[\forall i : p_{(i)} \geq c | \mathcal{H}_0\right] = \Pr\left[\forall i : L_i^n(c) \leq u_{(i)} \leq 1 | \mathcal{H}_0\right], \quad (5.1)$$

where $L_i^n(c)$ denotes the inverse of the regularized incomplete Beta function, satisfying

$$\Pr\left[\mathrm{Beta}(i, n-i+1) < L_i^n(c)\right] = c.$$

Procedures to compute $L_i^n(c)$ are available in most mathematical packages.

Note that under the null, the n unsorted variables are uniformly distributed, $U_i \overset{i.i.d.}{\sim} U[0,1]$, and hence their joint density equals 1 inside the n-dimensional box $[0,1]^n$. Given that there are $n!$ distinct permutations of n indices, the joint probability density of the random vector of sorted values $(U_{(1)},\ldots,U_{(n)})$ is

$$f(U_{(1)},\ldots,U_{(n)}) = \begin{cases} n! & \text{if } 0 \le U_{(1)} \le \ldots \le U_{(n)} \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

From this it readily follows that

$$\begin{aligned} \Pr\left[M_n^+ \ge c|\mathcal{H}_0\right] &= n! Vol\{(U_{(1)},\ldots,U_{(n)}) \mid \forall i : L_i^n(c) \le U_{(i)} \le U_{(i+1)}\} \\ &= n! \int_{L_n^n(c)}^{1} dU_{(n)} \int_{L_{n-1}^n(c)}^{U_{(n)}} dU_{(n-1)} \cdots \int_{L_2^n(c)}^{U_{(3)}} dU_{(2)} \int_{L_1^n(c)}^{U_{(2)}} dU_{(1)} \,. \end{aligned} \tag{5.2}$$

Eq. (5.2) is the key to fast calculation of p-values for M_n^+ or other one-sided tests. The idea is to evaluate this multiple integral, from right to left. The first integral yields a polynomial of degree 1 in $U_{(2)}$, the next integral yields a polynomial of degree 2 in $U_{(3)}$ and so on. While we have not found simple explicit formulas for the resulting polynomials, their numerical integration is straightforward. We store $d+1$ coefficients for the d-th degree polynomial, and its numerical integration takes $O(d)$ operations. Hence, the total time complexity is $O(n^2)$.

Still, there are some numerical difficulties with this approach: A naïve implementation suffers from a fast accumulation of numerical errors and breaks down completely at $n \approx 150$. Nonetheless, as described in the appendix, with a modified procedure and using extended precision (80-bit) floating point numbers, this accumulation of errors is significantly attenuated, allowing accurate calculation of one-sided p-values for up to $n \approx 50,000$ samples. The actual running time of our freely available C++ implementation is about one second for $n = 4000$ samples using a present-day PC.

The following theorem provides simple upper and lower bounds for the p-value of the two-sided M_n, in terms of its one-sided p-values,

Theorem 5.1. *For any $c \in [0,1]$, let $q_c := \Pr[M_n^+ \le c \mid \mathcal{H}_0]$. Then,*

$$2q_c - q_c^2 \le \Pr[M_n \le c \mid \mathcal{H}_0] \le 2q_c. \tag{5.3}$$

Furthermore, as $n \to \infty$,

$$\Pr\left[M_n \le c \,|\mathcal{H}_0\right] \xrightarrow{n\to\infty} 2q_c - q_c^2. \tag{5.4}$$

Remark 5.1. As mentioned above, our algorithm can compute the p-value of any supremum-type one-sided test statistic. The only difference lies in the coefficients $L_i^n(c)$ of Eq. (5.2), which depend on the specific test statistic. For example, the HC^{2008} test of Eq. (2.6) satisfies

$$\Pr\left[\mathrm{HC}^{2008} < c|\mathcal{H}_0\right] = \Pr\left[\forall i : \tfrac{i}{n} - c\sqrt{\tfrac{i}{n^2}\left(1-\tfrac{i}{n}\right)} < U_{(i)} \le U_{(i+1)} \middle| \mathcal{H}_0\right].$$

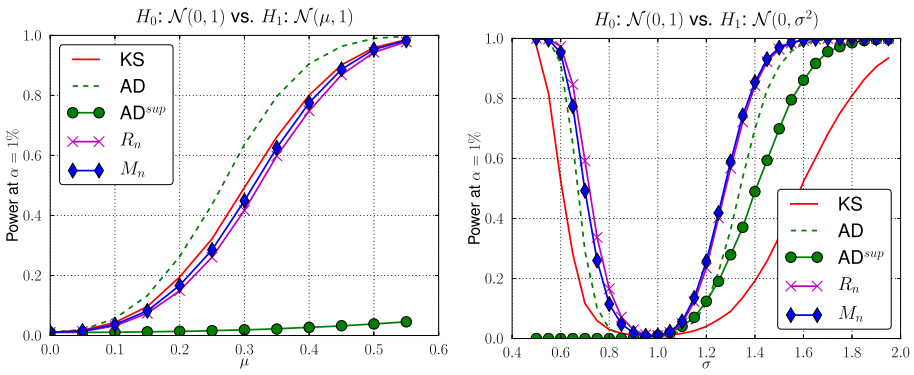

FIG 1. *Power comparisons of two-sided tests for detecting lack-of-fit to a standard Gaussian distribution (at significance level $\alpha = 1\%$) with $n = 100$ samples. (left panel) change in the mean of the distribution; (right panel) change in the variance.*

Thus, for this statistic, $L_i^n(c) = \frac{i}{n} - c\sqrt{\frac{i}{n^2}\left(1 - \frac{i}{n}\right)}$.

Remark 5.2. Historically, an equation similar to (5.2) was derived by Daniels (1945), in an entirely different context. His formula was used in later works to derive closed form expressions for the asymptotic distribution of the KS test statistic. See Durbin (1973) for a survey.

Remark 5.3. To the best of our knowledge, the only other $O(n^2)$ algorithm for computing p-values of L_∞-type one-sided test statistics is that of Kotel′nikova and Khmaladze (1982). Their method is based on a different recursive formula, which involves large binomial coefficients and also requires a careful numerical implementation.

6. Simulation results

6.1. Deviations from a standard Gaussian distribution

We consider a null hypothesis that $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and two alternatives: a shift in the mean, $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu,1)$, or a change in the variance $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$. The left and right panels of Figure 1 compare the power of M_n, KS, AD and AD^{sup} (=two-sided HC) under these two alternatives at a significance level of $\alpha = 1\%$.

For detecting a change in the mean, the M_n test is on par with KS, but the AD test outperforms both. The AD^{sup} test has close to zero power in this benchmark. For detecting a change in the variance, which strongly affects the tails, M_n has a higher detection power throughout the entire range of σ. In contrast, AD^{sup} performs poorly, and has power close to zero when $\sigma < 1$.

As we now show, the poor performance of AD^{sup}/HC stems from its specific normalization of the deviations at the extreme indices $u_{(1)}, u_{(2)}$, etc. To this end, recall that under the null, $\Pr\left[u_{(1)} < x\right] = 1 - (1-x)^n$. Hence, the probability

that the first order statistic is smaller than $1/cn \log\log n$, for some constant $c > 0$ is given by

$$\Pr\left[u_{(1)} < \frac{1}{cn \log\log n}\right] = \frac{1 + o(1)}{c \log\log n}.$$

For such values of $u_{(1)}$, the corresponding HC deviation at the first index is

$$\sqrt{n}\frac{\frac{1}{n} - u_{(1)}}{\sqrt{u_{(1)}(1 - u_{(1)})}} > \sqrt{c \log\log n}(1 + o(1)).$$

It is now instructive to plug in some specific number into the above equations. In particular, for $n = 100$ samples as in Figure 1, a value $c = 65.48$ gives that with probability of 1% the deviation of the first order statistic is at least $\sqrt{c \log\log n} \approx 10$.

Now suppose we conduct an HC test at a false alarm level of $\alpha = 1\%$. The above calculation has two important implications: First, the finite sample threshold of the HC test at $n = 100$ must clearly satisfy $t_\alpha > 10$. This value is significantly larger than its asymptotic value of $\sqrt{2 \log\log n}(1 + o(1)) \approx 1.74$ (see Theorem A.1). Since the decay of $1/\log\log n$ to zero is extremely slow, the above illustrates the very slow convergence of the AD^{sup} or HC distribution to its asymptotic limit. Second, such a high threshold prevents detection of significant deviations near the center of the distribution, as indeed is shown empirically in Figure 1. As an example, a significant deviation from the null of $u_{(n/2)} = 1/4$ which corresponds to about 5.8 standard deviations cannot be detected by the HC test at level $\alpha = 1\%$.

We remark that HC's problematic handling of $u_{(1)}$ was already noted by Donoho and Jin (2004), and discussed in several recent works (Walther, 2013; Li and Siegmund, 2015; Gontscharuk, Landwehr and Finner, 2015, 2016). Finally, we note that in our numerical example, removing $u_{(1)}$ from the HC test does not resolve the problem, since the next extreme order statistics $u_{(2)}, u_{(3)}$ etc., also have a non-negligible probability to induce very large HC values. In contrast, the M_n statistic puts all of these deviations on an equal scale.

6.2. Detecting Sparse Gaussian mixtures

Next, we consider the problem of detecting a sparse Gaussian mixture of the form (4.6), where the parameter μ is assumed positive. We hence compare the following four one-sided test statistics: $\max X_i$, $\sum X_i$, HC^{2004} and M_n^+.

Figure 2 compares the resulting Receiver Operating Characteristic (ROC) curves for two choices of ϵ and μ, both with $n = 10,000$ samples. The optimal curve is that of the likelihood-ratio test, which unlike the other statistics, is model specific and requires explicit knowledge of the values of ϵ and μ.

While asymptotically as $n \to \infty$, both HC^{2004} and M_n^+ achieve the same performance as that of the optimal LR test, for finite values of n, as seen in the figure, the gap in detection power may be large. Moreover, for some (μ, ϵ)

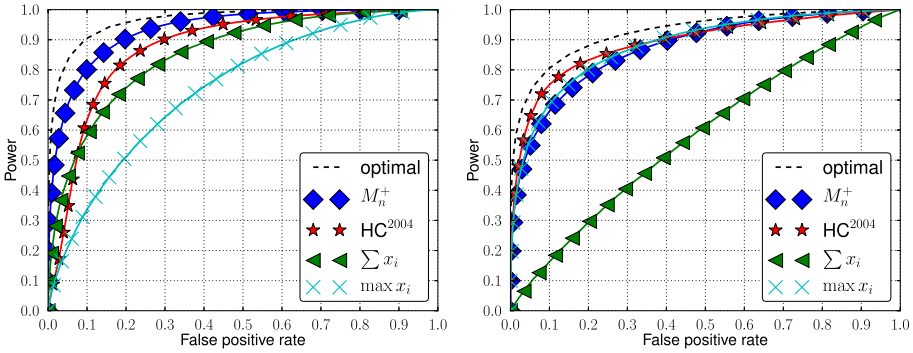

FIG 2. *ROC curves for the rare-weak Gaussian mixture model with* $n = 10,000$ *samples, and with sparsity and contamination levels:* $\epsilon = 0.01, \mu = 1.5$ *(left);* $\epsilon = 0.001, \mu = 3$ *(right).*

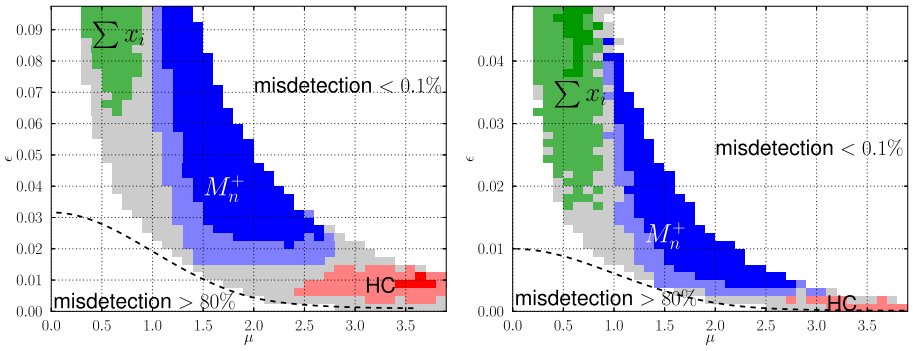

FIG 3. *(Best viewed in color) Comparison of tests for detecting rare-weak Gaussian mixtures* $(1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(\mu,1)$ *vs.* $\mathcal{N}(0,1)$*. Colored blobs represent regions where the misdetection rate of the second-best test divided by that of the best test was larger than 1.1. The dark centers signify regions where this ratio was larger than 1.5. The gray band delineates the zone where misdetection is in the range* $0.1\% - 80\%$*. The dotted line is the asymptotic detection boundary* (4.7) *when substituting* $\epsilon = n^{-\beta}$*,* $\mu = \sqrt{2r\log n}$*. Left panel:* $n = 1,000$*; right panel:* $n = 10,000$*.*

values, HC^{2004} achieves a higher ROC curve, whereas for others M_n^+ is better. A natural question thus follows: For a finite number of samples n, as a function of the two parameters ϵ and μ, which of these four tests has greater power? To study this question, we made the following extensive simulation: for many different values of (μ, ϵ), we empirically computed the detection power of the four tests mentioned above at a significance level of $\alpha = 5\%$, both for $n = 1000$ and for $n = 10,000$ samples. For each sparsity value ϵ and contamination level μ we declared that a test T_1 was a clear winner if it had a significantly lower misdetection rate, namely if $\min_{j=2,3,4} \Pr[T_j = \mathcal{H}_0|\mathcal{H}_1]/\Pr[T_1 = \mathcal{H}_0|\mathcal{H}_1] > 1.1$.

Figure 3 shows the regions in the (μ, ϵ) plane where different tests were declared as clear winners. First, as the figure shows, at the upper left part in the

(μ, ϵ) plane, $\sum X_i$ is the best test statistic. This is expected, since in this region ϵ is relatively large and leads to a significant shift in the mean of the distribution. At the other extreme, in the lower right part of the (μ, ϵ) plane, where ϵ is small but μ is large, very few samples are contaminated and here the HC^{2004} test statistic works best, with the max statistic being a close second. In the intermediate region, which would naturally be characterized as the rare/weak region, it is the M_n^+ test that has a higher power. Second, while not shown in the plot, we note that the M_n^+ test had similar power to that of the R_n^+ test. Finally, in this simulation the HC^{2008} test performed worse than at least one of the other tests for all values of (μ, ϵ).

Appendix A: Auxiliary lemmas

A.1. Asymptotics of the Beta distribution

As is well known, when both $\alpha, \beta \to \infty$, the $\mathrm{Beta}(\alpha, \beta)$ distribution approaches $\mathcal{N}(\mu, \sigma^2)$, where μ and σ are the mean and standard deviation of the Beta random variable. The following lemma quantifies the error in this approximation. For other approximations, see for example Peizer and Pratt (1968); Pratt (1968).

Lemma A.1. *Let $f_{\alpha,\beta}$ be the density of a* $\mathrm{Beta}(\alpha, \beta)$ *random variable and let $g_t(\alpha, \beta)$ be its value at t standard deviations from the mean. i.e.*

$$g_t(\alpha, \beta) = f_{\alpha,\beta}(\mu_{\alpha,\beta} + \sigma_{\alpha,\beta} \cdot t).$$

For any fixed t, as both $\alpha, \beta \to \infty$,

$$\begin{aligned} g_t(\alpha, \beta) = & \frac{e^{-t^2/2}}{\sqrt{2\pi} \cdot \sigma_{\alpha,\beta}} \times \exp\left[\frac{1}{\sqrt{\alpha+\beta+1}}\left(\sqrt{\frac{\alpha}{\beta}} - \sqrt{\frac{\beta}{\alpha}}\right) t\right] \\ & \times \exp\left[O\left(\frac{\beta}{(\alpha+\beta)\alpha} + \frac{\alpha}{(\alpha+\beta)\beta}\right) t^2\right] \\ & \times \exp\left[O\left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right) t^3\right] \times \left(1 + O\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)\right). \end{aligned} \tag{A.1}$$

Remark A.1. For any fixed t, as $\alpha, \beta \to \infty$ all error terms tend to zero, hence demonstrating that the distribution of non-extreme order statistics converges to a Gaussian. However, for this approximation to be accurate, all correction terms must be small, which may require huge sample sizes. As an example, with $t = 2$ standard deviations and $\alpha = n^{1/4}$, to have $|t^3|/\sqrt{\alpha} < 0.1$ we need $n > 1.7 \times 10^{15}$ samples, far beyond the reach of almost any scientific study.

Remark A.2. A closer inspection of the proof below shows that Lemma A.1 continues to hold even if $t = t(\alpha, \beta) \to \infty$, provided that $\alpha, \beta \to \infty$ and

$$t \cdot \max\left(\sqrt{\tfrac{\alpha}{(\alpha+\beta)\beta}}, \sqrt{\tfrac{\beta}{(\alpha+\beta)\alpha}}\right) \to 0\,. \tag{A.2}$$

This shall prove to be useful later on.

Proof of Lemma A.1. For convenience we denote

$$A := \alpha - 1, \quad B := \beta - 1, \quad n := \alpha + \beta - 1 = A + B + 1\,.$$

In terms of these variables, the mean and variance of $\mathrm{Beta}(\alpha, \beta)$ are

$$\mu = \frac{A+1}{n+1}, \quad \sigma^2 = \frac{(A+1)(B+1)}{(n+1)^2(n+2)},$$

whereas its density is $f(x) = \frac{n!}{A!B!} x^A (1-x)^B$. At $x = \mu + \sigma t$, we obtain

$$f(\mu + \sigma t) = \frac{n!}{A!B!} \mu^A (1-\mu)^B \left(1 + \tfrac{\sigma}{\mu} t\right)^A \left(1 - \tfrac{\sigma}{1-\mu} t\right)^B . \tag{A.3}$$

Using Stirling's approximation, that $n! = \sqrt{2\pi n}(n/e)^n (1 + O(1/n))$, and the fact that $\sigma = \sqrt{AB/n^3}(1 + O(1/A + 1/B))$, we obtain that as both $A, B \to \infty$,

$$\frac{n!}{A!B!} \mu^A (1-\mu)^B = \frac{1}{\sqrt{2\pi} \cdot \sigma} \left(1 + O\left(\frac{1}{A} + \frac{1}{B}\right)\right).$$

Next, we write the remaining terms in (A.3),

$$\left(1 + \tfrac{\sigma}{\mu} t\right)^A \left(1 - \tfrac{\sigma}{1-\mu} t\right)^B = \exp\left[A \ln(1 + \tfrac{\sigma}{\mu} t) + B \ln(1 - \tfrac{\sigma}{1-\mu} t)\right]. \tag{A.4}$$

Note that as $A, B \to \infty$, both σ/μ and $\sigma/(1-\mu)$ tend to zero. Hence, for either a fixed t, or $t = t(\alpha, \beta)$ slowly growing to ∞ such that Eq. (A.2) holds, we may replace the logarithms in Eq. (A.4) by their Taylor expansion with small approximation errors

$$\begin{aligned}\left(1 + \tfrac{\sigma}{\mu} t\right)^A \left(1 - \tfrac{\sigma}{1-\mu} t\right)^B &= \exp\left[\sigma t \left(\tfrac{A}{\mu} - \tfrac{B}{1-\mu}\right) - \tfrac{\sigma^2 t^2}{2} \left(\tfrac{A}{\mu^2} + \tfrac{B}{(1-\mu)^2}\right)\right] \\ &\quad \times \exp\left[O\left(\tfrac{A\sigma^3}{\mu^3} + \tfrac{B\sigma^3}{(1-\mu)^3}\right) t^3\right].\end{aligned} \tag{A.5}$$

Simple algebra gives

$$\sigma t \left(\tfrac{A}{\mu} - \tfrac{B}{1-\mu}\right) = \tfrac{t}{\sqrt{n+2}} \left(\sqrt{\tfrac{A+1}{B+1}} - \sqrt{\tfrac{B+1}{A+1}}\right).$$

Similarly,

$$\tfrac{\sigma^2 t^2}{2} \left(\tfrac{A}{\mu^2} + \tfrac{B}{(1-\mu)^2}\right) = \tfrac{t^2}{2} \left(1 + O\left(\tfrac{B}{nA} + \tfrac{A}{nB}\right)\right).$$

Finally, as $A, B \to \infty$, the cubic term in Eq. (A.5) is of order $O\left(\frac{1}{\sqrt{A}} + \frac{1}{\sqrt{B}}\right) t^3$. Combining all of these results concludes the proof of the lemma. □

We present a simple corollary of Lemma A.1, which shall prove useful below in studying the asymptotic behavior of M_n.

Corollary A.1. *let $\{\alpha_n\}$ be a sequence of numbers converging to infinity. Let μ_n, σ_n^2 and f_n denote the mean, variance and density of a Beta$(\alpha_n, n-\alpha_n+1)$ distribution, respectively. Furthermore, let $g(n)$ be any positive function satisfying $g(n) = o(\min\{\alpha_n, n-\alpha_n\})$. Then, for large values of n we have a lower bound on the density near the mean,*

$$f_n(\mu_n + \sigma_n \cdot t) \geq \frac{e^{-t^2/2}}{\sqrt{2\pi}\cdot\sigma_n}\left(1 - \frac{t^3}{\sqrt{g(n)}} - \frac{1}{g(n)}\right). \tag{A.6}$$

Proof. Follows from an inspection of the various error terms in Eq. (A.1). □

A.2. Supremum of the standardized empirical process

The standardized empirical process plays a central role in our analysis of the M_n statistic. We begin with its definition followed by several known results regarding the magnitude and location of its supremum.

Definition A.1. Let $X_1, \ldots, X_n$ be i.i.d. random variables from some continuous distribution F and let $\hat{F}_n(x) = \frac{1}{n}\sum_i \mathbf{1}(X_i \leq x)$ denote their empirical cdf. The normalized empirical process is defined as

$$V_n(x) = \sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{F(x)(1-F(x))}} \quad \text{for} \quad 0 < F(x) < 1\,. \tag{A.7}$$

Similarly, the standardized empirical process is

$$\hat{V}_n(x) = \sqrt{n}\frac{\hat{F}_n(x) - F(x)}{\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}} \quad \text{for} \quad 0 < F_n(x) < 1\,. \tag{A.8}$$

Of particular interest to us is the supremum of $\hat{V}_n$. Theorem 1 of Eicker (1979) gives the asymptotic distribution of this supremum, see also Csörgő et al. (1986):

Theorem A.1. *Let $U_1, \ldots, U_n \overset{i.i.d.}{\sim} U[0,1]$. As $n \to \infty$,*

$$\Pr\left[\sup_{U_{(1)}<u<U_{(n)}} \hat{V}_n(u) < \sqrt{2\log\log n} + \frac{\log\log\log n}{2\sqrt{2\log\log n}} + \frac{1}{\sqrt{2\log\log n}}\cdot t\right] \to e^{-e^{-t}/\sqrt{\pi}}.$$

Furthermore, the next lemma, which follows from the main Theorem of Jaeschke (1979), implies that this supremum is rarely attained at one of the extreme order statistics.

Lemma A.2. *Let $k > 0$ and let I be the union of intervals containing the first and last $\log^k n$ order statistics, $I = (U_{(1)}, U_{(\log^k n)}] \cup [U_{(n-\log^k n)}, U_{(n)})$. Then*

$$\Pr\left[\sup_{u\in I}\hat{V}_n(u) < \sup_{U_{(1)}<u<U_{(n)}}\hat{V}_n(u)\right] \xrightarrow{n\to\infty} 1\,. \tag{A.9}$$

Appendix B: Proofs of theorems

B.1. Proof of Theorem 4.2

Let $t_0 \in \mathbb{R}$ be some point that satisfies

$$|G(t_0) - F(t_0)| = \|G - F\|_\infty .$$

Without loss of generality, we assume that $F(t_0) < G(t_0)$ and derive an upper bound on M_n^+ (in the opposite case the same upper bound would be obtained on M_n^-). Let i_* denote the number of random variables X_i smaller than t_0. Since for all i, $\Pr[X_i < t] = G(t)$, the random variable i_* follows a binomial distribution,

$$i_* \sim Binomial(n, G(t_0)) .$$

Since $F(t_0) < G(t_0)$, for any fixed $0 < \lambda < 1$,

$$\Pr\left[\frac{i_*}{n+1} > \lambda G(t_0) + (1-\lambda)F(t_0)\right] \xrightarrow{n\to\infty} 1 .$$

This implies that with probability tending to one,

$$\frac{i_*}{n+1} - F(t_0) > \lambda\left(G(t_0) - F(t_0)\right). \tag{B.1}$$

We show that this implies Eq. (4.3) of the theorem. To this end, recall that

$$M_n^+ \le p_{(i_*)} = \Pr\left[\mathrm{Beta}(i_*, n - i_* + 1) < U_{(i_*)}\right] .$$

By definition, $X_{(i_*)} < t_0$ and therefore $U_{(i_*)} := F(X_{(i_*)}) < F(t_0)$. Thus

$$\begin{aligned} M_n^+ &< \Pr\left[\mathrm{Beta}(i_*, n - i_* + 1) < F(t_0)\right] \\ &= \Pr\left[\frac{i_*}{n+1} - \mathrm{Beta}(i_*, n - i_* + 1) > \frac{i_*}{n+1} - F(t_0)\right] . \end{aligned} \tag{B.2}$$

Hence, from (B.1) follows that with probability tending to one,

$$\begin{aligned} M_n^+ &< \Pr\left[\frac{i_*}{n+1} - \mathrm{Beta}(i_*, n - i_* + 1) > \lambda\left(G(t_0) - F(t_0)\right)\right] \\ &< \Pr\left[\left|\frac{i_*}{n+1} - \mathrm{Beta}(i_*, n - i_* + 1)\right| > \lambda\left(G(t_0) - F(t_0)\right)\right] . \end{aligned}$$

Recall that the expectation of $\mathrm{Beta}(i_*, n - i_* + 1)$ is $i_*/(n+1)$ and that for any $1 \le i_* \le n$ its standard deviation is smaller than $1/2\sqrt{n}$. Therefore, by Chebyshev's inequality

$$M_n^+ < \frac{1}{4n\lambda^2\|G - F\|_\infty^2} .$$

Setting $\lambda = 1/\sqrt{1+\epsilon}$ concludes the proof. □

B.2. Proof of Theorem 4.4

We start with two technical lemmas.

Lemma B.1. *Let μ and σ^2 denote the mean and variance of a Beta$(i, n-i+1)$ random variable. For any $x \in [0,1]$,*

$$\frac{\mu - x}{\sigma} = \sqrt{n}\frac{\frac{i}{n} - x}{\sqrt{\frac{i}{n}(1-\frac{i}{n})}}\left(1 + O\left(\frac{1}{n}\right)\right).$$

Proof. Follows by straightforward algebraic manipulations. □

Lemma B.2. *Let $\epsilon > 0$ and $a > 0$ be constants. Let $f : \mathbb{R} \to \mathbb{R}$ be a non-negative function that satisfies the following conditions,*

1. $\int f(x)dx = 1$.
2. $f(x) \geq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}(1-\epsilon)$ *in the range* $x \in [-a, a]$.

Then for any $t \in [0, a]$,

$$\int_{-\infty}^{-t} f(x)dx \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-t^2/2} + \frac{1}{\sqrt{2\pi}}\frac{1}{a}e^{-a^2/2} + \epsilon.$$

Proof.

$$\int_{-\infty}^{-t} f(x)dx \leq 1 - \int_{-t}^{a} f(x)dx \leq 1 - (1-\epsilon)\int_{-t}^{a}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx. \tag{B.3}$$

A simple bound on the Gaussian tail is given by

$$\int_{t}^{\infty} e^{-x^2/2}dx \leq \int_{t}^{\infty}\frac{x}{t}e^{-x^2/2}dx = \frac{1}{t}e^{-t^2/2}. \tag{B.4}$$

Therefore,

$$\begin{aligned}\int_{-t}^{a}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx &= 1 - \int_{a}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx - \int_{-\infty}^{-t}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx \\ &\geq 1 - \frac{1}{\sqrt{2\pi}}\frac{e^{-a^2/2}}{a} - \frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t}.\end{aligned}$$

Plugging this bound into Eq. (B.3) concludes the proof. □

We are now ready to present the main part of the proof. We base our result on the proof of Theorem 4 by Cai and Wu (2014). They show that under the alternative, there exists a fixed $0 < s < 1$ such that

$$\Pr\left[V_n(\sqrt{2s\log n}) > \sqrt{(2+\delta)\log\log n}\right] \xrightarrow{n\to\infty} 1\,, \tag{B.5}$$

where V_n is the normalized empirical process of Eq. (A.7).

Let $i_* = |\{j \,|X_j > \sqrt{2s\log n}\}|$ be the (random) number of observations above $\sqrt{2s\log n}$. Then, from Eq. (B.5) follows that with high probability,

$$\sqrt{n}\frac{i_*/n - U_{(i_*)}}{\sqrt{U_{(i_*)}(1-U_{(i_*)})}} > \sqrt{(2+\delta)\log\log n}\,. \tag{B.6}$$

Let μ_* and σ_*^2 denote the mean and variance of a Beta$(i_*, n-i_*+1)$ random variable and let

$$\tau_* := \frac{\mu_* - U_{(i_*)}}{\sigma_*}$$

be the z-score of $U_{(i_*)}$. With a change of variables $t = (x-\mu_*)/\sigma_*$, we obtain

$$p_{(i_*)} = \int_{-\mu_*/\sigma_*}^{-\tau_*} f_*(\mu_* + \sigma_* t)\cdot\sigma_* dt \le \int_{-\infty}^{\tau_*} f_*(\mu_* + \sigma_* t)\cdot\sigma_* dt.$$

By Lemma B.1,

$$\tau_* = \hat{V}_n(U_{(i_*)})(1+O(1/n)). \tag{B.7}$$

For any $\epsilon_1 > 0$, the following statements hold with probability $> 1-\epsilon_1$ for large values of n:

1. $\log^{12} n < i_* < n - \log^{12} n$ (from Lemma A.2).
2. $\tau_* < \log n$ (from Theorem A.1 and Eq. (B.7)).

By Corollary A.1, for any $t \in [-\log n, \log n]$, if n is large enough, then

$$f_*(\mu_* + \sigma_*\cdot t) \ge \frac{e^{-t^2/2}}{\sqrt{2\pi}\cdot\sigma_*}\left(1 - \frac{1}{\log^2 n}\right).$$

Hence, we may apply Lemma B.2 with $a = \log n$ and obtain that the following holds with high probability,

$$\begin{aligned} p_{(i_*)} &\le \frac{1}{\sqrt{2\pi}}\frac{1}{\tau_*}e^{-\frac{1}{2}\tau_*^2} + \frac{1}{\sqrt{2\pi}}\frac{1}{\log n}e^{-\log^2 n/2} + \frac{1}{\log^2 n} \\ &\le \frac{1}{\sqrt{2\pi}}\frac{1}{\tau_*}e^{-\frac{1}{2}\tau_*^2} + \frac{2}{\log^2 n}. \end{aligned} \tag{B.8}$$

From (B.6) it follows that $\tau_* > \sqrt{(2+\delta)\log\log n}(1+O(1/n))$, and therefore

$$\Pr\left[M_n \le\ p_{(i_*)} < \frac{1+o(1)}{\sqrt{(2+\delta)\log\log n}\,(\log n)^{1+\delta/2}}\right] \xrightarrow{n\to\infty} 1\,.$$

Setting $\epsilon = \delta/2$ concludes the proof. □

B.3. Proof of Theorem 5.1

Lemma B.3. *The null distributions of M_n^+ and M_n^- are identical.*

Proof. Assume w.l.o.g. that $U[0,1]$ is the null distribution and let $U_1,\ldots,U_n$ be distributed according to the null. It is easy to show that $M_n^+(U_1,\ldots,U_n) = M_n^-(1-U_1,\ldots,1-U_n)$. The claim follows from the fact that the vectors $(U_1,\ldots,U_n)$ and $(1-U_1,\ldots,1-U_n)$ have the same distribution. □

The right inequality in Eq. (5.3) follows directly from the union bound and the last lemma.

$$\Pr[M_n \le c \mid \mathcal{H}_0] \le \Pr[M_n^+ \le c \mid \mathcal{H}_0] + \Pr[M_n^- \le c \mid \mathcal{H}_0] = 2q_c\,.$$

We now prove the left inequality in Eq. (5.3). Let $U_1,\ldots,U_n \overset{i.i.d.}{\sim} U[0,1]$ and denote the joint density of their order statistics by $\mathbf{U} = (U_{(1)},\ldots,U_{(n)})$. Denote the events $M_n^+ > c$ and $M_n^- > c$ by A and B respectively. According to proposition 3.11 from Karlin and Rinott (1980), the random vector $\mathbf{U}$ is multivariate totally positive of order 2. It is easy to show that the indicator functions $\mathbf{1}_A(\mathbf{U})$ and $\mathbf{1}_{B^c}(\mathbf{U})$ are monotone-increasing in $\mathbb{R}^n$, hence by Karlin and Rinott (1980, Theorem 4.2), we have

$$\mathbb{E}[\mathbf{1}_A(\mathbf{X})\mathbf{1}_{B^c}(\mathbf{X})] \ge \mathbb{E}[\mathbf{1}_A(\mathbf{X})]\mathbb{E}[\mathbf{1}_{B^c}(\mathbf{X})].$$

Equivalently, $\Pr[A \wedge B^c] \ge \Pr[A]\cdot\Pr[B^c]$. Therefore,

$$\begin{aligned}\Pr[M_n > c] &= \Pr[A\wedge B] = \Pr[A] - \Pr[A\wedge B^c]\\ &\le \Pr[A] - \Pr[A]\Pr[B^c] = \Pr[A]\Pr[B] = (1-q_c)^2.\end{aligned}$$

From this follows the left inequality of Eq. (5.3). Finally, Eq. (5.4) follows from the asymptotic distributions of M_n^+, M_n^- and M_n given in Theorem 4.1. □

Appendix C: One-sided p-value computation

Let $x_1,\ldots,x_n$ be n observations with a one-sided value $M_n^+(x_1,\ldots,x_n) = c$. A direct approach to compute the corresponding p-value is to recursively evaluate the $n-1$ integrals in Eq. (5.2)

$$f_0(t) = 1, \quad f_1(t) = \int_{L_1}^{t} f_0(x)dx, \quad \ldots, \quad f_n(t) = \int_{L_n}^{t} f_{n-1}(x)dx, \tag{C.1}$$

where for notational simplicity we use the shorthand L_i for $L_i^n(c)$. The p-value of the M_n^+ test is then given by

$$\Pr\left[M_n^+ < c \,\middle|\, \mathcal{H}_0\right] = 1 - n!f_n(1)\,. \tag{C.2}$$

By definition, the various functions f_d in Eq. (C.1) are polynomials of increasing degree, $f_d(x) = \sum_{k=0}^{d} c_{d,k}x^k$, whose coefficients $c_{d,0},\ldots,c_{d,d}$ are sums of various products of $L_1,\ldots,L_d$. The second column of Table 1 lists explicit symbolic

TABLE 1
Comparison of symbolic expressions for $f_n(1)$ resulting from direct integration vs. computation using translated polynomials. L_i is shorthand for $L_i^n(c)$.

n	straightforward integration	translated polynomials
1	$1 - L_1$	$1 - L_1$
2	$\frac{1}{2} - L_1 - \frac{1}{2}L_2^2 + L_1L_2$	$\frac{1}{2}(1-L_1)^2 - \frac{1}{2}(L_2 - L_1)^2$
3	$\frac{1}{6} - \frac{1}{2}L_1 - \frac{1}{2}L_2^2 + L_1L_2 - \frac{1}{6}L_3^3$ $-\frac{1}{2}L_1L_3^2 - \frac{1}{2}L_2^2L_3 + L_1L_2L_3$	$\frac{1}{6}(1-L_1)^3 - \frac{1}{2}(L_2-L_1)^2(1-L_3)$ $-\frac{1}{6}(L_3-L_1)^3$
4	$\frac{1}{24} - \frac{1}{6}L_1 + (\frac{1}{2}L_1L_2 - \frac{1}{4}L_2^2)$ $-L_1L_2L_3 + \frac{1}{2}L_2^2L_3 + \frac{1}{2}L_1L_3^2$ $-\frac{1}{6}L_3^3 + L_1L_2L_3L_4 - \frac{1}{2}L_2^2L_3L_4$ $-\frac{1}{2}L_1L_3^2L_4 + \frac{1}{6}L_3^3L_4 - \frac{1}{2}L_1L_2L_4^2$ $+\frac{1}{4}L_2^2L_4^2 + \frac{1}{6}L_1L_4^3 - \frac{1}{24}L_4^4$	$\frac{1}{24}(1-L_1)^4 - \frac{1}{4}(L_2-L_1)^2(1-L_3)^2$ $-\frac{1}{6}(L_3-L_1)^3(1-L_4) - \frac{1}{24}(L_4-L_1)^4$ $+\frac{1}{4}(L_2-L_1)^2(L_4-L_3)^2$

expressions for the resulting $f_n(1)$ for small values of n. Clearly, the number of terms in the exact symbolic representation grows rapidly with n, and unfortunately we have not found a simple closed-form formula for its coefficients. Nonetheless, one can iteratively evaluate the coefficients of the polynomials $\{f_d\}_{d=1}^n$ numerically, since

$$f_d(t) = \int_{L_d}^{t} f_{d-1}(x)dx = \int_{L_d}^{t} \sum_{k=0}^{d-1} c_{d-1,k}x^k dx = \sum_{k=1}^{d} \frac{c_{d-1,k-1}}{k} t^k - \sum_{k=1}^{d} \frac{c_{d-1,k-1}}{k} L_d^k .$$

Thus, at each iteration we store the numerical values of $c_{d,0}, \ldots, c_{d,d}$ and update them according to the following formula

$$c_{d,0} = -\sum_{k=1}^{d} \frac{c_{d-1,k-1}}{k} L_d^k \quad \text{and} \quad \forall k \geq 1 : c_{d,k} = \frac{c_{d-1,k-1}}{k} . \tag{C.3}$$

While seemingly straightforward to evaluate, a naïve implementation using standard (80-bit) long double floating-point accuracy suffers from a fast accumulation of numerical errors and breaks down completely at $n \approx 150$. The heart of the problem is the formula for the constant term $c_{d,0}$ of f_d. As seen from Eq. (C.3), at each iteration the term $c_{d+1,0}$ accumulates errors from all previous coefficients $\{c_{d,j}\}_{j=0}^d$. These errors propagate to the higher order coefficients in the next iteration, and are again amplified when computing $c_{d+2,0}$, etc.

Computation using translated polynomials. To attenuate the accumulation of numerical errors we perform the calculations in a different basis for the space of degree d polynomials. Instead of the standard basis, for each degree d we use a basis of translated monomials $(x + t_{d,k})^k$, where the constants $t_{d,k}$ are yet to be determined,

$$f_d(x) = c_{d,0} + \sum_{k=1}^{d} c_{d,k} \left(x + t_{d,k}\right)^k . \tag{C.4}$$

As in (C.1), $f_0(t) = 1$, which is represented as $c_{0,0} = 1$. Using the representation (C.4), each integration step yields

$$f_d(t) = \int_{L_d}^{t} f_{d-1}(x)dx = c_{d-1,0}(t - L_d) + \sum_{k=2}^{d} \frac{c_{d-1,k-1}}{k} \left(t + t_{d-1,k-1}\right)^k - \sum_{k=2}^{d} \frac{c_{d-1,k-1}}{k} \left(L_d + t_{d-1,k-1}\right)^k .$$

Given the above form we define $t_{d,k}$ as follows,

$$t_{d,1} = -L_d, \quad \text{and} \quad t_{d,k} = t_{d-1,k-1} \quad \forall k \in \{2, \ldots, d\}.$$

Then, the coefficients of the translated polynomial f_d satisfy

$$c_{d,0} = -\sum_{k=2}^{d} \frac{c_{d-1,k-1}}{k} \left(L_d + t_{d-1,k-1}\right)^k \quad \text{and} \quad c_{d,k} = \frac{c_{d-1,k-1}}{k} \quad \text{for } k = 1, \ldots, d. \tag{C.5}$$

In contrast to (C.3), in this update rule the constant term $c_{d,0}$ does not depend on the term $c_{d-1,0}$ of the previous iteration. Thus, the error accumulation in this recursion is slower than in (C.3), and empirically, the update rule (C.5) is significantly more stable. In summary, numerical integration of (C.5), using extended double-precision (80-bit), allows accurate calculation of one-sided p-values for up to $n \approx 50,000$ samples. C++ source code for this procedure is freely available at `http://www.wisdom.weizmann.ac.il/~amitmo`.

Acknowledgements

The authors thank Yoav Benjamini, Ya'acov Ritov, Art Owen, Jiashun Jin, Guenther Walther and Jonathan Rosenblatt for interesting discussions. This work was supported by a Texas A&M-Weizmann research grant from Paul and Tina Gardner, and by a grant from the Citi Foundation.

Supplementary Material

Supplementary to "On the exact Berk-Jones statistics and their p-value calculation"
(doi: 10.1214/16-EJS1172SUPP; .pdf).

References

Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W. and Stine, R. A. (2014). The power to see: a new graphical test of normality (correction). *Amer. Statist.* **68** 318. MR3280623

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statistics* **23** 193–212. MR0050238

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.* **49** 765–769. MR0069459

Barnett, I. J. and Lin, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika* **101** 964–970. MR3286929

Berk, R. H. and Jones, D. H. (1978). Relatively optimal combinations of test statistics. *Scand. J. Statist.* **5** 158–162. MR509452

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. MR521531

Brown, J. R. and Harvey, M. E. (2008a). Arbitrary Precision Mathematica Functions to Evaluate the One-Sided One Sample K-S Cumulative Sampling Distribution. *Journal of Statistical Software* **26** 1–55.

Brown, J. R. and Harvey, M. E. (2008b). Rational Arithmetic Mathematica Functions to Evaluate the Two-Sided One Sample K-S Cumulative Sampling Distribution. *Journal of Statistical Software* **26** 1–40.

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

Calitz, F. (1987). An alternative to the Kolmogorov-Smirnov test for goodness of fit. *Comm. Statist. Theory Methods* **16** 3519–3534. MR930353

Csörgő, M., Csörgő, S., Horváth, L. and Mason, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85. MR815960

Daniels, H. E. (1945). The statistical theory of the strength of bundles of threads. I. *Proc. Roy. Soc. London. Ser. A.* **183** 405–435. MR0012388

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences* **105** 14790–14795.

Duembgen, L. and Wellner, J. A. (2014). Confidence Bands for Distribution Functions: A New Look at the Law of the Iterated Logarithm. *ArXiv e-prints*.

Durbin, J. (1973). *Distribution theory for tests based on the sample distribution function*. Society for Industrial and Applied Mathematics, Philadelphia, Pa. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 9. MR0305507

Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *Ann. Statist.* **7** 116–138. MR515688

Friedrich, T. and Schellhaas, H. (1998). Computation of the percentage points and the power for the two-sided Kolmogorov-Smirnov one sample test. *Statist. Papers* **39** 361–375. MR1649359

Gontscharuk, V. and Finner, H. (2016). Asymptotics of goodness-of-fit tests based on minimum P-value statistics. *Communications in Statistics - Theory and Methods*.

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

Gontscharuk, V., Landwehr, S. and Finner, H. (2016). Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli* **22** 1331–1363. MR3474818

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

Jaeschke, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.* **7** 108–115. MR515687

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962

Janssen, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* **28** 239–253. MR1762910

Kaplan, D. M. and Goldman, M. (2014). True equality (of pointwise sensitivity) at last: a Dirichlet alternative to Kolmogorov–Smirnov inference on distributions. Technical Report.

Karlin, S. and Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10** 467–498. MR599685

Keilson, J. and Sumita, U. (1983). A decomposition of the beta distribution, related order and asymptotic behavior. *Ann. Inst. Statist. Math.* **35** 243–253. MR716035

Khmaladze, E. and Shinjikashvili, E. (2001). Calculation of noncrossing probabilities for Poisson processes and its corollaries. *Adv. in Appl. Probab.* **33** 702–716. MR1860097

Kotel′nikova, V. F. and Khmaladze, È. V. (1982). Calculation of the probability of an empirical process not crossing a curvilinear boundary. *Teor. Veroyatnost. i Primenen.* **27** 599–607. MR673936

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000–1005. MR1294744

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*, third ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

Li, J. and Siegmund, D. (2015). Higher criticism: p-values and criticism. *Ann. Statist.* **43** 1323–1350. MR3346705

Marsaglia, G., Tsang, W. W. and Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software* **8** 1–4.

Mary, D. and Ferrari, A. (2014). A non-asymptotic standardization of binomial counts in Higher Criticism. In *Information Theory (ISIT), 2014 IEEE International Symposium on* 561–565. IEEE.

Mason, D. M. and Schuenemeyer, J. H. (1983). A modified Kolmogorov-Smirnov test sensitive to tail alternatives. *Ann. Statist.* **11** 933–946. MR707943

Moscovich, A. and Nadler, B. (2016). Fast calculation of boundary crossing probabilities for Poisson processes. *ArXiv e-prints*.

Neyman, J. (1937). Smooth test for goodness of fit. *Skand. Aktuarie Tidskv.* **20** 149-199.

Noé, M. (1972). The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics. *Ann. Math. Statist.* **43** 58–64. MR0300379

Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function. *J. Amer. Statist. Assoc.* **90** 516–521. MR1340504

Peizer, D. B. and Pratt, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities. I. *J. Amer. Statist. Assoc* **63** 1416–1456. MR0235650

Pratt, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities. II. *J. Amer. Statist. Assoc.* **63** 1457–1483. MR0235651

Rainer, J. C. W., Thas, O. and Best, D. J. (2009). *Smooth Tests of Goodness of Fit Using R*, 2nd ed. Wiley.

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. Inst. Math. Statist., Beachwood, OH. MR3202643