The Annals of Statistics
2018, Vol. 46, No. 6A, 3014–3037
https://doi.org/10.1214/17-AOS1647

TWO-SAMPLE KOLMOGOROV–SMIRNOV-TYPE TESTS REVISITED: OLD AND NEW TESTS IN TERMS OF LOCAL LEVELS[1]

BY HELMUT FINNER* AND VERONIKA GONTSCHARUK†

*Institute for Biometrics and Epidemiology, German Diabetes Center** *and Institute for Health Services Research and Health Economics, German Diabetes Center*†

From a multiple testing viewpoint, Kolmogorov–Smirnov (KS)-type tests are union-intersection tests which can be redefined in terms of local levels. The local level perspective offers a new viewpoint on ranges of sensitivity of KS-type tests and the design of new tests. We study the finite and asymptotic local level behavior of weighted KS tests which are either tail, intermediate or central sensitive. Furthermore, we provide new tests with approximately equal local levels and prove that the asymptotics of such tests with sample sizes m and n coincides with the asymptotics of one-sample higher criticism tests with sample size $\min(m, n)$. We compare the overall power of various tests and introduce local powers that are in line with local levels. Finally, suitably parameterized local level shape functions can be used to design new tests. We illustrate how to combine tests with different sensitivity in terms of local levels.

1. Introduction. Local levels of one-sample goodness-of-fit (GOF) tests were introduced in [17] (also cf. [16]) in order to yield a better understanding of the asymptotic and finite behavior of higher criticism (HC) statistics, among others, in connection with sparse signals and detectability. The main focus in [17] is on the local level behavior of the original HC statistic introduced in [9]. It is shown in [17] that local levels of the original HC test are almost all asymptotically equal. One-sample GOF tests defined in terms of equal local levels are studied extensively in [15] and [16]. It is also indicated in [17] that local levels may serve as a useful tool for designing new GOF tests. In this paper, we adopt this idea and investigate a large class of (nonparametric) two-sample Kolmogorov–Smirnov (KS)-type tests in terms of local levels. KS tests have its origins in the pathbreaking papers [21], [26] and [27].

Received May 2016; revised September 2017.

[1]Supported by the Ministry of Science and Research of the State of North Rhine-Westphalia (MIWF NRW) and the German Federal Ministry of Health (BMG).

MSC2010 subject classifications. Primary 62G10, 62G30; secondary 62G20, 60F99.

Key words and phrases. Goodness-of-fit, higher criticism test, local levels, multiple hypotheses testing, nonparametric two-sample tests, order statistics, weighted Brownian bridge.

Let F and G denote two continuous (unknown) cumulative distribution functions (CDFs) on $\mathbb{R}$ and let $X_1, \ldots, X_m \sim F$ and $Y_1, \ldots, Y_n \sim G$ denote two independent i.i.d. samples. Two-sample KS-type tests for

$$H^{=} : F = G \quad \text{or} \quad H^{\geq} : F \geq G$$

rely on the difference of the corresponding empirical CDFs (ECDFs) $\hat{F}_m$ and $\hat{G}_n$. Consider the local hypotheses $H_t^{=} : F(t) = G(t)$ and $H_t^{\geq} : F(t) \geq G(t)$ for $t \in \mathbb{R}$. Then

$$H^{=} = \bigcap_{t \in \mathbb{R}} H_t^{=} \quad \text{and} \quad H^{\geq} = \bigcap_{t \in \mathbb{R}} H_t^{\geq}.$$

In what follows, H_t denotes either $H_t^{=}$ or $H_t^{\geq}$. We restrict attention to tests related to the union-intersection principle with local test statistics $T_t = T_t(\hat{F}_m(t), \hat{G}_n(t))$ for testing H_t, $t \in \mathbb{R}$. The global null hypothesis is rejected if H_t is rejected for at least one $t \in \mathbb{R}$. Once suitable test statistics are defined, we can define local levels α_t (say) as the rejection probability of the local test for H_t under the global null hypothesis $H^{=}$. Unfortunately, the local levels α_t depend on the value of $F(t)$ under $H^{=}$. In order to obtain distribution-free local levels, a key step is to redefine KS-type tests in terms of conditional tests related to 2×2 table tests. Note that $m\hat{F}_m(t) \sim \text{Binl}(m, F(t))$ and $n\hat{G}_n(t) \sim \text{Binl}(n, G(t))$, where $\text{Binl}(N, p)$ denotes the binomial distribution with parameters N and p. Hence, each H_t can be tackled by some (conditional, unconditional, exact or asymptotic) test developed in the area of two-sample binomial testing problems. The conditional point of view allows us to define appropriate (conditional) distribution-free local levels α_s given that $(m+n)\hat{H}_{m+n}(t) = s$, $s \in \{1, \ldots, m+n\}$, where $\hat{H}_{m+n}$ denotes the ECDF of the combined sample. The local levels α_s can be computed in terms of the underlying hypergeometric distributions.

On the one hand, local levels can be viewed as an interesting characteristic of union-intersection related GOF tests indicating in which area we can expect high or low sensitivity. On the other hand, we can design new GOF tests by choosing suitable local levels reflecting our wishes concerning the sensitivity in specific areas. Clearly, larger local levels result in larger local power. We study local levels of well established KS-type tests, especially weighted KS tests, and show how to design new tests in terms of local levels or local level shape functions. A further focus is on the asymptotics of local levels. Thereby, it depends heavily on the relation between the sample sizes m and n whether or not the asymptotics reflects the finite local level behavior. For example, for m close to n, the local level behavior of the two-sample HC statistics differs drastically from the local level behavior of one-sample HC statistics. On the other hand, unequal sample sizes may result in undesirable local level and power behavior of some tests. The overall power behavior of two-sample KS-type tests also depends on the relation between the sample sizes m and n and may lead to weird effects. For example, if we increase

one sample size, the power may decrease, or, if we exchange m and n, the power may change drastically. We will illustrate that a look at the underlying local levels helps to explain and to avoid such phenomena.

The paper is organized as follows. In Section 2, we introduce and review basic concepts and issues including a hypergeometric perspective, an important inherent structural property of GOF tests (called *proper*) related to Barnard-convexity, a formal look at local levels and local level shape functions. In Section 3, we consider the class of weighted KS two-sample GOF tests in more detail. In Section 3.1, we show that a large class of weight functions leads to proper two-sample GOF tests and illustrate their local level behavior for the well-known weight functions $w(t) = (t(1-t))^{\nu}$, $t \in (0,1)$, $\nu \in [0,1]$. In Section 3.2, we derive the asymptotics of local levels related to weighted KS statistics of this type. We distinguish three cases: (i) $\nu \in [0, 0.5)$, (ii) $\nu = 0.5$ and (iii) $\nu \in (0.5, 1]$, which lead to different types of asymptotic distributions and different local level behavior. Section 4 is concerned with two-sample GOF tests with (approximately) equal local levels. In Section 4.1, we consider two-sample minimum p-value (minP) tests and study their finite properties, and in Section 4.2, we derive the minP asymptotics. Section 5 provides some power considerations. In Section 5.1, we compare GOF tests from Sections 3 and 4 with respect to overall power. By means of numerical simulations, we give some hints which of the tests considered here yield a good overall performance and which of them are most likely to beat the original KS test. Some thoughts on local power are outlined in Section 5.2. Section 6 provides some concluding remarks. Among others, we discuss the possibility to construct new GOF tests by combining local levels of different tests. As an example, we consider two combinations of KS and minP tests. Proofs are deferred to Supplement A, that is, [11]. In Supplement B, that is, [12], we provide some animated graphics in order to illustrate the local level behavior of various KS-type tests.

2. Two-sample tests revisited.

2.1. *A hypergeometric perspective.* Setting $S_{m+n}(t) = m\hat{F}_m(t) + n\hat{G}_n(t)$, the ECDF of the combined sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is given by

$$\hat{H}_{m+n}(t) = \frac{1}{m+n} S_{m+n}(t), \qquad t \in \mathbb{R}.$$

Without loss of generality we assume that the ordered jump points t_s (say) of $\hat{H}_{m+n}$ satisfy $t_1 < \cdots < t_{m+n}$. Note that $S_{m+n}(t_s) = s$. Let $V_{m,s}$ denote the number of ranks related to the first sample being not larger than s, that is,

$$V_{m,s} = m\hat{F}_m(t_s), \qquad s = 1, \ldots, m+n.$$

Since $V_{m,m+n} = m$ for any $m, n \in \mathbb{N}$, we restrict attention to $s \in I_{m,n}$, where

$$I_{m,n} \equiv \{1, \ldots, m+n-1\}.$$

Given that $H^=$ is true, $V_{m,s}$ follows a hypergeometric distribution with probability mass function

$$f(x|s,m,n) = \binom{m}{x}\binom{n}{s-x}\Big/\binom{m+n}{s}, \qquad \max(0,s-n) \le x \le \min(s,m).$$

The related CDF is denoted by $F_{\mathrm{Hyp}}(\cdot|s,m,n)$. Below, $\mathbb{P}_0$ and $\mathbb{E}_0$ denote the probability measure and expectation under the global null hypothesis $H^=$.

The random vector $V_m \equiv (V_{m,s} : s \in I_{m,n})$ contains all the information about the ranks of both samples. Moreover,

$$\hat{G}_n(t_s) - \hat{F}_m(t_s) = \frac{m+n}{mn}\big(\mathbb{E}_0[V_{m,s}] - V_{m,s}\big), \qquad \mathbb{E}_0[V_{m,s}] = \frac{ms}{m+n}, \tag{2.1}$$

that is, test statistics in terms of $\hat{G}_n - \hat{F}_m$ can be rewritten in terms of V_m; see also [28]. In what follows, local test statistics $V_{m,s}$, $s \in I_{m,n}$, play a key role.

2.2. *Proper two-sample GOF tests.* In this paper, we restrict attention to GOF tests with acceptance regions for V_m of the form

$$A_{m,n} = \big\{x \in \mathbb{N}_0^{m+n-1} : c_s \le x_s \le d_s, s \in I_{m,n}\big\}. \tag{2.2}$$

In the one-sided case, we assume that the upper critical values are given by $d_s = \min(s,m)$, $s \in I_{m,n}$. It can easily be shown that for any acceptance region $A_{m,n}$ with $\mathbb{P}_0(V_m \in A_{m,n}) > 0$ there exists a unique acceptance region $\widetilde{A}_{m,n} \subseteq A_{m,n}$ of the form (2.2) with critical values $\tilde{c}_s, \tilde{d}_s \in \{0,\dots,m\}$, $s \in I_{m,n}$, satisfying $\mathbb{P}_0(V_m \in A_{m,n}) = \mathbb{P}_0(V_m \in \widetilde{A}_{m,n})$ and

$$\max(0,s-n) \le \tilde{c}_s \le \tilde{d}_s \le \min(s,m), \qquad s \in I_{m,n}, \tag{2.3}$$

$$c_s \le \tilde{c}_s \le \tilde{d}_s \le d_s, \qquad s \in I_{m,n}, \tag{2.4}$$

$$\tilde{c}_{s+1} \in \{\tilde{c}_s, \tilde{c}_s+1\} \quad \text{and} \quad \tilde{d}_{s+1} \in \{\tilde{d}_s, \tilde{d}_s+1\}, \qquad s = 1,\dots,m+n-2. \tag{2.5}$$

The latter property is a consequence of $V_{m,s+1} \in \{V_{m,s}, V_{m,s}+1\}$ for $s \in I_{m,n}$. We denote critical values satisfying (2.3) and (2.5) as *proper* critical values and the corresponding tests as *proper* GOF tests. Typically, acceptance regions are defined in terms of a test statistic $M = \max_{s\in I_{m,n}} M_s(V_{m,s})$ or $M = \min_{s\in I_{m,n}} M_s(V_{m,s}))$ with local test statistics M_s. Without loss of generality let $A_{m,n} = \{M \le c\}$. Then M is said to be proper if for all c with $\mathbb{P}_0(M \le c) > 0$ there exist proper critical values c_s, d_s, such that for all $s \in I_{m,n}$ and $\max(0,s-n) \le x \le \min(s,m)$ it holds

$$M_s(x) \le c \quad \text{iff} \quad c_s \le x \le d_s.$$

REMARK 2.1 (Proper GOF tests versus 2×2-table tests). In the area of 2×2-table tests, (2.5) is often referred to as Barnard-convexity according to Barnard's ideas in [3], for example, cf. the discussion in [13]. Any Barnard-convex unconditional 2×2-table test for the comparison of two independent binomial samples

yields a proper GOF test and vice versa. However, given an acceptance region $A_{m,n}$ with proper critical values, the global level of the corresponding proper GOF test and the unconditional level of the resulting 2×2-table test are two different things.

Typically, the effective global level $\mathbb{P}_0(V_m \notin A_{m,n})$ is smaller than the prespecified α, that is, two-sample GOF tests are not α-exhaustive. This is a general issue with discrete distributions. Especially, in case $m = n$, KS-type tests can be rather conservative. For the computation of the global level of tests with acceptance regions (2.2), we refer to Section A2 in [11], which provides a recursive algorithm based on properties of V_m. We also refer to [18] for further formulas and interesting discussions around the overall significance level of the two-sample KS test.

Finally, it may be worth to mention that all proper two-sample GOF tests can be complemented with simultaneous confidence bands for a shift function $\triangle$ defined by $G(x + \triangle(x)) = F(x)$, $x \in \mathbb{R}$, by applying the method of Doksum and Sievers; cf. [8].

2.3. *Local levels of two-sample GOF tests.* As mentioned in the Introduction, we define (conditional) local levels as probabilities under $H^=$ to reject local hypotheses H_t given $S_{m+n}(t) = s$, $s \in I_{m,n}$. More precisely, lower and upper local levels of a proper two-sample GOF test with acceptance region of the form (2.2) are defined by

$$\alpha_s^{\text{low}} = \mathbb{P}_0(V_{m,s} < c_s) \quad \text{and} \quad \alpha_s^{\text{up}} = \mathbb{P}_0(V_{m,s} > d_s), \qquad s \in I_{m,n},$$

respectively. The corresponding two-sided local levels are given by

$$\alpha_s = \alpha_s^{\text{low}} + \alpha_s^{\text{up}}, \qquad s \in I_{m,n}.$$

It is important to note that these local levels do not depend on t, F and G and can easily be computed in terms of the underlying hypergeometric distributions. Since nonproper critical values may yield artificial small local levels, we consider local levels based on proper critical values only.

Obviously, local levels of proper GOF tests are bounded by the global level, that is, $\alpha_s \le \mathbb{P}_0(V_{m,n} \notin A_{m,n})$, $s \in I_{m,n}$. Many GOF tests possess additional symmetry properties which result in corresponding symmetry properties of local levels. For example, if $d_s = m - c_{m+n-s}$, then $\alpha_s^{\text{low}} = \alpha_{m+n-s}^{\text{up}}$. If the latter property holds for all $s \in I_{m,n}$, then $\alpha_s = \alpha_{m+n-s}$ for all $s \in I_{m,n}$. Moreover, if $m = n$ and $c_s = s - m + c_{2m-s}$, then $\alpha_s^{\text{low}} = \alpha_{2m-s}^{\text{low}}$, $s \in I_{m,m}$.

Figure 1 illustrates local levels of the two-sample two-sided original KS test for $\alpha = 0.05$, $m = 20$ and $n = 80$. Thereby, one-sided local levels fulfill $\alpha_s^{\text{low}} = \alpha_{m+n-s}^{\text{up}}$, two-sided local levels are symmetric in $s \in I_{m,n}$ and the effective level is 0.0445. KS-type statistics are studied in more detail in Section 3.

In order to get a feeling for the behavior of local levels of various proper GOF tests, asymptotic considerations for $m, n \to \infty$ turn out to be helpful. As in [16]

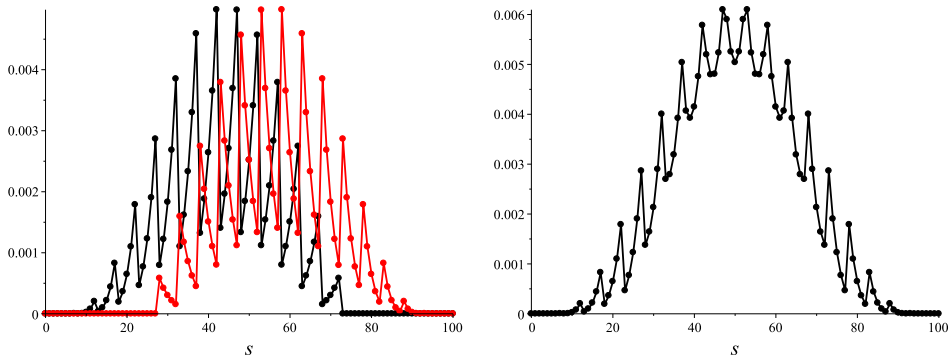

FIG. 1. *Lower* (*red points in the left graph*), *upper* (*black points in the left graph*) *and two-sided* (*right graph*) *local levels related to the level* α *two-sided KS test with* $\alpha = 0.05$, $m = 20$ *and* $n = 80$. *Thereby, the KS critical value is* $b = 1.3$.

and [17], we distinguish three types of s-values w.r.t. their asymptotics, that is, (a) $s/(m+n) \to \eta \in (0,1)$ (central range), (b) $s/(m+n) \to \eta \in \{0,1\}$ with $\min\{s, m+n-s\} \to \infty$ (intermediate range), (c) $s/(m+n) \to \eta \in \{0,1\}$ with $\min\{s, m+n-s\}$ fixed (extreme range). Thereby, we are interested in sets of s-values leading to asymptotic exhaustion of the overall level α, that is, we are looking for small sets $J_{m,n} \subset I_{m,n}$ such that $\mathbb{P}_0(\bigcup_{s\in J_{m,n}}\{V_{m,s} \notin [c_s, d_c]\}) \to \alpha$ as $m, n \to \infty$. We refer to such sets loosely as *sensitivity ranges*.

2.4. *Local level shape functions*. Noting that a given set of local levels yields a corresponding set of critical values, one may design two-sample GOF tests in terms of so-called *local level shape functions* (LLSFs). First, we restrict attention to lower LLSFs $\tilde{\alpha}_\kappa^{\text{low}} : [0,1] \to [0,1]$, which are assumed to be monotone in the tuning parameter κ. Then the largest critical values c_s leading to $\alpha_s^{\text{low}} \le \tilde{\alpha}_\kappa^{\text{low}}(s/(m+n))$, $s \in I_{m,n}$, define a (one-sided) GOF test of the form (2.2). In order to get a level α test, we choose κ such that α is maximally exhausted. Similarly, upper critical values can be defined by an upper LLSF $\tilde{\alpha}_\kappa^{\text{up}}$. For two-sided tests we may choose LLSFs with $\tilde{\alpha}_\kappa^{\text{up}}(\eta) = \tilde{\alpha}_\kappa^{\text{low}}(1-\eta)$ for $\eta \in [0,1]$ leading to $d_s = m - c_{m+n-s}$ for $s \in I_{m,n}$. In order to get more symmetry, we may choose $\tilde{\alpha}_\kappa^{\text{up}} = \tilde{\alpha}_\kappa^{\text{low}} = \tilde{\alpha}_\kappa$ for some symmetric LLSF $\tilde{\alpha}_\kappa$.

Typically, asymptotic local levels of conventional KS-type tests yield an asymptotic LLSFs; cf. Section 3. For example, classical KS tests result in the asymptotic (symmetric) LLSF $\tilde{\alpha}_\kappa(\eta) = \Phi(-\kappa/\sqrt{\eta(1-\eta)})$ which yields a neat reflection of the local levels of KS tests for larger sample sizes; cf. the animation in Figure B1 in [12]. LLSFs may be viewed as a tool for designing the sensitivity of GOF tests. In this paper, we mainly restrict attention to local levels and the resulting LLSFs of special weighted KS-type test statistics.

3. Weighted two-sample KS tests. In this section, we are concerned with weighted two-sample KS tests of the form

$$\sup_{t\in\mathbb{R}} \sqrt{\frac{mn}{m+n}} \frac{\langle \hat{G}_n(t) - \hat{F}_m(t)\rangle}{w(\hat{H}_{m+n}(t))}, \tag{3.1}$$

where $w : (0, 1) \to \mathbb{R}^+$ is a nonnegative continuous weight function. The notation $\langle\cdot\rangle$ indicates either the one- or two-sided test statistic, that is, $\langle a\rangle$ is either a or $|a|$, respectively. Clearly, $w(t) \equiv 1$ leads to the classical two-sample KS statistics and $w(t) = \sqrt{t(1-t)}$ yields the supremum version of the two-sample Anderson–Darling statistic. During the past decade, the one-sample supremum version of Anderson–Darling statistics has gained a lot of attention as a higher criticism (HC) statistic, for example, cf. [9] and [10]. Therefore, we refer to two-sample weighted KS tests based on $w(t) = \sqrt{t(1-t)}$ as HC tests, too. Although, the classical one-sample HC approach leads to a series of innovative asymptotic results, especially with respect to sparsity and detectability, the statistic itself has serious drawbacks. This was already indicated by Canner in [6] by looking at the critical values of two-sided HC tests. He found that "These results are rather shocking ...". On the other hand, Canner found that the two-sample statistic is less problematic.

3.1. *Finite considerations and local levels.* Since extrema of the difference of two ECDFs are taken in jump points of the combined ECDF $\hat{H}_{m+n}$ and $\hat{H}_{m+n}(t_s) = s/(m+n)$, a statistic defined in (3.1) is almost surely equal to $\mathrm{KS}^{w,\langle\rangle}_{m,n} = \sup_{s\in I_{m,n}} \mathrm{KS}^{w,\langle\rangle}_{m,n,s}$ with

$$\mathrm{KS}^{w,\langle\rangle}_{m,n,s} = \sqrt{\frac{m+n}{mn}} \frac{\langle (sm)/(m+n) - V_{m,s}\rangle}{w(s/(m+n))}. \tag{3.2}$$

The null hypothesis is rejected if $\mathrm{KS}^{w,\langle\rangle}_{m,n} > b$ for some $b > 0$. Obviously, weighted KS tests have acceptance regions of the form (2.2).

LEMMA 3.1. *If the weight function w is continuous and concave, then the weighted KS statistic $\mathrm{KS}^{w,\langle\rangle}_{m,n}$ is proper.*

Formulas for proper critical values can be found in the proof of Lemma 3.1 in [11]. If the weight function is symmetric, we get $d_s = m - c_{m+n-s}$, $s \in I_{m,n}$, and, in case of $m = n$, $c_s = s - m + c_{2m-s}$, $s \in I_{m,m}$. In what follows, we restrict attention to the following symmetric, continuous and concave weight functions

$$w_\nu(t) = \big(t(1-t)\big)^\nu, \qquad \nu \in [0, 1]. \tag{3.3}$$

The corresponding weighted KS statistics can be represented as

$$\mathrm{KS}^{\nu,\langle\rangle}_{m,n} = \sup_{s\in I_{m,n}} \sqrt{\frac{m+n}{m+n-1}} \left(\frac{mn}{m+n-1}\right)^{\nu-0.5} \frac{\langle \mathbb{E}_0[V_{m,s}] - V_{m,s}\rangle}{(\mathbb{V}\mathrm{ar}_0[V_{m,s}])^\nu}, \tag{3.4}$$

where $\mathbb{V}\mathrm{ar}_0[V_{m,s}]$ denotes the variance of $V_{m,s}$ under $H^=$.

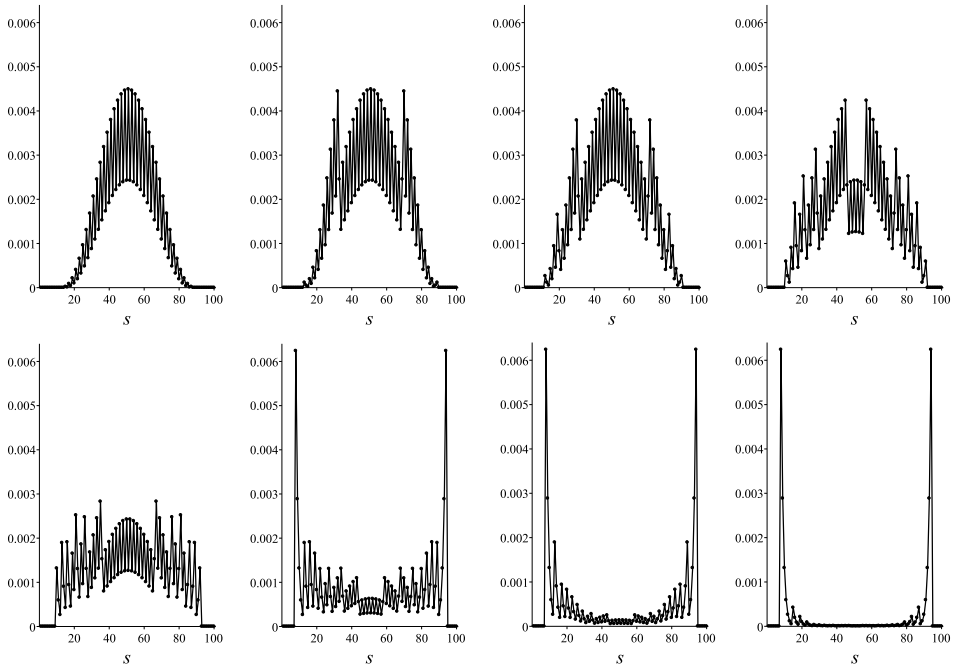

FIG. 2. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to level* α *two-sided weighted KS tests based on* (3.4) *and critical value* b *for* $m = n = 50$ *and* $\alpha = 0.05$. *First row*: $\nu = 0, 0.125, 0.25, 0.375$, $b = 1.3, 1.5699, 1.9116, 2.3602$, $\mathbb{P}_0(\mathrm{KS}_{m,n}^{\nu,||} > b) = 0.0392, 0.0487, 0.0494, 0.0495$; *second row*: $\nu = 0.5, 0.625, 0.75, 0.875$, $b = 2.9489, 3.846, 5.1844, 7.4265$, $\mathbb{P}_0(\mathrm{KS}_{m,n}^{\nu,||} > b) = 0.0453, 0.0466, 0.035, 0.0262$ (*from left to right in each row*).

Figure 2 illustrates the shape of lower local levels related to two-sided level α weighted KS tests for $m = n = 50$, $\alpha = 0.05$ and various ν-values. We observe that smaller ν-values lead to larger local levels in the central range, and larger ν-values lead to larger local levels in the tails. Figure 2 also illustrates that the actual global level $\mathbb{P}_0(\mathrm{KS}_{m,n}^{\nu,||} > b)$ may be much smaller than the prespecified level α for ν-values larger than (and not too close to) 0.5. This is due to the fact that we typically get discrete asymptotic distributions for $\nu \in (0.5, 1]$; see Theorem 3.3 and its discussion.

Figure 3 illustrates lower local levels for unequal sample sizes $m = 20$, $n = 80$ and $\alpha = 0.05$ and $\nu = 0.25, 0.5, 0.75$. For $\nu = 0$, we refer to Figure 1. We observe that lower local levels are only slightly asymmetric for $\nu = 0.0, 0.25$ and extremely asymmetric for $\nu = 0.5, 0.75$. Extremely asymmetric local levels have serious consequences with respect to power; cf. Section 5.

3.2. *Asymptotics of weighted two-sample KS tests*. In the one-sample case, the asymptotic behavior of weighted KS statistics with weight functions defined in

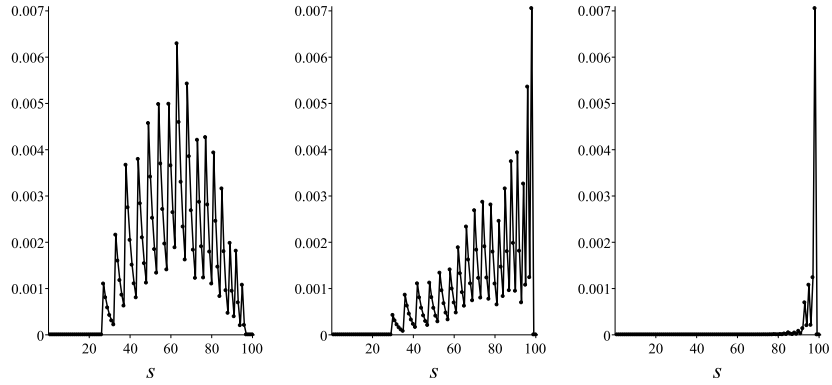

FIG. 3. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to level* α *two-sided weighted KS tests based on test statistic* (3.4) *and critical value* b *for* $m = 20$, $n = 80$, $\alpha = 0.05$ *and* $\nu = 0.25, 0.5, 0.75$ (*from left to right*). *Thereby*, $b = 1.93, 3.134, 7.64$ *and* $\mathbb{P}_0(\text{KS}_{m,n}^{\nu,||} > b) = 0.0489, 0.04727, 0.0158$, *respectively.*

(3.3) is closely related to the asymptotic behavior of weighted Brownian bridges and normalized Poisson processes. A summary of main results can be found in [19]. It therefore stands to reason that these results should carry over to the two-sample case. However, this is not the case for all ν-values. In order to derive asymptotic results, we consider the cases $\nu \in [0, 0.5)$, $\nu = 0.5$ and $\nu \in (0.5, 1]$ separately. Proofs of the asymptotic results are deferred to Section A3 in [11].

The case $\nu \in [0, 0.5)$. Let $\mathbb{B}$ denote a standard Brownian bridge on $[0, 1]$ and let the corresponding weighted Brownian bridges be defined by

$$\mathbb{B}^\nu(t) = \frac{\mathbb{B}(t)}{(t(1-t))^\nu}, \qquad t \in (0,1), \nu \in [0, 0.5].$$

The next theorem shows that the asymptotic behavior of weighted KS statistics with $\nu \in [0, 0.5)$ is the same as in the one-sample case.

THEOREM 3.1. *Let* $\nu \in [0, 0.5)$.

(a) *Under* $H^=$, *the test statistic* $\text{KS}_{m,n}^{\nu,\langle\rangle}$ *converges in distribution to* $\sup_{t\in(0,1)} \langle \mathbb{B}^\nu(t) \rangle$ *as* $m, n \to \infty$.

(b) *For* $s \in I_{m,n}$ *with* $\lim_{m,n\to\infty} s/(m+n) = \eta$ *for some* $\eta \in [0, 1]$, *lower local levels related to weighted* (*one- or two-sided*) *KS tests based on some critical value* $b \in \mathbb{R}$ *fulfill*

$$\begin{aligned} &\lim_{m,n\to\infty} \alpha_s^{\text{low}} = 0, \qquad \eta \in \{0, 1\}, \\ &\lim_{m,n\to\infty} \alpha_s^{\text{low}} = 1 - \Phi\big(b(\eta(1-\eta))^{\nu-0.5}\big), \qquad \eta \in (0, 1). \end{aligned} \tag{3.5}$$

(c) *The sensitivity range of weighted KS tests*, *that is*, *a range of* s*-values leading to the asymptotic exhaustion of the level* α, *coincides with the central range.*

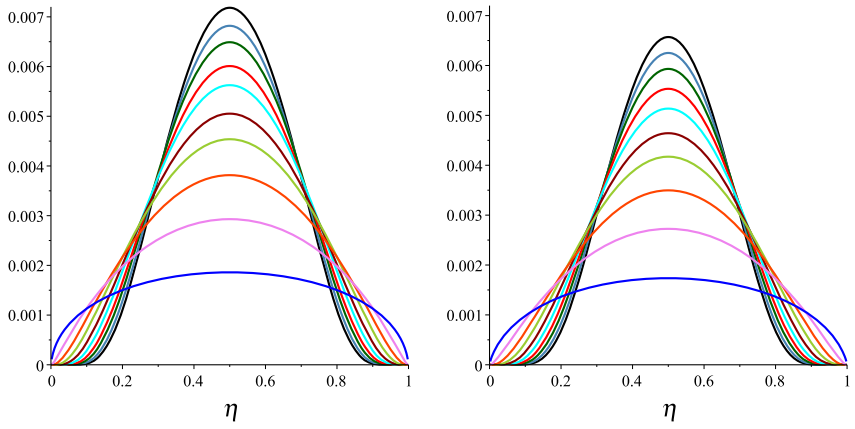

FIG. 4. *Asymptotic lower local levels* $\lim_{m,n\to\infty}\alpha_s^{\mathrm{low}}$ *related to asymptotic level* α *one-sided tests* (*left graph*) *and asymptotic two-sided local levels* $\lim_{m,n\to\infty}\alpha_s$ *related to asymptotic level* α *two-sided tests* (*right graph*) *based on* (3.4) *for* $\alpha = 0.05$ *and* $\nu = 0.0(0.05)0.45$ (*from top to bottom in* $\eta = 0.5$). *Thereby,* $b = 1.224, 1.322, 1.427, 1.546, 1.672, 1.819, 1.977, 2.167, 2.399, 2.707$ (*left graph*) *and* $b = 1.359, 1.465, 1.580, 1.708, 1.846, 2.002, 2.171, 2.372, 2.609, 2.922$ (*right graph*) *are simulated* $(1-\alpha)$*-quantiles related to* $\sup_{t\in(0,1)}\langle \mathbb{U}_\nu(t)\rangle$ *with* $\nu = 0.0(0.05)0.45$.

REMARK 3.1. To the best of our knowledge, for $\nu \in (0, 0.5)$, a manageable formula for the distribution of $\sup_{t\in(0,1)}\langle \mathbb{B}^\nu(t)\rangle$ is not available. Approximate critical values may be obtained by simulation. However, in order to simulate the supremum of a continuous process with sufficient accuracy, some care is necessary.

Figure 4 shows asymptotic one- and two-sided local levels based on (simulated by 10^6 repetitions) $(1-\alpha)$-quantiles of $\sup_{t\in(0,1)}\langle \mathbb{B}^\nu(t)\rangle$ for $\alpha = 0.05$ and various values of ν. Observe that the two-sided local levels are slightly smaller than their one-sided counterparts.

Note that the right-hand side in (3.5) yields asymptotic (lower as well as upper) LLSFs that are unimodal and symmetric at $\eta = 1/2$; cf. left graph in Figure 4. Such asymptotic LLSFs induce new modified weighted KS tests in the finite setting. Modified weighted KS tests coincide asymptotically with the original counterparts defined by (3.4). However, in the finite case modified tests typically differ from the related original tests. In some cases, modified KS tests may be better than their counterparts defined by (3.4). For instance, for $m = n$, $m = 30, \ldots, 1000$, $\nu = 0$ and $\alpha = 0.05$, we observed that modified KS tests have a much better α-exhaustion than the related original KS tests.

Figure 5 shows exact local levels of weighted KS level α tests together with the asymptotic local levels for $\alpha = 0.05$ and $\nu = 0.25$. It seems that the asymptotic local levels yield a neat reflection of the shape of the exact local levels even for the asymmetric case $m \neq n$. Similar pictures can be observed for a lot of ν-values being not too closed to 0.5. Moreover, asymptotic and exact critical values seem

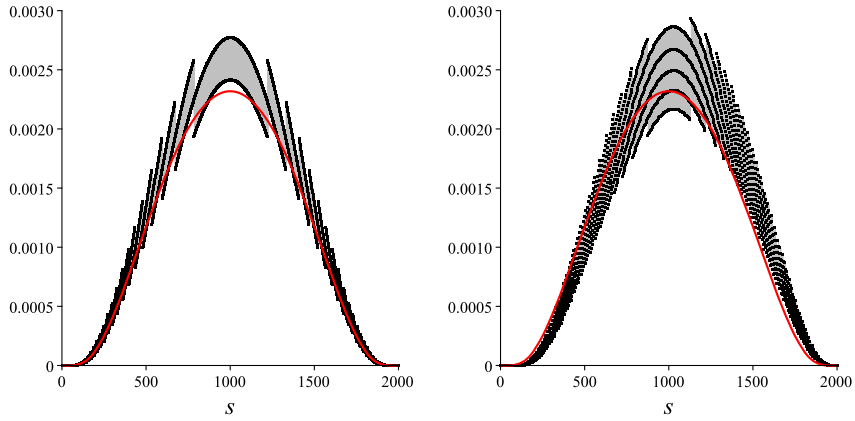

FIG. 5. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to asymptotic and exact level* α *two-sided tests based on* (3.4) *for* $\alpha = 0.05$ *and* $\nu = 0.25$. *Asymptotic local levels* (*red curves*) *are defined by* (3.5) *with* b *being the simulated* $(1-\alpha)$*-quantile of the corresponding asymptotic distribution, here,* $b = 2.002$. *Exact local levels* α_s^{low}, $s \in I_{m,n}$, (*black points*) *correspond to critical values* $b = 1.98504$ *for* $m = n = 1000$ (*left graph*) *and* $b = 1.98431$ *for* $m = 400$, $n = 1600$ (*right graph*).

to be nearly equal. The global level is almost exhausted at least for the exact tests considered here.

The case $\nu = 0.5$. Now we show that two-sample HC statistics, that is, weighted KS statistics defined in (3.4) with $\nu = 0.5$, coincide asymptotically in distribution with the one-sample HC statistic related to the smaller sample.

Let $x_\alpha^+ = -\log(-\log(1-\alpha))$, $x_\alpha = -\log(-\log(1-\alpha)/2)$ and

$$b_m(x) = \sqrt{2\log_2(m)} + \bigl(\log_3(m) - \log(\pi) + 2x\bigr)/\bigl(2\sqrt{2\log_2(m)}\bigr),$$

with $\log_2(m) = \log(\log(m))$ and $\log_3(m) = \log(\log_2(m))$. Moreover, $x_\alpha^{\langle\rangle}$ denotes x_α^+ or x_α. Note that

$$\lim_{m\to\infty} \mathbb{P}_0\Bigl(\sup_{t\in T_m} \bigl|\mathbb{B}^{0.5}(t)\bigr| \le b_m\bigl(x_\alpha^{\langle\rangle}\bigr)\Bigr) = 1-\alpha$$

for $T_m = (\log(m)^5/m, 1 - \log(m)^5/m)$, for example, cf. (11)–(13) together with (15), (16) in [16].

THEOREM 3.2. *Let* $\nu = 0.5$ *and* $n \equiv n(m) \ge m$, $m \in \mathbb{N}$.

(a) *It holds*

$$\lim_{m\to\infty} \mathbb{P}_0\bigl(\mathrm{KS}_{m,n}^{0.5,\langle\rangle} \le b_m\bigl(x_\alpha^{\langle\rangle}\bigr)\bigr) = 1-\alpha.$$

(b) *All local levels of asymptotic level* α *two-sample HC tests converge to zero. Moreover, almost all HC local levels are asymptotically equal to*

$$\alpha_m^* \equiv \frac{-\log(1-\alpha)}{2\log(m)\log_2(m)} \tag{3.6}$$

in the sense that for $s \in I_{m,n}$ fulfilling

$$\lim_{m\to\infty} \frac{\min\{s, m+n-s\}}{(m+n)\log_2^3(m)/m} = \infty \tag{3.7}$$

we get

$$\lim_{m\to\infty} \alpha_s/\alpha_m^* = 1$$

for one- and two-sided tests. In addition, in the two-sided case we get

$$\lim_{m\to\infty} \alpha_s^{\mathrm{low}}/\alpha_m^* = \lim_{m\to\infty} \alpha_s^{\mathrm{up}}/\alpha_m^* = 1/2.$$

Thereby, the convergence of local levels is uniform in s fulfilling (3.7).

(c) *The two-sample HC sensitivity range considered in $s/(m+n)$ on $(0,1)$ coincides with the sensitivity range of the one-sample HC statistic. More precisely, two-sample HC tests are sensitive in the intermediate range of s-values which fulfill* (3.7).

Since all HC local levels in the sensitivity range are equal, a constant asymptotic LLSF is a reasonable choice and leads to the so-called minP tests studied in Section 3.2.

Lower HC local levels related to asymptotic and exact level α HC tests with $\alpha = 0.05$ are given in Figure 6. We observed that the level α is nearly exhausted if m and n are not too small. It seems that the asymptotic local level (3.6) can rather be seen as an upper bound for discrete (exact) local levels if $m = n$. In the case $m \neq n$, it looks like the most of exact local levels are much smaller and a few local levels are much larger than the asymptotic local level. Although the two-sided HC asymptotics is slow, it is much better than the one-sided HC asymptotics, for example, cf. [6].

The case $\nu \in (0.5, 1]$. For $\nu \in (0.5, 1]$, we consider renormalized KS statistics

$$\overline{\mathrm{KS}}_{m,n}^{\nu,\langle\rangle} = \left(\frac{m+n}{mn}\right)^{\nu-0.5} \mathrm{KS}_{m,n}^{\nu,\langle\rangle}. \tag{3.8}$$

In order to derive asymptotic results we have to define some further random variables and processes. Let Z_i, $i \in \mathbb{N}$, be i.i.d. Bernoulli random variables with parameter $p \in (0,1)$. The corresponding binomial process is given by $Y_s = \sum_{i=1}^{s} Z_i$, $s \in \mathbb{N}$. Let $\widetilde{Y}_s$ denote an independent copy of Y_s and let N and $\widetilde{N}$ be independent standard (right-continuous) Poisson processes.

The next theorem shows that the asymptotics of renormalized KS tests with $\nu \in (0.5, 1]$ differs from the one-sample asymptotics unless one sample is much larger than the other. To this end, define

$$Q_{\nu,p}^{\langle\rangle} = \frac{1}{(p(1-p))^{\nu}} \max\left\{\sup_{s\in\mathbb{N}} \frac{\langle sp - Y_s\rangle}{s^{\nu}}, \sup_{s\in\mathbb{N}} \frac{\langle \widetilde{Y}_s - sp\rangle}{s^{\nu}}\right\}, \qquad p \in (0,1),$$

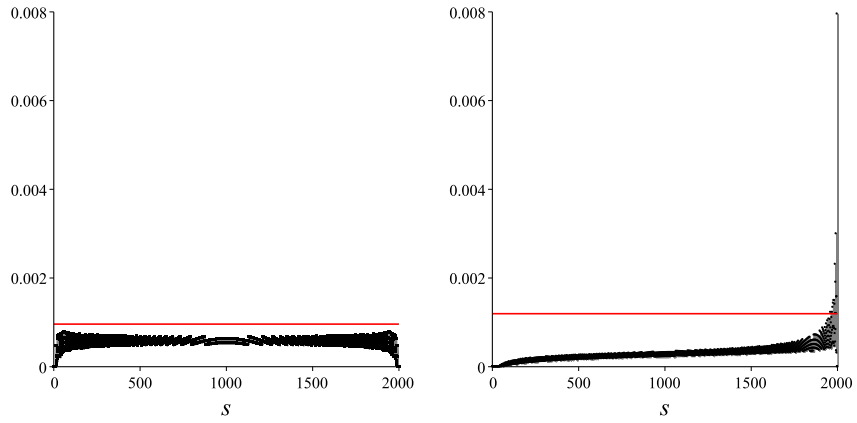

FIG. 6. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to asymptotic and exact level* α *two-sided HC tests, that is, tests based on* (3.4) *with* $\nu = 0.5$. *Here,* $\alpha = 0.05$. *Asymptotic local levels* (*red lines*) *are equal to* $\alpha_m^*/2$ *with* α_m^* *defined in* (3.6), *while the exact* α_s^{low}, $s \in I_{m,n}$, (*black points*) *correspond to critical values* $b = 3.2451, 3.4505$ *for* $m = n = 1000$ (*left graph*) *and* $m = 400, n = 1600$ (*right graph*), *respectively.*

and

$$Q_{\nu,p}^{\langle\rangle} = \max\left\{\sup_{t>0} \frac{\langle t - N(t)\rangle}{t^\nu}, \sup_{t>0} \frac{\langle \widetilde{N}(t) - t\rangle}{t^\nu}\right\}, \qquad p \in \{0, 1\}.$$

THEOREM 3.3. *Let* $\nu \in (0.5, 1]$. *Under* $H^=$, *the test statistic* $\overline{\mathrm{KS}}_{m,n}^{\nu,\langle\rangle}$ *converges in distribution to* $Q_{\nu,p}^{\langle\rangle}$ *as* $m, n \to \infty$ *and* $m/(m+n) \to p \in [0, 1]$.

The distribution of $Q_{\nu,p}^{\langle\rangle}$ with $p \in (0, 1)$ seems to be discrete, and hence even asymptotic weighted KS tests are typically not α-exhaustive and the effective level may be much smaller than the prespecified α. For $p \in \{0, 1\}$, we obtain asymptotically α-exhaustive tests at least for $\alpha < 0.5$ and the CDF of $Q_{\nu,p}^{\langle\rangle}$ can be calculated with formulas given in [22]. However, a simple analytical representation is only available for $\nu = 1$; cf. results in [25].

For $\nu \in (0.5, 1]$, almost all asymptotic local levels related to asymptotic level α weighted KS tests are equal to zero. Many cases have to be distinguished in order to identify all positive asymptotic local levels. We omit this here and give a brief hint only. For $p \in (0, 1)$, only some extreme local levels are positive in the left and/or right tail, and hence the sensitivity range of such tests lies in the extreme tails. For $p \in \{0, 1\}$, we get that only α_s with $\lim_{m,n\to\infty} \min\{s, m+n-s\} \min\{m, n\}/(m+n) \in (0, \infty)$ may be asymptotically positive.

Two examples of local level behavior in the right tail are displayed in Figure 7. In the case $m = 400, n = 1600$ local levels are zero in the left tail, while for $m = n = 1000$ local levels are symmetric in $s \in I_{m,n}$. Surprisingly, asymptotic and exact

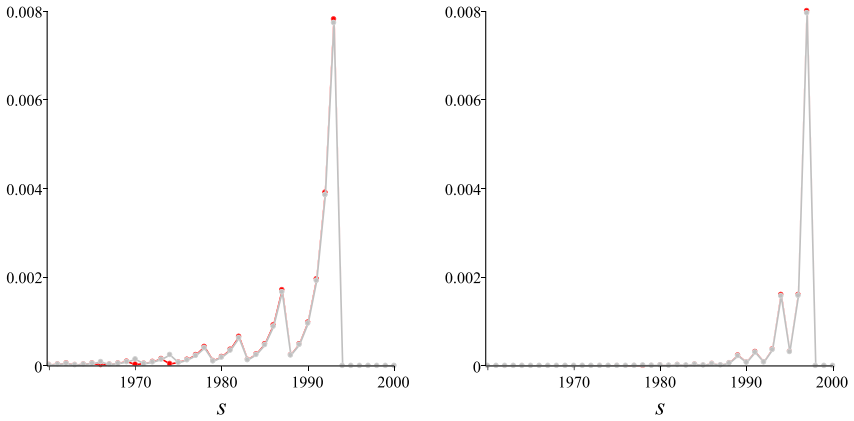

FIG. 7. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to asymptotic and exact level* α *two-sided tests based on* (3.4) *with* $\nu = 0.75$ *and* $\alpha = 0.05$. *The exact* α_s, $s \in I_{m,n}$, (*gray*) *correspond to the critical values* $b = 10.48996, 15.91735$ *and* $\mathbb{P}_0(\mathrm{KS}_{m,n}^{\nu,||} > b) = 0.03598, 0.01771$ *for* $m = n = 1000$ (*left graph*) *and* $m = 400, n = 1600$ (*right graph*), *respectively. Asymptotic local levels* (*red*) *are defined via Theorem* 3.3. *Thereby, the asymptotic critical values are* $b = 10.46635, 15.90542$ *and the asymptotic global levels are* 0.03632, 0.01787, *respectively.*

local levels are very close to each other. This seems to be true in general for ν not too close to 0.5.

REMARK 3.2. The proofs of Theorems 3.1–3.2 also show that the empirical process $\mathrm{KS}_{m,n}^{\nu}(t)$ converges in the sense of Hungarian constructions to a suitably weighted Brownian bridge $\mathbb{B}^{\nu}(t)$ uniformly on some subinterval of $(0, 1)$, cf. (A3.11) in [11] for $\nu \in [0, 0.5)$ and (A3.14) in [11] for $\nu = 0.5$. Analogously, the proof of Theorem 3.3 implies uniform convergence of the renormalized KS process to suitably weighted Binomial processes for $p \in (0, 1)$ and Poisson processes for $p \in \{0, 1\}$.

4. Two-sample GOF tests with approximately equal local levels. In this section, we provide new GOF tests that can be viewed as promising alternatives for the HC tests. In the one-sample case, the concept of equal local levels was recently introduced and studied in [16], [17] and [15]. The requirement of equal local levels leads to so-called minimum p-value (minP) statistics. In [4], Berk and Jones delivered a general theory for (one-sample) minP test statistics with respect to Bahadur efficiency. Thereby, they denoted minP statistics as *minimum level attained* statistics. Moreover, Berk and Jones studied various one-sample GOF tests including minP tests with respect to Bahadur efficiency and asymptotic properties; cf. [5]. The tail sensitive confidence bands introduced in [1] (see also [2]) correspond to the minP test in the sense of duality between tests and confidence sets. Recent investigations have shown that one-sample minP GOF tests are asymptotically equivalent to one-sample HC tests but have favorable finite properties; cf.

the discussions in [16], [17] and [15]. It seems that at least in the one-sample case minP GOF tests yield a compromise between KS (larger local levels in the central range) and HC tests (larger local levels in the tails).

In Section 4.1, we introduce a two-sample minP test and study its finite properties. In Section 4.2, we provide minP asymptotics.

4.1. *Two-sample minP GOF tests*. Formally, the minimum of any set of p-values can be seen as a minP test statistic. Hence, minP GOF tests depend heavily on the choice of p-values for testing local null hypotheses. First, we define local p-values in terms of $V_{m,s}$. Conditionally on $S_{m+n}(t) = s$, one-sided p-values for testing local null hypotheses $H_t^{\geq}$ are defined by

$$p_s = F_{\mathrm{Hyp}}(V_{m,s}|s, m, n), \qquad s \in I_{m,n},$$

and two-sided p-values for testing $H_t^{=}$ are defined by

$$p_s = 2\min\bigl\{1/2, F_{\mathrm{Hyp}}(V_{m,s}|s, m, n), 1 - F_{\mathrm{Hyp}}(V_{m,s} - 1|s, m, n)\bigr\}, \qquad s \in I_{m,n}.$$

In both cases, the minP statistic is defined by $\min_{s\in I_{m,n}} p_s$ and the corresponding global null hypothesis is rejected if $\min_{s\in I_{m,n}} p_s \leq \alpha_{m,n}^{\mathrm{loc}}$. In order to exhaust the prespecified level α as sharp as possible, $\alpha_{m,n}^{\mathrm{loc}}$ is chosen as large as possible and equal to the maximum of all local levels. Note that $\alpha_{m,n}^{\mathrm{loc}}$-values typically differ in the one- and two-sided case. The minP tests considered here can be rewritten in terms of the acceptance region $A_{m,n}$ of the form (2.2) with critical values c_s, d_s, $s \in I_{m,n}$, defined by

$$c_s = \max\bigl\{x \in \{0, \ldots, m\} : F_{\mathrm{Hyp}}(x - 1|s, m, n) \leq \alpha_{m,n}^{\mathrm{loc}}\bigr\} \tag{4.1}$$

in the one-sided case and

$$c_s = \max\bigl\{x \in \{0, \ldots, m\} : F_{\mathrm{Hyp}}(x - 1|s, m, n) \leq \alpha_{m,n}^{\mathrm{loc}}/2\bigr\}, \tag{4.2}$$

$$d_s = \min\bigl\{x \in \{0, \ldots, m\} : F_{\mathrm{Hyp}}(x|s, m, n) \geq 1 - \alpha_{m,n}^{\mathrm{loc}}/2\bigr\} \tag{4.3}$$

in the two-sided case. It is obvious that minP critical values fulfill $d_s = m - c_{m+n-s}$, $s \in I_{m,n}$, and, if $m = n$, $c_s = s - m + c_{2m-s}$, $s \in I_{m,m}$. Consequently, we get $\alpha_s^{\mathrm{low}} = \alpha_{m+n-s}^{\mathrm{up}}$ and $\alpha_s^{\mathrm{low}} = \alpha_{2m-s}^{\mathrm{low}}$ for $s \in I_{m,n}$ respectively, for resulting minP local levels.

REMARK 4.1. It can easily be seen that one-sided minP tests based on $\alpha_{m,n}^{\mathrm{loc}}$ can be obtained by derandomizing conditional one-sided UMPU tests at level $\alpha_{m,n}^{\mathrm{loc}}$ for the (one-sided) comparison of two binomial distributions with sample sizes m and n. Two-sided minP tests can be represented as a combination of two one-sided level $\alpha_{m,n}^{\mathrm{loc}}/2$ tests. Therefore, the theory of UMPU tests in 2×2-tables can be a useful tool for studying various properties of minP tests.

The next lemma is a consequence of results obtained in [13] concerning structural properties of UMPU tests in 2×2-tables.

LEMMA 4.1. *The critical values defined in* (4.1)–(4.3) *are proper. Hence, the corresponding minP tests are proper.*

Due to the discreteness of the hypergeometric distributions, equal local levels of all local tests are not possible. However, minP local levels may be viewed as approximately equal if the sample sizes m and n are large enough. In any case, for one-sided minP tests we get that $\alpha_s^{\text{low}} \le \alpha_{m,n}^{\text{loc}}$, $s \in I_{m,n}$, and there exists at least one $s_0 \in I_{m,n}$ such that $\alpha_{s_0}^{\text{low}} = \alpha_{m,n}^{\text{loc}}$. In the two-sided case, we obtain $\alpha_s^{\text{low}}, \alpha_s^{\text{up}} \le \alpha_{m,n}^{\text{loc}}/2$, $s \in I_{m,n}$, and there exists at least one $s_0 \in I_{m,n}$ such that $\alpha_{s_0}^{\text{low}} = \alpha_{m+n-s_0}^{\text{up}} = \alpha_{m,n}^{\text{loc}}/2$.

Figure 8 shows lower local levels of level α two-sided minP GOF tests with $\alpha = 0.05$ and some sample sizes. Note that minP local levels are identical to HC local levels in the case $m = n = 50$, cf. $\nu = 0.5$ in Figure 2. This is often the case for smaller m- and n-values. In general, HC and minP GOF local levels seem to be similar, while minP and HC local levels show a completely different behavior if $m \ne n$; cf. Figure 3 for $\nu = 0.5$.

REMARK 4.2. An alternative version of two-sided minP GOF tests can be constructed by derandomizing conditional level $\alpha_{m,n}^{\text{loc}}$ UMPU tests for the two-sided comparison of two binomial distributions. For $m = n$, we get the same two-sided minP tests, no matter whether we start with one- or two-sided UMPU tests. In contrast, for $m \ne n$, different versions of minP tests are possible. For example,

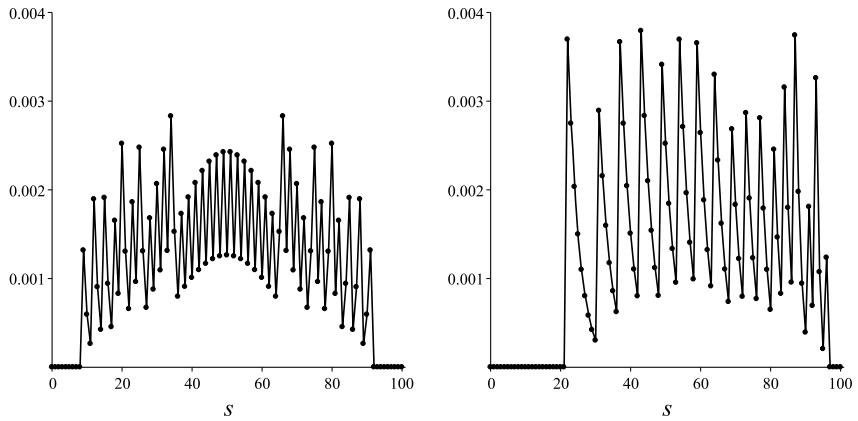

FIG. 8. *Lower local levels* α_s^{low}, $s \in I_{m,n}$, *related to level* α *two-sided minP GOF tests with* $\alpha = 0.05$. *Left graph*: $m = 50$, $n = 50$, $\alpha_{m,n}^{\text{loc}} = 0.00283$, $\mathbb{P}_0(\min_{s \in I_{m,n}} p_s \le \alpha_{m,n}^{\text{loc}}) = 0.453$; *right graph*: $m = 20, n = 80$, $\alpha_{m,n}^{\text{loc}} = 0.0038$, $\mathbb{P}_0(\min_{s \in I_{m,n}} p_s \le \alpha_{m,n}^{\text{loc}}) = 0.0492$.

local p-values can be defined as the smallest possible level α' such that the derandomized version of the conditional level α' two-sided UMPU test leads to rejection for the observed $V_{m,s}$. This leads to an alternative minP test with asymmetric local levels. There is some evidence that the asymptotic behavior of this minP test coincides with the minP test based on one-sided tests. From [13], we get that two-sided (as well as one-sided, see Lemma 4.1) UMPU tests lead to proper critical values for $V_{m,s}$.

4.2. *Asymptotics of two-sample minP GOF tests.* In this subsection, we provide the asymptotics of two-sample minP GOF tests defined by critical values (4.1)–(4.3) and compare exact and corresponding asymptotic minP local levels. We first provide asymptotic minP critical values.

THEOREM 4.1. *For $m, n \in \mathbb{N}$, $\alpha^{\mathrm{loc}}_{m,n} \in (0,1)$ and $\alpha^*_{m,n} \equiv \alpha^*_{\min\{m,n\}}$ with α^*_m defined in (3.6) we get*

$$\lim_{m,n\to\infty} \mathbb{P}_0\Big(\min_{s\in I_{m,n}} p_s > \alpha^{\mathrm{loc}}_{m,n}\Big) = 1-\alpha \quad \textit{iff} \quad \lim_{m,n\to\infty} \frac{\alpha^{\mathrm{loc}}_{m,n}}{\alpha^*_{m,n}} = 1.$$

It follows that two-sample minP local levels and almost all HC local levels are asymptotically equal; cf. (b) in Theorem 3.2. That is, minP and HC tests coincide asymptotically. Moreover, the two-sample minP asymptotics coincides with the one-sample minP asymptotics related to the smaller sample.

Figure 9 shows lower minP local levels of asymptotic and exact level α minP GOF tests as well as the corresponding HC local levels from Figure 6. Similarly, as

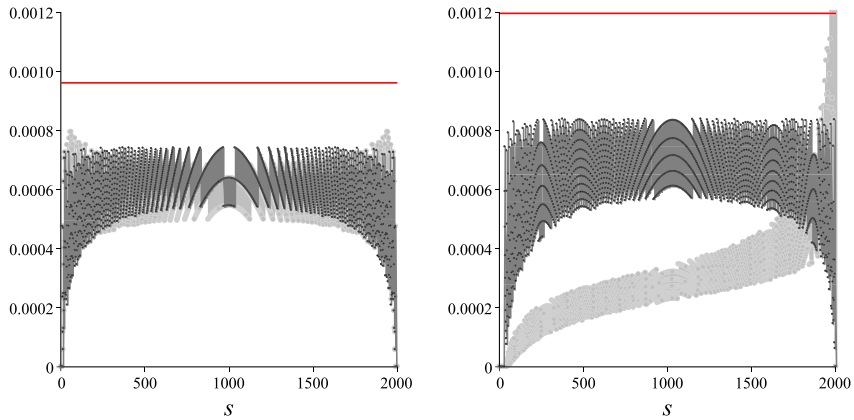

FIG. 9. *Lower local levels α^{low}_s, $s \in I_{m,n}$, related to asymptotic (red lines) and exact (black points) level α two-sided minP tests together with the corresponding HC local levels for $\alpha = 0.05$. Asymptotic local levels are equal to $\alpha^*_m/2$ with α^*_m defined in (3.6), exact α^{low}_s, $s \in I_{m,n}$, are based on critical values $\alpha^{\mathrm{loc}}_{m,n} = 0.000741, 0.000838$ for $m = n = 1000$ (left graph) and $m = 400$, $n = 1600$ (right graph), respectively. HC local levels (light gray curves) are the same as in Figure 6.*

in the HC case, the global level of minP GOF tests seems to be nearly exhausted for sample sizes being not too small. Moreover, it looks like minP and HC local levels are very close to each other for equal (or nearly equal) sample sizes, while local levels differ considerably if sample sizes are unequal. Unfortunately, the minP asymptotics seems to be only slightly better than the HC asymptotics in the case of equal sample sizes. However, for $m \neq n$ it appears as if the exact minP local levels are much closer to the asymptotic ones than HC local levels. All in all, minP GOF tests can be seen as more balanced HC tests in two- as well as one-sample settings.

5. Power considerations. In Section 5.1, we briefly study and compare the (overall) power of the two-sample GOF tests considered in Sections 3 and 4. In Section 5.2, we discuss some versions of local power.

5.1. *Power.* The power of a two-sample GOF test with acceptance region (2.2) is defined by $\mathbb{P}_{F,G}(V_m \notin A_{m,n})$. Due to the lack of general formulas for the power of two-sample GOF tests (with few exceptions, cf., e.g., [28] for Lehmann alternatives), simulation seems the method of choice. Some general results on the global power function of GOF tests can be found in [20].

Given two different GOF tests, we will typically find different pairs of distributions such that the first test dominates the second test and vice versa. Assuming that we have no prior knowledge where the CDFs F and G may differ, the choice of a test can only be a compromise in the sense that the power of the chosen test should behave reasonably well over a large class of CDFs. From a practical point of view, it seems a good strategy to avoid tests with excellent power under specific alternatives and extremely poor power under other possible alternatives.

First of all, it is worth to mention that, given $m \neq n$, the overall power typically differs if we exchange m and n; cf. Table 1. This issue becomes more serious if tests have nonsymmetric one-sided local levels. A general observation is that unequal sample sizes may lead to strange power behavior. For example, an increasing sample size may lead to decreasing power; cf. the HC tests in Table 1. Here, a look

TABLE 1
Power of the two-sided KS ($\nu = 0$), *HC* ($\nu = 0.5$) *and minP tests for* $X_1, \ldots, X_m$ *being i.i.d.* $N(0, 1)$ *distributed and* $Y_1, \ldots, Y_n$ *being i.i.d.* $N(0, 3)$ *distributed* ($\alpha = 0.05$)

	KS ($\nu = 0$)	**HC** ($\nu = 0.5$)	**minP**
$m = 20, n = 20$	0.29	0.62	0.62
$m = 20, n = 80$	0.55	0.31	0.83
$m = 80, n = 20$	0.67	0.99	0.97

at local levels yields an explanation of such phenomena, for example, cf. Figure 3. More precisely, lower and upper local levels may be large in the wrong tails. HC tests have the most skewed local levels while KS and minP local levels are more (although not perfect) symmetric. Therefore, we recommend to avoid tests with extremely nonsymmetric local levels, for example, weighted KS tests with $\nu \geq 0.5$, if sample sizes differ considerably. In this case a minP test seems a good compromise. If one prefers weighted KS tests with $\nu < 0.5$, for example, because their local levels in the central range are even bounded away from zero asymptotically, we recommend the corresponding modified version based on the LLSFs induced by the asymptotics in (3.5); cf. Section 2.4.

If $m \approx n$, local levels of weighted KS tests with $\nu \leq 0.5$ and even ν's slightly larger that 0.5 are nearly symmetric, that is, right and left tails obtain approximately similar weights. Thereby, HC tests come close to the corresponding minP tests.

We simulated various location-scale normal models with equal as well as unequal sample sizes. We observed that the power of weighted KS tests as a function of ν is more or less unimodal. Thereby, the power is typically maximal for some $\nu \in [0.3, 0.6]$. Surprisingly, $\nu \in (0, 0.4]$ seems to beat $\nu = 0$ so that we do not recommend original KS tests. One reason for this behavior may be that the size of the acceptance regions, for example, $\sum_{s \in I_{m,n}} (d_s - c_s)$, seems to be minimal for some $\nu \in [0.3, 0.4]$. For approaches to minimize the acceptance region of one-sample GOF tests, we refer to [14] and [29]. Furthermore, we observe that minP tests (and hence HC tests in case $m \approx n$) can be moderately more (rather for location-scale or pure scale models) or less (rather for pure location models) powerful than weighted KS tests with $\nu \in [0.0, 0.4]$. In total, it seems that minP tests have a stable high power at least for most of the normal models considered in our simulations. However, if one expects scale alternatives, one may choose a ν slightly larger than 0.5 in order to weight the tails a little more. This may lead to an increased power compared to HC and minP. Thereby, one should check the resulting local levels in order to avoid overweighting of the tails. Weighted KS tests with larger ν's may be an option to test a reference sample against outliers in a second sample.

In case $m = n$, the effective level of the classical KS test can be considerably lower than the prespecified α which may result in an unnecessary low overall power. Typically, if m and n are not relatively prime, there may be room for improvement. We observed that the effective level of the KS variant based on the corresponding LLSF typically comes closer to the prespecified α. Therefore, if one insists on the KS test, it is always worth to check whether the test based on the corresponding LLSF yields a tighter effective level. If so, one should replace KS by the latter one. For example, for $m = n = 120$ and $\alpha = 0.05$ we observe $\alpha_{\text{eff}} \approx 0.035$ for KS while the $\alpha_{\text{eff}} \approx 0.0498$ for the KS variant based on the LLSF. A further option is to choose a ν slightly larger than 0.

5.2. *Power of local tests.* A referee of this paper suggested to study some kind of local power function. In order to receive an impression of the behavior of different tests, one may study, for example, the power of local tests for testing H_t. First, we may consider *local t-power* defined as local rejection probability $\beta_1(t) = \mathbb{P}_{F(t),G(t)}(H_t \text{ is rejected})$ for each t. This results in the computation of the unconditional power of two-sample binomial tests and can be time consuming. Note that $\sup_t \beta_1(t)$ yields a lower bound for the overall power. For illustration, one may plot $H(t) = \eta F(t) + (1-\eta)G(t)$ versus $\beta_1(t)$ with $\eta = m/(m+n)$. Second, we may consider *local s-power* defined as local rejection probability $\beta_2(s) = \mathbb{P}_{F,G}(V_{m,s} \notin [c_s, d_s])$ for each $s \in I_{m,n}$. In this case, it seems hard to say anything about the distribution of $V_{m,s}$ which depends on F and G. Nevertheless, given F and G, we can easily simulate $V_{m,s}$ as well as the test decision in s, and hence the local s-power $\beta_2(s)$. In this case, one may plot $s/(m+n)$ versus $\beta_2(s)$. Typically, both local t- and s-powers lead to very similar shapes indicating where local powers are small or large.

One may also look at least favorable distributions w.r.t. the overall power. Suppose for a moment we restrict attention to the case $F \le G$ and test $H^=$ versus the one-sided alternative $F(t) < G(t)$ for at least one t. Then the *least favorable overall power* for a proper one-sided test φ defined as

$$\beta(t_0, q, \delta, \varphi) = \inf_{F \le G : F(t_0) = q, G(t_0) \ge q+\delta} \mathbb{P}_{F,G}(\varphi = 1), \qquad q \in (0, 1-\delta), \tag{5.1}$$

may be of interest. Clearly, the local t-power $\beta_1(t_0)$ evaluated under $G(t_0) = q + \delta$ yields a lower bound for $\beta(t_0, q, \delta, \varphi)$. Noting that the power decreases if F (G) increases (decreases), an upper bound may be obtained for distributions F and G with (i) $F(t) = G(t)$ for $t < t_0$ and $t \ge t_0 + \varepsilon$, and (ii) $F(t) = q$, $G(t) = q + \delta$ for $t \in [t_0, t_0 + \varepsilon)$ with $\varepsilon > 0$. Altogether, we get that

$$\beta_1(t_0) \le \beta(t_0, q, \delta, \varphi) \le \beta_1(t_0) + \mathbb{P}_0(\varphi = 1),$$

which yields a nice connection between power of local tests and least favorable overall power. It seems hard if not impossible to cover the two-sided case in a similar way.

6. Concluding remarks. Local levels of KS-type tests indicate which type of s-values (e.g., extremes, intermediates or central) contribute to the overall level α. Asymptotic local levels for extreme and central s-values may be bounded away from zero or may tend to zero. Typically, local levels of intermediate s-values tend to zero. While all local levels of minP versions tend to zero, weighted KS tests with weight function (3.3) have asymptotically positive local levels in the central range for $\nu \in [0, 0.5)$ and, if $m/(m+n) \to p \in (0,1)$, asymptotically positive local levels in the extreme range for $\nu \in (0.5, 1]$. In any case, large local levels for s-values moving away from the center become more and more expensive with respect to the consumption of the overall level α and the extreme tails are extremely

expensive. Thereby, due to the underlying hypergeometric distributions, local levels of s-values close to zero or $m+n$ are typically zero for conventional α-values. Taking all these points into account, one may design new KS-type tests in terms of local levels or a LLSF in order to improve the power against specific alternatives.

The LLSFs $\tilde{\alpha}_b(\eta) = \Phi(-b[\eta(1-\eta)]^{\nu-1/2})$, which are induced by the asymptotics of weighted KS-type tests with $\nu < 0.5$, have the interesting property, that the (right and left hand) derivatives in 0 and 1 are zero for $\nu \in [0, 0.5)$ and $b > 0$. One may also consider shape functions with different behavior in 0 and 1 in order to give more weight to the tails. For example, consider the LLSFs $\tilde{\alpha}_\kappa(\eta) = \kappa[\eta(1-\eta)]^\nu$ with tuning parameter $\kappa > 0$ for $\nu > 0$. Now the corresponding derivatives in 0 and 1 are $\pm\infty$ for $\nu \in (0,1)$, $\pm\kappa$ for $\nu = 1$ and 0 for $\nu > 1$.

We may also combine the specific advantages of different types of sensitivity behavior in terms of local levels. For a combination of classical test statistics in order to overcome the poor sensitivity in the tails of the one-sample KS test, we refer to [23] together with [24]. Alternatively, we may combine local levels of different tests. As an example, we consider two combinations of (asymptotic) minP and original KS tests. The asymptotic KS-related LLSFs $\tilde{\alpha}_\kappa^{\mathrm{KS}}(\eta)$ are given by the right-hand side of (3.5) with b being the critical value of the asymptotic level κ two-sided original KS test. For the minP part (assuming that $m = \min\{m,n\}$) we choose LLSFs $\alpha_\kappa^{\mathrm{minP}}(\eta) \equiv \alpha_m^*$ defined in (3.6) with κ instead of α. Then the (symmetric) LLSFs $\tilde{\alpha}_\kappa^{\mathrm{low}} = \tilde{\alpha}_\kappa^{\mathrm{up}} = \max\{\tilde{\alpha}_\kappa^{\mathrm{KS}}, \alpha_\kappa^{\mathrm{minP}}\}$ and $\tilde{\alpha}_\kappa^{\mathrm{low}} = \tilde{\alpha}_\kappa^{\mathrm{up}} = \tilde{\alpha}_\kappa^{\mathrm{KS}} + \alpha_\kappa^{\mathrm{minP}}$ lead to two new GOF tests. In both cases, we choose κ as large as possible such that the resulting combined tests are level α tests.

Figure 10 shows two-sided local levels of the aforementioned combined tests for $m = n = 1000$ and $\alpha = 0.05$. The two versions lead to very similar shapes of

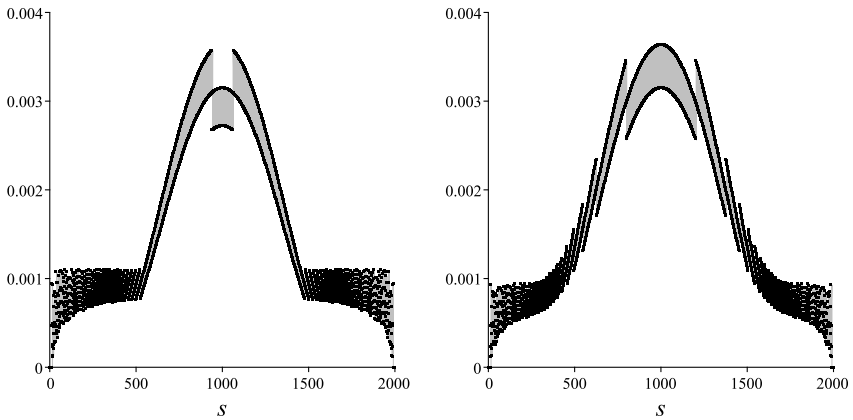

FIG. 10. *Two-sided local levels* α_s, $s \in I_{m,n}$, *related to level* α *two-sided GOF tests that are a combination of original KS and minP tests with* $\alpha = 0.05$ *and* $m = n = 1000$. *Left graph*: *maximum version*; *right graph*: *sum version*.

the local levels. Moreover, these tests coincide asymptotically under the null hypothesis and $\kappa = \alpha/2$ leads to a level α test asymptotically. That is, the probability to reject the true null hypothesis in the tails as well as in the central range tends to $\alpha/2$ for increasing sample sizes. We may also choose different weights for the tests to be combined.

We conclude with some general remarks on the connection between weight functions w and LLSFs $\tilde{\alpha}_\kappa$. Several classes of weight functions including Chibisov–O'Reilly functions and the larger class of Erdős–Feller–Kolmogorov–Petrovski (EFKP) upper-class functions were extensively studied, for example, in [7] in connection with the uniform empirical process (and hence one-sample GOF tests), the uniform quantile process and the Brownian bridge. We note that the weight functions w_ν studied in Section 3 are Chibisov–O'Reilly functions for $\nu \in [0, 0.5)$. For an EFKP upper-class weight function w, weighted KS statistics defined in (3.2) converge under $H^=$ to the supremum of the corresponding weighted Brownian bridge $Z = \sup_{\eta\in(0,1)} \langle \mathbb{B}(\eta)\rangle / w(\eta)$, which is a nondegenerate random variable; cf. Theorem 4.2.3 in [7]. In such cases, the corresponding (upper and lower) asymptotic LLSFs are given by $\tilde{\alpha}_\kappa(\eta) = \Phi(-\kappa w(\eta)/\sqrt{\eta(1-\eta)})$, where κ is the asymptotic critical value. Except rare cases, explicit formulas for the distribution of Z are not available. However, one may simulate the distribution of Z and the critical value κ in order to get a glimpse of the shape of the local levels and to judge whether they may lead to useful GOF tests. Finally, any LLSF $\tilde{\alpha}_\kappa$ defines a bounding function b_κ for the Brownian Bridge via $\mathbb{P}(\mathbb{B}(\eta) \le b_\kappa(\eta)) = \tilde{\alpha}_\kappa(\eta)$ and vice versa. It may be of general interest to characterize LLSFs leading to $b_{\kappa(\alpha)}$ such that $\mathbb{P}(\mathbb{B}(\eta) \le b_{\kappa(\alpha)}(\eta), \eta \in (0, 1)) = 1 - \alpha$ for all $\alpha \in (0, 1)$.

Acknowledgments. The authors are grateful to the referee and the Associate Editor for their extremely valuable and constructive comments and suggestions. Special thanks are due to the Editor, E. I. George, for handling the manuscript.

SUPPLEMENTARY MATERIAL

Supplement A: Proofs and computation of global levels (DOI: 10.1214/17-AOS1647SUPPA; .pdf). In Section A1, we prove Lemma 3.1. Section A2 focuses on the computation of global levels. Proofs of asymptotic results in Sections 3.2 and 4.2 are given in Section A3. Section A4 provides technical results for proofs in Section A3.

Supplement B: Animated graphics of local levels (DOI: 10.1214/17-AOS1647SUPPB; .pdf). In this supplement, we illustrate the convergence of local levels related to weighted KS as well as minP tests to the corresponding asymptotic counterparts by means of animated graphics.

REFERENCES

[1] ALDOR-NOIMAN, S., BROWN, L. D., BUJA, A., ROLKE, W. and STINE, R. A. (2013). The power to see: A new graphical test of normality. *Amer. Statist.* **67** 249–260. MR3303820

[2] ALDOR-NOIMAN, S., BROWN, L. D., BUJA, A., ROLKE, W. and STINE, R. A. (2014). Correction to: "The power to see: A new graphical test of normality." [*Amer. Statist.* **67**(4) (2013) 249–260. MR3303820] *Amer. Statist.* **68** 318. MR3280623

[3] BARNARD, G. A. (1947). Significance test for 2×2 tables. *Biometrika* **34** 123–138. MR0019285

[4] BERK, R. H. and JONES, D. H. (1978). Relatively optimal combinations of test statistics. *Scand. J. Stat.* **5** 158–162. MR0509452

[5] BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47–59. MR0521531

[6] CANNER, P. L. (1975). A simulation study of one- and two-sample Kolmogorov–Smirnov statistics with a particular weight function. *J. Amer. Statist. Assoc.* **70** 209–211.

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

[8] DOKSUM, K. A. and SIEVERS, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63** 421–434. MR0443210

[9] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[10] DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751

[11] FINNER, H. and GONTSCHARUK, V. (2018). Supplement A to "Two-sample Kolmogorov-Smirnov type tests revisited: Old and new tests in terms of local levels": Proofs and computation of global levels. DOI:10.1214/17-AOS1647SUPPA.

[12] FINNER, H. and GONTSCHARUK, V. (2018). Supplement B to "Two-sample Kolmogorov-Smirnov type tests revisited: Old and new tests in terms of local levels": Animated graphics of local levels. DOI:10.1214/17-AOS1647SUPPB.

[13] FINNER, H. and STRASSBURGER, K. (2002). Structural properties of UMPU-tests for 2×2-tables and some applications. *J. Statist. Plann. Inference* **104** 103–120. MR1900521

[14] FREY, J. (2008). Optimal distribution-free confidence bands for a distribution function. *J. Statist. Plann. Inference* **138** 3086–3098. MR2442227

[15] GONTSCHARUK, V. and FINNER, H. (2017). Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Comm. Statist. Theory Methods* **46** 2332–2342. MR3576717

[16] 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

[17] 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

[18] HODGES, J. L. (1958). The significance probability of the Smirnov two-sample test. *Ark. Mat.* **3** 469–486. MR0097136

[19] JAGER, L. and WELLNER, J. A. (2004). On the "Poisson boundaries" of the family of weighted Kolmogorov statistics. In *A Festschrift for Herman Rubin. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 319–331. IMS, Beachwood, OH. MR2126907

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

[21] KOLMOGOROV, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari* **4** 83–91. Translated by Q. Meneghini as On the empirical determination of a distribution function. In *Breakthroughs in Statistics* **II**. *Springer Series in Statistics* (*Perspectives in Statistics*) (S. Kotz and N. L. Johnson, eds.) 106–113. Springer, New York. MR1182795

[22] MASON, D. M. (1983). The asymptotic distribution of weighted empirical distribution functions. *Stochastic Process. Appl.* **15** 99–109. MR0694539

[23] MASON, D. M. and SCHUENEMEYER, J. H. (1983). A modified Kolmogorov–Smirnov test sensitive to tail alternatives. *Ann. Statist.* **11** 933–946. MR0707943

[24] MASON, D. M. and SCHUENEMEYER, J. H. (1992). Correction to: "A modified Kolmogorov–Smirnov test sensitive to tail alternatives." [*Ann. Statist.* **11**(3) (1983) 933–946. MR0707943] *Ann. Statist.* **20** 620–621. MR1150371

[25] PYKE, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Stat.* **30** 568–576. MR0107315

[26] SMIRNOFF, N. (1939). Sur les écarts de la courbe de distribution empirique. *Rec. Math. N.S.* [*Mat. Sbornik*] **6**(48) 3–26. MR0001483

[27] SMIRNOV, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow Univ. Math. Bull.* **2** 3–16. MR0002062

[28] STECK, G. P. (1969). The Smirnov two sample tests as rank tests. *Ann. Math. Stat.* **40** 1449–1466. MR0246473

[29] XU, X., DING, X. and ZHAO, S. (2009). The reduction of the average width of confidence bands for an unknown continuous distribution function. *J. Stat. Comput. Simul.* **79** 335–347. MR2522355

INSTITUTE FOR BIOMETRICS
AND EPIDEMIOLOGY
GERMAN DIABETES CENTER (DDZ)
LEIBNIZ CENTER FOR DIABETES RESEARCH
AT HEINRICH HEINE UNIVERSITY DÜSSELDORF
AUF'M HENNEKAMP 65
40225 DÜSSELDORF
GERMANY
E-MAIL: finner@ddz.uni-duesseldorf.de

INSTITUTE FOR HEALTH SERVICES RESEARCH
AND HEALTH ECONOMICS
GERMAN DIABETES CENTER (DDZ)
LEIBNITZ CENTER FOR DIABETES RESEARCH
AT HEINRICH HEINE UNIVERSITY DÜSSELDORF
AUF'M HENNEKAMP 65
40225 DÜSSELDORF
GERMANY
E-MAIL: veronika@gontscharuk.de