Received: 3 July 2020 | Revised: 8 March 2021 | Accepted: 13 May 2021

DOI: 10.1111/sjos.12544

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Cramér-von Mises tests for change points

Rasmus Erlemann[1] | Richard Lockhart[2] | Rihan Yao[2]

[1]Department of Mathematical Sciences, NTNU, Norway

[2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada

Correspondence
Rasmus Erlemann, Kilu 17, Tallinn 13516, Estonia.
Email: rasmus.erlemann@uncc.edu

Abstract

We study two nonparametric tests of the hypothesis that a sequence of independent observations is identically distributed against the alternative that at a single change point the distribution changes. The tests are based on the Cramér–von Mises two-sample test computed at every possible change point. One test uses the largest such test statistic over all possible change points; the other averages over all possible change points. Large sample theory for the average statistic is shown to provide useful p-values much more quickly than bootstrapping, particularly in long sequences. Power is analyzed for contiguous alternatives. The average statistic is shown to have limiting power larger than its level for such alternative sequences. Evidence is presented that this is not true for the maximal statistic. Asymptotic methods and bootstrapping are used for constructing the test distribution. Performance of the tests is checked with a Monte Carlo power study for various alternative distributions.

KEYWORDS

asymptotic distribution, change point detection, Cramér–von Mises two-sample test, Monte Carlo simulation, nonparametric test statistics

1 | INTRODUCTION

Consider a sequence of independent observations $X_1, \dots, X_n$. We propose tests of the null hypothesis that the X_i are independent and identically distributed (iid) with unknown continuous distribution H against the change point alternative that there is some (unknown) c with $1 \leq c < n$ such that $X_1, \dots, X_c$ are iid with continuous distribution F and then $X_{c+1}, \dots, X_n$ are iid with some other continuous distribution G. We will consider tests based on two sample empirical distribution function tests for equality of distribution, focusing on the two-sample Cramér–von Mises test.

If the time c of the potential change point were specified in advance we could test the hypothesis that $F = G = H$ using any two sample test for equality of two distributions. The two-sample Cramér–von Mises test is one well-known possibility. Notation may be simpler to read if we used the shorthand $d = n - c$. Let

$$F_c(x) = \frac{1}{c}\sum_{i=1}^{c} 1(X_i \leq x), \tag{1}$$

be the empirical distribution function of the first c observations and

$$G_d(x) = \frac{1}{d}\sum_{i=c+1}^{n} 1(X_i \leq x), \tag{2}$$

be the empirical distribution function of the remaining d observations. The combined empirical distribution function H_n of the entire sample is

$$H_n(x) = \frac{cF_c(x) + dG_d(x)}{n}. \tag{3}$$

The two-sample Cramér–von Mises test of the hypothesis $F = G$ is based on the statistic

$$W_n(X_1, \dots, X_n; c) \equiv W_n(c) = \frac{cd}{n}\int_{-\infty}^{\infty} \{F_c(x) - G_d(x)\}^2 dH_n(x).$$

For a thorough discussion of this nonparametric test and a simple computing formula in terms of the ranks of the first c values of X in the whole sample see Anderson (1962). The distribution of the test statistic does not depend on H under the null hypothesis provided H is a continuous function.

A number of authors have suggested adapting this statistic to the change point problem. See, for instance, Picard (1985) and Brodsky and Darkhovsky (1993) where the two natural possible test statistics considered herein are suggested and studied briefly. The first of these tests can be used both to assess the existence of a change point and to estimate the location of the change if it exists. The statistic in question is

$$W_{\max,n} \equiv \max_{1 \leq c \leq n-1} W_n(c).$$

We shall also use $W_{\max,n}$ to define the estimated change point

$$\hat{c}_n = \underset{1 \leq c \leq n-1}{\arg\max} W_n(c);$$

thus $\hat{c}_n$ is the value of c achieving the maximum. (The statistic $W_n(c)$ is discrete and in small samples there is some modest probability that $\hat{c}_n$ will not be unique; this lack of uniqueness plays no role in the hypothesis testing problem.)

We prefer, however, the statistic

$$\overline{W}_n(X_1, \dots ,X_n) = \overline{W}_n \equiv \frac{1}{n-1}\sum_{c=1}^{n-1} W_n(c).$$

We offer several potential rationales for our choice:

- In many goodness-of-fit contexts quadratic statistics like ours outperform maximal statistics. For instance, the Cramér–von Mises goodness-of-fit test is generally more powerful than the Kolmogorov–Smirnov test; see, for instance, Stephens (1986).
- Quadratic statistics such as we propose often have simpler large sample theory than do maximal statistics like the Kolmogorov–Smirnov test. Generally speaking the former have limiting distributions which are linear combination of chi-squares while the latter have limiting laws which are those of the supremum of a Gaussian process. The actual laws of these suprema are known only in special cases (although inequalities can often provide useful upper bounds on p-values).
- The large sample theory in question often provides a more accurate approximation for quadratic statistics than it does for maximal statistics. For example, see Razali and Yap (2011) and Büning (2002).

In Section 2 we present large sample distribution theory for $\overline{W}_n$ under the null hypothesis. We show how to compute p-values based on this large sample theory and demonstrate that the asymptotic approximations are quite accurate for $n \geq 100$.

In Section 3 we discuss large sample behavior of $W_{\max,n}$ and $\hat{c}_n$. For $W_{\max,n}$ we do not have complete large sample theory; instead we present some evidence that a centered and scaled version of $W_{\max,n}$ has a limiting extreme value distribution. We show that $\hat{c}_n$ tends, when the null holds, to occur near $c = 1$ or near $c = n$. We also show that, when the null holds, $W_{\max,n}$ tends to ∞ in probability. Thus complete large sample theory would require some sort of rescaling.

Section 4 presents a short Monte Carlo power study showing that over a range of alternatives the statistic $\overline{W}_n$ is more powerful than $W_{\max,n}$; exceptions occur when the change point occurs close to the beginning or close to the end of the sequence and sometimes when there is more than 1 change point. Section 5 presents asymptotic power calculations for $\overline{W}_n$ against contiguous sequences of alternatives; these permit useful approximations to the power of $\overline{W}_n$ in cases where the null is not obviously false. By contrast, the limit theory for $W_{\max,n}$ does not lend itself to easy power calculations. We conjecture, however, that in the context of the contiguous alternatives of this section the statistic $W_{\max,n}$ has the defect that, unlike $\overline{W}_n$, its power converges to its level as $n \to \infty$. Also in Section 5, we present some further Monte Carlo studies relevant to contiguous sequences of alternatives. Finally we present some discussion in Section 7. We give proofs and evidence for our conjectures in the Appendix.

2 | NULL LIMIT THEORY: $\overline{W}_n$

Suppose that the null hypothesis holds and the $X_1, \dots ,X_n$ are iid with *continuous* cdf H. Then for all c we have

$$W_n(X_1, \dots ,X_n; c) = W_n(H(X_1), \dots ,H(X_n); c).$$

Thus in computing distribution theory under the null we may, and will, assume that H is the uniform distribution; to emphasize the point we let $U_1, U_2, \ldots$ be an iid sequence of Uniform random variables; the joint law of $(H(X_1), \ldots, H(X_n))$ is the same as that of $(U_1, \ldots, U_n)$.

We now describe large sample theory for our statistics based on weak convergence results for associated processes. We use $\rightsquigarrow$ for weak convergence; where necessary we specify the topology involved more precisely.

Large sample theory for the two sample Cramér–von Mises statistic is well known: if c depends on n in such a way that $c/n \to u \in (0,1)$ (or even just $\min\{c, n-c\} \to \infty$) then

$$W_n(c) \rightsquigarrow \sum_{j=1}^{\infty} \frac{Z_j^2}{\pi^2 j^2},$$

where the Z_i are iid standard normal; see Anderson (1962, p. 1152). (Notice that the limit law does not depend on u.). Our statistic has a related limit given as follows.

Theorem 1. *As $n \to \infty$ we have, under the null hypothesis,*

$$\overline{W}_n \rightsquigarrow \overline{W}_\infty \equiv \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{Z_{jk}^2}{j(j+1)\pi^2 k^2}, \tag{4}$$

where the Z_{jk} are iid standard normal.

The theorem is a consequence, as usual, of a suitable weak convergence result which we now present; the Gaussian process limit we derive is mentioned in Picard (1985, p. 843); the specific weights in Theorem 1 do not seem to have been previously described.

We begin by defining the partial sum empirical process, van der Vaart and Wellner (1996, p. 225), for $(s,t) \in [0,1]^2$, by

$$\mathbb{Z}_n(s,t) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le ns} \{1(U_i \le t) - t\}\,. \tag{5}$$

Our statistic can be described in terms of this process. Notice that

$$F_c(t) = \frac{\sqrt{n}}{c}\mathbb{Z}_n(c/n, t) + t,$$

and that

$$G_d(t) = \frac{\sqrt{n}}{d}\{\mathbb{Z}_n(1,t) - \mathbb{Z}_n(c/n,t)\} + t.$$

Thus

$$F_c(t) - G_d(t) = \sqrt{n}\left\{\frac{\mathbb{Z}_n(c/n,t)}{c} - \frac{\mathbb{Z}_n(1,t) - \mathbb{Z}_n(c/n,t)}{d}\right\}.$$

We now define a process $\mathbb{W}_n(s,t)$ for $0 < s < 1$ and $0 \le t \le 1$ by

$$\mathbb{W}_n(s,t) = \sqrt{s(1-s)}\left\{\frac{\mathbb{Z}_n(s,t)}{s} - \frac{\mathbb{Z}_n(1,t)-\mathbb{Z}_n(s,t)}{1-s}\right\} = \frac{\mathbb{Z}_n(s,t)-s\mathbb{Z}_n(1,t)}{\sqrt{s(1-s)}}. \tag{6}$$

For given c our two-sample test statistic is given by

$$W_n(c) = \int_0^1 \{\mathbb{W}_n(c/n,t)\}^2 \, dH_n(t).$$

The processes $\mathbb{Z}_n$ and $\mathbb{W}_n$ have well-known weak limits given in the following theorem. Details can be found at van der Vaart and Wellner (1996, p. 226). The weak convergence results happen either in $\ell^\infty([0,1]^2)$ or in $\ell^{\infty,\text{loc}}((0,1)\times[0,1])$. The latter space is the set of all functions on $T = (0,1)\times[0,1]$ which are uniformly bounded on every compact subset of T; the space is endowed with the topology of uniform convergence on compacts. This is a special case of the space denoted by $\ell^\infty(T_1, T_2, \dots)$ in van der Vaart and Wellner (1996, pp. 43–44); for us the subsets T_k of T may be taken to be $T_k = [\epsilon_k, 1-\epsilon_k]\times[0,1]$ for some sequence $1/2 > \epsilon_1 > \epsilon_2 > \dots$ which decreases to 0. It will prove useful in some of our proofs to introduce the notation

$$\mathbb{B}_n(s,t) = \mathbb{Z}_n(s,t) - s\mathbb{Z}_n(1,t).$$

Theorem 2. *Under the null hypothesis:*

1. *As $n \to \infty$,*

$$\mathbb{Z}_n \rightsquigarrow \mathbb{Z}_\infty,$$

in $\ell^\infty([0,1]^2)$ where $\mathbb{Z}_\infty$ is a tight mean 0 Gaussian Process with continuous sample paths and covariance function

$$\rho_Z(s,t;s',t') = s \wedge s'\psi(t,t'),$$

where $\psi(t,t') = t\wedge t' - tt'$;

2. *As $n \to \infty$,*

$$\mathbb{B}_n \rightsquigarrow \mathbb{B}_\infty,$$

in $\ell^\infty([0,1]^2)$ where $\mathbb{B}_\infty$ is the tight mean 0 Gaussian Process defined for $0 \le s,t \le 1$ by

$$\mathbb{B}_\infty(s,t) = \mathbb{Z}_\infty(s,t) - s\mathbb{Z}_\infty(1,t).$$

The process $\mathbb{B}_\infty$ has continuous sample paths and covariance function

$$\rho_B(s,t;s',t') = \psi(s,s')\psi(t,t').$$

3. *As $n \to \infty$,*

$$\mathbb{W}_n \rightsquigarrow \mathbb{W}_\infty,$$

in $\ell^{\infty,\text{loc}}((0,1)\times[0,1])$ where for $0 < 1 < 1$ and $0 \le t \le 1$, we have

$$\mathbb{W}_\infty(s,t) = \frac{\mathbb{B}(s,t)}{\sqrt{s(1-s)}}.$$

The process $\mathbb{W}_\infty$ is a mean 0 Gaussian Process with covariance function

$$\rho_W(s,t;s',t') = \chi(s,s')\psi(t,t'),$$

where

$$\chi(s,s') = \frac{\psi(s,s')}{\sqrt{s(1-s)s'(1-s')}}. \tag{7}$$

The restriction of $\mathbb{W}_\infty$ to a compact $K \subset (0,1) \times [0,1]$ is tight in $\ell^\infty(K)$.

The process $\mathbb{B}$ is called a Brownian pillow by some writers or a four-side tied down Brownian motion; see, for instance Zhang (2014) or McKeague and Sun (1996). The process $\mathbb{Z}$ is a Blum-Kiefer-Rosenblatt process; see Blum et al. (1961).

We now record well-known facts about the eigenvalues of the covariance ρ_W. The covariance kernel ψ is that of a Brownian Bridge. It has eigenvalues of the form $1/(\pi^2k^2)$ for $k = 1, 2, \ldots$ with corresponding orthonormal eigenfunctions $f_{\psi,k}(u) = \sqrt{2}\sin(\pi k u)$. The covariance kernel χ arises in the study of the Anderson–Darling goodness-of-fit test. It has eigenvalues of the form $1/\{j(j+1)\}$ for $j = 1, 2, \ldots$. The corresponding orthonormal eigenfunctions are associated Legendre functions. The jth eigenfunction is

$$f_{\chi,j}(u) = 2\sqrt{\frac{2j+1}{j(j+1)}}\sqrt{s(1-s)}q_j(2s-1),$$

where the q_j are polynomials of degree $j-1$ defined recursively as follows:

$$q_1(u) = 1, \quad q_2(u) = 3u,$$

and for $j \geq 2$

$$q_{j+1}(u) = \frac{1}{j}\left\{(2j+1)uq_j(u) - (j+1)q_{j-1}(u)\right\}.$$

The covariance function ρ_W is the tensor product of χ and ψ. It follows that the eigenvalues of ρ_W consist of all possible products

$$\lambda_{jk} = \frac{1}{j(j+1)\pi^2k^2}, \tag{8}$$

with corresponding eigenfunctions

$$f_{\chi,j}(s)f_{\psi,k}(t).$$

The expansion (4) in Theorem 1 is then Parseval's identity with

$$Z_{jk} = \int_0^1 \int_0^1 \mathbb{W}(s,t) f_{\chi,j}(s) f_{\psi,k}(t)\, ds\, dt.$$

2.1 | Numerical work

The distribution of $\overline{W}_\infty$ can be computed numerically to provide approximate, asymptotically valid, p-values. Our desired approximation to the p-value is

$$P(\overline{W}_n > w_{\text{obs}}) \approx P(\overline{W}_\infty > w_{\text{obs}}),$$

where w_{obs} is the value of $\overline{W}_n$ observed in the data. In practice, we truncate the infinite sum in (4) defining $\overline{W}_\infty$, retaining the terms with the largest values of λ_{jk}, and replace the neglected terms by their expected value. So we write

$$\begin{aligned}\overline{W}_\infty &= \overline{W}_{\infty,M} + T_{\infty,M} \\ &= \sum_{jk \le M} \lambda_{jk} Z_{jk}^2 + \sum_{jk > M} \lambda_{jk} Z_{jk}^2.\end{aligned}$$

We then approximate $T_{M,\infty}$ by its expected value:

$$\mu_M \equiv \sum_{jk>M} \lambda_{jk} \mathrm{E}\left(Z_{jk}^2\right) = \sum_{jk>M} \lambda_{jk}.$$

Since the mean of $\overline{W}_\infty$ is

$$\sum_{j,k} \lambda_{jk} = \frac{1}{6},$$

the mean of $T_{M,\infty}$ may be computed by

$$\frac{1}{6} - \sum_{jk \le M} \lambda_{jk}.$$

Our approximation becomes

$$P(\overline{W}_n > w_{\text{obs}}) \approx P(\overline{W}_{M,\infty} + \mu_M > w_{\text{obs}}).$$

The latter quantity may now be computed by using numerical Fourier inversion following Imhof (1961). The `R` package `CompQuadForm` (see Duchesne & de Micheaux, 2010) implements this computation in the function `imhof`; we use this software in our numerical work below retaining the 1024 largest eigenvalues and adjusting the mean for truncation as described above.

We have evaluated the quality of our asymptotic approximation to the null distribution of $\overline{W}_n$ in a small Monte Carlo study. Since this distribution does not depend on H when the null hypothesis holds we generated $N = 100{,}000$ samples of size $n = 100, 200, 500$. Figure 1 shows a Q-Q plot for these 100,000 values for $n = 100$ to check the uniformity of their distribution. Specifically, we

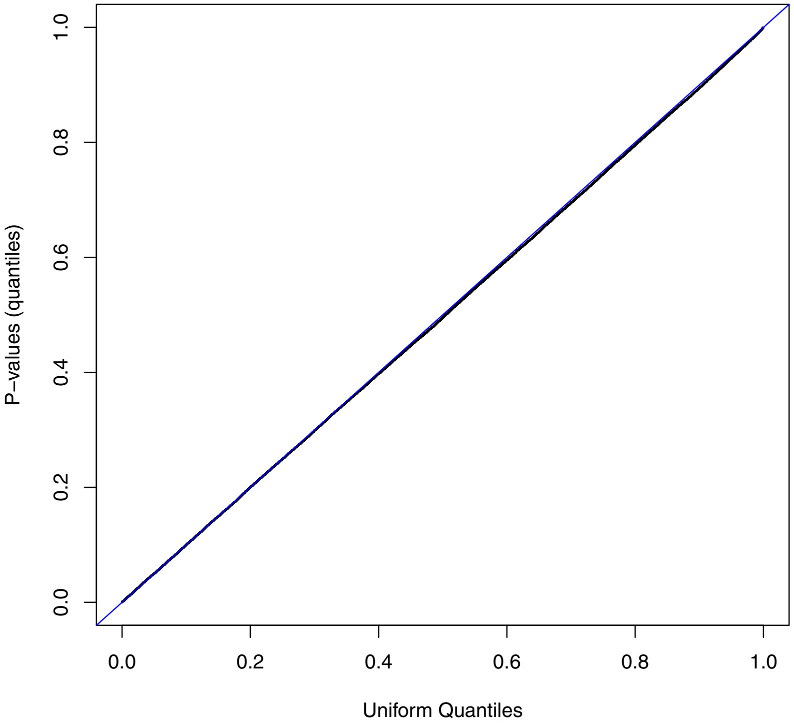

FIGURE 1 Ordered p-values simulated under the null hypothesis for sample size $n = 100$ and plotted against uniform quantiles for 100,000 iid Monte Carlo samples from a continuous distribution. The blue line is the uniform cumulative distribution function; exact p-values have a uniform distribution. The graph shows this approximation is extremely good [Colour figure can be viewed at wileyonlinelibrary.com]

plot the order statistics against the uniform plotting points $1/(N+1), \dots, N/(N+1)$. Our asymptotic approximation is clearly excellent; the same pattern is observed for all the larger values of n we tried.

3 | NULL DISTRIBUTION: $W_{\max,n}$ AND $\hat{c}_n$

We do not have complete large sample distribution theory for $W_{\max,n}$. The statistic $W_{\max,n}$ is more challenging to analyze because the weak convergence result in Theorem 2 asserts convergence in $\ell^{\infty,\mathrm{loc}}((0,1)\times[0,1])$ and the supremum norm is not necessarily finite even for a continuous function belonging to $\ell^{\infty,\mathrm{loc}}((0,1)\times[0,1])$. Our proof of Theorem 1 shows that our statistic, $\overline{W}_n$, is a continuous functional on a subset of $\ell^{\mathrm{loc}}_{\infty}((0,1)\times[0,1])$ to which sample paths of $\mathbb{W}_\infty$ are almost sure to belong. We are not able to establish the corresponding result for $W_{\max,n}$. Traditionally this problem has been handled either by fixing a small $\epsilon > 0$ and redefining $W_{\max,n}$ by maximizing only over $\{c : \epsilon \le c/n \le 1-\epsilon\}$ or by careful analysis of the behavior of the process and the test statistic for c/n close to 0 or to 1. For instance, Jaeschke (1979) considers a weighted Kolmogorov–Smirnov test for the uniform distribution and shows that the supremum of the weighted empirical process has, after suitable normalization, an extreme value distribution. We conjecture that this happens here, too.

Conjecture 1. *There are constants a_n and b_n with a_n and b_n both going to ∞ such that*

$$\frac{W_{\max,n} - b_n}{a_n} \rightsquigarrow E,$$

where E is an extreme value random variable with

$$P(E \le x) = \exp\{-\exp(-x)\}.$$

There are several motivations for the conjecture. The first lies in the analogy cited to Jaeschke's work. Jaeschke (1979) shows that the supremum over [0, 1] of the usual empirical process divided by its pointwise SD, namely,

$$\sup_{0\le x\le 1} \frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}},$$

is achieved near x=0 or near $x = 1$. He then uses this and the behavior of a similarly scaled Brownian Bridge to deduce the extreme value limit.

We examined $W_{\max,n}$ in a small Monte Carlo study. First, for various values of n we examined plots of the order statistics, $Y_{(i)}, i = 1, \dots, M$, of the values of $W_{\max,n}$ computed from $M = 100{,}000$ Monte Carlo samples of size n against the quantiles $-\log(-\log(i/(M + 1)))$ of the extreme value distribution. It will be seen in Figure 2 that the plot is quite linear; we note, moreover, that about 90% of the 100,000 plotted points actually touch the line in the figure. We have used ordinary least squares to add a line in blue whose slope and intercept would lead to reasonable values for the parameters a_n and b_n for this value of n.

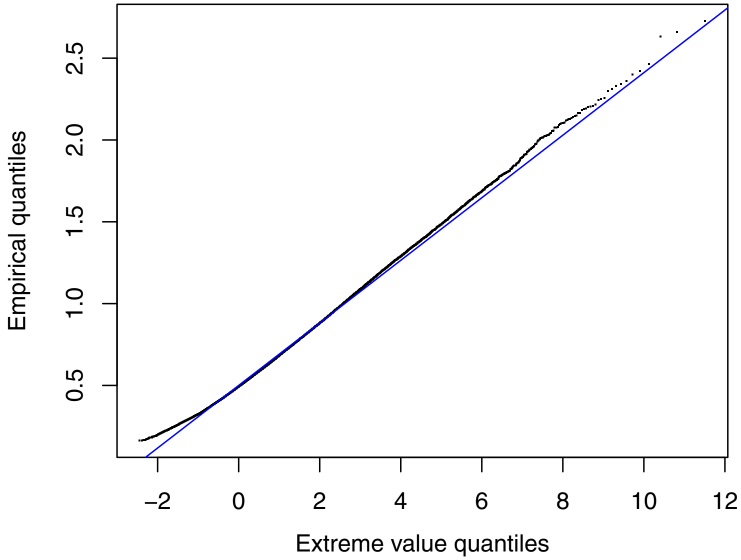

FIGURE 2 Quantile-quantile plot of the P-values for $W_{\max,n_n}$ using $M = 100{,}000$ Monte Carlo samples under the null hypothesis against quantiles for the standard extreme value distribution. The ith order statistic of the 100,000 P-values is plotted against $-log(-log(i/(M + 1)))$. The blue line was fitted using ordinary least squares [Colour figure can be viewed at wileyonlinelibrary.com]

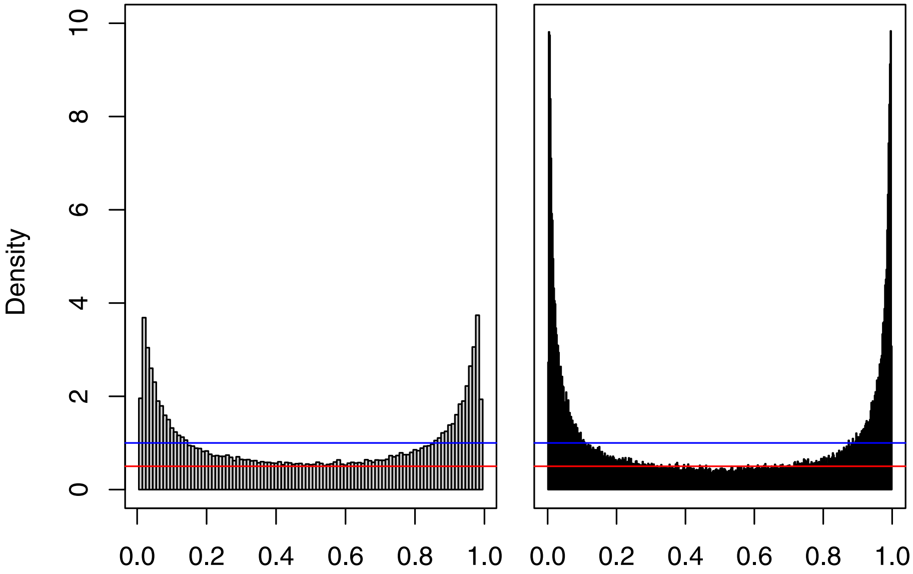

FIGURE 3 Histograms of values of estimated change points for sample sizes $n = 100$ on the left and $n = 500$ on the right. The null hypothesis is true and 100,000 samples were used for each histogram. The x-axis shows $\hat{c}/n$ and the y-axis is a probability density scale. We have used one bin for each possible value of $\hat{c}_n$. The two figures have the same scales on each axis. Horizontal lines at height 1 (blue) and 0.5 (red) are provided to help see the extent to which the distribution on the right is more concentrated around 0 and 1 than the distribution on the left [Colour figure can be viewed at wileyonlinelibrary.com]

We then looked at the law of $\hat{c}_n$ in a similar Monte Carlo study. We generated 100,000 samples of size 100 and 500 from the null hypothesis. In Figure 3 we plot histograms of the value $\hat{c}$ which maximizes $W_n(c)$ over $1 \le c \le n - 1$. Observe that as the sample size grows the histogram concentrates near 0 and 1 (though the convergence is slow). We prove in the Appendix:

Proposition 1. *Under the null hypothesis we have:*

$$\min\left\{\frac{\hat{c}}{n}, \frac{n-\hat{c}}{n}\right\} \to 0 \ \textit{in probability}; \tag{9}$$

and

$$\hat{c}/n \rightsquigarrow \textit{Bernoulli}(0.5); \tag{10}$$

and

$$W_{\max,n} \to +\infty \ \textit{in probability}. \tag{11}$$

The last assertion implies that any asymptotic limit law for $W_{\max,n}$ will require recentering or rescaling or both. For some evidence concerning how slow this convergence is, we record in Table 3 the mean and SD of $W_{\max,n}$ for various values of n, based on $M = 100{,}000$ Monte Carlo samples for each case. It will be seen that both the mean and the SDs are rising slowly. The Monte Carlo SE in the mean is 3 or 4 in the fourth digit, roughly, so the observed differences are real, though they are, of course, small.

The following table displays the mean and SD of $W_{\max,n}$ for various sample sizes n based on $M = 100{,}000$ null hypothesis samples.

	$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
Mean	0.373	0.487	0.529	0.565	0.610
SD	0.145	0.208	0.225	0.234	0.245

We have not pursued the task of proving a limit theorem for $W_{\max,n}$; it may be worth quoting from Jaeschke (1979) who says the convergence rate to the extreme value law is so slow that "we would not encourage anyone to use the confidence intervals based on the asymptotic analysis." We ourselves use Monte Carlo critical points in the studies which follow.

4 | MONTE CARLO POWER STUDY

We undertook a variety of Monte Carlo simulation studies to compare the power of $\overline{W}_n$ to that of $W_{\max}$. In Table 1 we show the percentage of samples rejected in 10,000 trials by the two methods at the levels $\alpha = 0.05$ and $\alpha = 0.1$. We consider samples of size $n \in \{20, 50, 100\}$. In each case we used Monte Carlo critical values based on 100,000 Monte Carlo samples. Powers for $\overline{W}$ using our asymptotic approximation are very similar.

In one experiment recorded in the table we generated data from the Gamma distributions where the parameters change at $c = n/2$. In another experiment we change from the Gamma distribution to the Normal distribution at $c = n/2$; in this case neither the mean nor the variance changes. While our tests are designed to detect single change points we have included two trials in which there are three segments which change between various Gamma distributions. One changes from shape 1, scale 2 to shape 2, scale 1 at the 40% point and then to shape 0.5, scale 4 at the 60% point. All three of these have the same mean. The other changes from shape 1, scale 2 to shape 2, scale 3, and back to shape 1, scale 2; the changes happen after 30% and then 70% of the data. Finally we present two experiments with samples from the normal distribution; in one the mean changes at $c = n/2$ and in the other the SD changes at the same point. In all these trials the parameter values in the distributions in a given segment do not change as the sample size changes; this may be compared with the further Monte Carlo results in Section 5.

It will be seen that, except for very small samples, when there is a single change point the test using $\overline{W}_n$ has better power than $W_{\max,n}$ unless the change point is near the one end or the other of the series. Since it is also far faster to compute p-values for $\overline{W}_n$ using the highly accurate asymptotic law we recommend $\overline{W}_n$ over $W_{\max,n}$ except if one has a priori (i.e., of course, before looking at the data) reason to expect the change to be near one end or the other. At the same time we observe that the procedure is specifically designed to choose between 1 change point and no change points and not to estimate and find multiple change points. In particular, for one of the alternatives in Table 1 with 2 change points the statistic $W_{\max,n}$ is usually more sensitive than $\overline{W}_n$. The last alternative demonstrates that when the change happens in the end of the sample, test statistic $W_{\max,n}$ performs better in detecting it.

The results presented here show how the powers grow with sample size when the two distributions are fixed. More Monte Carlo power calculations are presented in Section 6 below with a focus on contiguous alternatives.

TABLE 1 Powers (percentage) from various alternative distributions and significance levels 0.1 and 0.05. Critical points were calculated with 100,000 and Powers by 10,000 Monte Carlo simulations. The notation Gamma(α, β) indicates sampling from a Gamma distribution with shape α and scale β. The parameters in the normal distribution are mean and variance. The exponential distribution parameter is the mean

		$\alpha = 0.1$		$\alpha = 0.05$	
Alternative	**Sample size**	$W_{\max}$	$\overline{W}_n$	$W_{\max}$	$\overline{W}_n$
$X_1, \dots, X_{0.5n} \sim$ Gamma$(1, 2)$,	$n = 20$	47.9	50.7	35.0	37.5
$X_{0.5n+1}, \dots, X_n \sim$ Gamma$(2, 2)$	$n = 50$	82.3	85.7	73.9	77.4
	$n = 100$	98.3	98.9	96.3	96.9
$X_1, \dots, X_{0.5n} \sim$ Gamma$(1, 2)$,	$n = 20$	12.9	13.7	6.9	7.2
$X_{0.5n+1}, \dots, X_n \sim \mathcal{N}(2, 2)$	$n = 50$	16.1	19.2	9.0	11.2
	$n = 100$	22.1	31.2	13.7	19.0
$X_1, \dots, X_{0.4n} \sim$ Gamma$(1, 2)$,	$n = 20$	17.5	16.5	10.0	9.2
$X_{0.4n+1}, \dots, X_{0.6n} \sim$ Gamma$(2, 1)$	$n = 50$	24.6	25.5	15.5	15.9
$X_{0.6n+1}, \dots, X_n \sim$ Gamma$(0.5, 4)$	$n = 100$	38.3	42.8	27.3	28.5
$X_1, \dots, X_{0.3n} \sim$ Gamma$(1, 2)$,	$n = 20$	29.0	20.6	15.8	7.9
$X_{0.3n+1}, \dots, X_{0.7n} \sim$ Gamma$(2, 3)$	$n = 50$	72.3	71.6	54.4	48.1
$X_{0.7n+1}, \dots, X_n \sim$ Gamma$(1, 2)$	$n = 100$	98.3	98.6	94.1	94.6
$X_1, \dots, X_{0.5n} \sim \mathcal{N}(0, 1)$,	$n = 20$	18.2	22.0	10.8	11.3
$X_{0.5n+1}, \dots, X_n \sim \mathcal{N}(0, 3)$	$n = 50$	29.6	56.0	17.0	33.0
	$n = 100$	66.3	93.4	45.0	81.2
$X_1, \dots, X_{0.9n} \sim Exp(1)$,	$n = 20$	15.3	16.2	9.9	11.2
$X_{0.9n+1}, \dots, X_n \sim Exp(5)$	$n = 50$	57.4	41.4	42.5	26.1
	$n = 100$	95.9	80.2	90.4	56.5

5 | POWER: CONTIGUOUS ALTERNATIVES

We now compute approximate distribution theory for $\overline{W}_n$ when the null hypothesis is false and the extent of the change at the change point is big enough to be detectable but not obvious; that is, we study situations where the best possible power in large samples stays away from 1. To do so we consider a sequence of alternatives indexed by n and assume that these alternatives are contiguous to a sequence for which the null hypothesis of no change holds. To be specific our null hypothesis sequence will have X_i iid for $1 \leq i \leq n$ with density h_n and cdf H_n. For the alternative we suppose that there is a value c_0 such that for $1 \leq i \leq c_0$, the X_i are iid with density f_n and that for $c_0 + 1 \leq i \leq n$ the X_i are iid with density g_n. All of f_n, g_n, h_n, and the true change point $c_{0,n}$ may depend on n. Under the null hypothesis the joint density of $X_1, \dots, X_n$ is

$$\mathbf{f}_{0n}(x_1, \dots, x_n) = \prod_{i=1}^{n} h_n(x_i).$$

Under the alternative the joint density becomes

$$\mathbf{f}_{1n}(x_1, \dots, x_n) = \prod_{i=1}^{c_{0,n}} f_n(x_i) \prod_{c_{0,n}+1}^{n} g_n(x_i).$$

The log-likelihood ratio of these two is

$$\begin{aligned}\Lambda_n &= \ln\left\{\mathbf{f}_{1,n}(X_1, \dots, X_n)/\mathbf{f}_{0n}(X_1, \dots, X_n)\right\}\\ &= \sum_{i=1}^{c_{0,n}} \ln\left\{f_n(X_i)/h_n(X_i)\right\} + \sum_{i=c_{0,n}+1}^{n} \ln\left\{g_n(X_i)/h_n(X_i)\right\}.\end{aligned}$$

The sequence of alternatives $\mathbf{f}_{1n}$ is contiguous to the null sequence $\mathbf{f}_{0n}$ if, computing under the null hypothesis, we have

$$\Lambda_n \rightsquigarrow N(-\tau^2/2, \tau^2), \tag{12}$$

for some $0 \le \tau < \infty$. If we define $U_i = H_n(X_i)$ then under the null hypothesis the U_i are iid Uniform[0,1]. Under the alternative $U_1, \dots, U_{c_{0,n}}$ are iid with density $\tilde{f}_n(u) = f_n(H_n^{-1}(u))/h_n\left(H_n^{-1}(u)\right)$ while $U_{c_{0,n}+1}, \dots, U_n$ are iid with density $\tilde{g}_n(u) = g_n(H^{-1}(u))/h_n\left(H_n^{-1}(u)\right)$. The likelihood ratio becomes

$$\tilde{\Lambda}_n = \sum_{i=1}^{c_{0,n}} \ln\left\{\tilde{f}_n(U_i)\right\} + \sum_{i=c_{0,n}+1}^{n} \ln\left\{\tilde{g}_n(U_i)\right\}.$$

Theorem 3. *Assume*

A1 *There are two functions ϕ_f and ϕ_g in $L_2[0,1]$ such that*

$$\lim_{n\to\infty} \sqrt{n}(\tilde{f}_n - 1) = \phi_f,$$

and

$$\lim_{n\to\infty} \sqrt{n}(\tilde{g}_n - 1) = \phi_g,$$

in $L_2[0,1]$.

A2 *There is a $u \in (0,1)$ such that*

$$\lim_{n\to\infty} \frac{c_{0,n}}{n} = u.$$

Then as $n \to \infty$ we have, under the sequence of alternative hypotheses specified by f_n, g_n, and $c_{0,n}$,

1. *The log-likelihood ratio satisfies*

$$\Lambda_n = S_n + o_P(1) \rightsquigarrow N(-\tau^2/2, \tau^2),$$

where

$$S_n = \sum_{i=1}^{c_0} \phi_f(U_i)/\sqrt{n} + \sum_{i=c_0+1}^{n} \phi_g(U_i)/\sqrt{n}, \tag{13}$$

and

$$\tau^2 = u\int_0^1 \phi_f^2(t)\,dt + (1-u)\int_0^1 \phi_g^2(t)\,dt.$$

2. *The process* $\mathbb{W}_n$ *converges weakly in* $\ell^{\infty,\mathrm{loc}}((0,1)\times[0,1])$ *to a Gaussian process with continuous sample paths, covariance* ρ*, and mean*

$$\mu(s,t) = \mu_\chi(s)\mu_\psi(t),$$

where

$$\mu_\chi(s) = \sqrt{s(1-s)}\left\{\frac{1-u}{1-s}1(s\le u) + \frac{u}{s}1(s>u)\right\},$$

and

$$\mu_\psi(t) = \left[\mathrm{E}\left\{\phi_f(U)1(U\le t)\right\} - \mathrm{E}\left\{\phi_g(U)1(U\le t)\right\}\right].$$

3. *and*

$$\overline{W}_n \rightsquigarrow \overline{W}_\infty \equiv \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{\left(Z_{jk}+\eta_j\tau_k\right)^2}{j(j+1)\pi^2k^2},$$

where the Z_{jk} *are iid standard normal,*

$$\eta_j = \int_0^1 \mu_\chi(s)f_{j,\chi}(s)\,ds,$$

and

$$\tau_k = \int_0^1 \mu_\psi(t)f_{j,\psi}(t)\,dt.$$

As with the null distribution, this limiting alternative distribution for $\overline{W}_n$ can be computed using the R package CompQuadForm. As an example we take f_n to be standard normal and g_n to be normal with mean μ_n and SD σ_n. The two parameters are assumed to depend on n in such a way that

$$\sqrt{n}\mu_n \to \gamma_1 \quad \text{and} \quad \sqrt{n}(\sigma_n - 1) \to \gamma_2.$$

It is convenient to take $h_n = f_n$. Under the null the data $X_1, \dots, X_n$ are iid standard normal. The functions $\tilde{f}_n$ and $\tilde{g}_n$ are then given by $\tilde{f}_n \equiv 0$ and

$$\tilde{g}_n(u) = \frac{\phi\left\{\frac{\Phi^{-1}(u)-\mu_n}{\sigma_n}\right)}{\phi\left\{\Phi^{-1}(u)\right\}}.$$

Under these conditions we may check that condition **A1** holds with $\phi_f = 0$ and

$$\phi_g(u) = \gamma_2\left[\left\{\Phi^{-1}(u)\right\}^2 - 1\right] + \gamma_1\Phi^{-1}(u).$$

6 | LARGE SAMPLE BEHAVIOR OF $W_{\max,n}$

Proposition 1 shows that under the null hypothesis

$$\min\left\{\frac{\hat{c}_n}{n}, \frac{n-\hat{c}_n}{n}\right\} \to 0,$$

in probability. By the definition of contiguity this conclusion also holds under any sequence of contiguous alternatives. This means that, even for data from detectable (but not obvious) alternatives, our test statistic $W_{\max,n}$ usually compares the distribution of a tiny fraction of the data to that of the vast majority of the data even when the true change point is in the middle of the sequence. We note, however, that the rate of convergence for this assertion appears from Monte Carlo work to be quite slow so that the practical impact of the conclusion must be assessed more carefully.

Lockhart (1991) shows that the statistic studied in Jaeschke (1979) has the property that the difference between the power and the level of the corresponding test converges to 0 for any sequence of contiguous alternatives. This motivates us to conjecture:

Conjecture 2. *For any sequence of contiguous alternatives the difference between the power and the level of a test based on $W_{\max,n}$ goes to 0 as $n \to \infty$.*

In the Appendix we provide partial details showing how we would hope to prove our conjecture, if we could.

Here is some Monte Carlo evidence from a simulation study. In Tables 2 and 3 we study four alternatives at sample sizes $n = 10, 50, 100, 200, 500$. For each sample size we draw 10,000 samples of size n. The first c observations in each sample have some parameter of the form $a + b/\sqrt{n}$ and the remaining $n - c$ have parameter a. We used the Gamma distribution and the normal distribution and tried $c = 0.5n$ and $c = 0.3n$ for each distribution. In the Gamma case we tried changing the shape parameter with $a = 1$ while holding the scale parameter at 1. The tables show the expected convergence (although we have not computed the power predicted by our theory in Section 5).

For the statistic $W_{\max,n}$ the tables show, in the normal case, the power declining towards the level (which is 5% here). For the Gamma cases studied here the power is rising but slowly for distant alternatives (large values of b) and declining very slowly for less distant alternatives (smaller values of b). Our experience in general is that for more distant alternatives it requires larger (sometimes much larger) sample sizes before the power of $W_{\max,n}$ begins to drop.

TABLE 2 Powers (percentage) for change from Gamma (shape $= 1 + b/\sqrt{n}$, scale=1) to Gamma(1,1) at the indicated breakpoint, $n/2$ in the top and $3n/10$ in the bottom. Powers are based on 10,000 samples and either use Monte Carlo critical points (based on 100,000 samples) or asymptotic critical points as indicated by 'MC' or 'Asym'. All tests are at the level $\alpha = 0.05$

Gamma, shape = $1 + b/\sqrt{n}$, break at $n/2$							
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
$b = 2$	$\overline{W}_n$	MC	11.70	13.96	14.83	14.71	15.91
	$\overline{W}_n$	Asym	11.79	13.59	14.61	14.67	15.70
	$W_{\max,n}$	MC	12.13	12.00	12.36	11.41	11.80
$b = 3$	$\overline{W}_n$	MC	18.50	25.18	26.48	27.74	29.52
	$\overline{W}_n$	Asym	18.72	24.73	26.12	27.66	29.25
	$W_{\max,n}$	MC	18.62	22.05	21.84	21.34	21.88
$b = 5$	$\overline{W}_n$	MC	34.95	52.67	57.39	61.28	65.62
	$\overline{W}_n$	Asym	35.26	51.97	57.06	61.18	65.35
	$W_{\max,n}$	MC	35.48	47.60	50.07	52.76	54.46
Gamma, shape = $1 + b/\sqrt{n}$, break at $3n/10$							
$b = 2$	$\overline{W}_n$	MC	9.24	11.29	11.73	11.83	13.21
	$\overline{W}_n$	Asym	9.42	10.86	11.47	11.79	13.10
	$W_{\max,n}$	MC	10.00	10.60	10.48	9.86	10.37
$b = 3$	$\overline{W}_n$	MC	13.41	20.04	20.26	21.80	23.15
	$\overline{W}_n$	Asym	13.54	19.56	19.92	21.66	22.98
	$W_{\max,n}$	MC	14.81	18.07	17.38	18.00	17.97
$b = 5$	$\overline{W}_n$	MC	22.42	41.53	45.54	48.59	53.34
	$\overline{W}_n$	Asym	22.75	40.87	45.11	48.52	53.07
	$W_{\max,n}$	MC	26.43	39.44	41.36	43.09	45.72

7 | DISCUSSION

It is a general principle that procedures with optimal frequency properties are found by searching among Bayes procedures. It is also generally the case that optimal Bayes procedures involve averaging rather than maximizing. These heuristics motivate considering testing for change points by using test statistics which are averages over possible change points rather than maxima. In this paper we have used this heuristic to motivate an average two-sample goodness-of-fit statistic when we are concerned about general changes in distribution, rather than simple changes in mean, in a sequence of independent data points. We have shown the resulting test statistic has computable large sample theory which can be used to provide very accurate p-values. Moreover we have shown that averaging over possible change points is generally more sensitive to alternatives than maximizing over possible change points.

Exceptions to the last conclusion arise when it is suspected a priori that a change might occur near the end of the interval. In this case one might prefer to modify the statistics to focus more weight near the ends of the interval. The statistic $W_{\max,n}$ already focuses substantially on the ends

TABLE 3 Powers (percentage) for change from Normal($0,\sigma = 1 + b/\sqrt{n}$) to Normal(0,1) at the indicated breakpoint, namely, $n/2$ in the top and $3n/10$ in the bottom. Powers are based on 10,000 samples and either use Monte Carlo critical points (based on 100,000 samples) or asymptotic critical points as indicated by 'MC' or 'Asym'. All tests are at the level $\alpha = 0.05$

Normal, $\sigma = 1 + b/\sqrt{n}$, break at $n/2$							
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
$b = 2$	$\overline{W}_n$	MC	5.61	5.97	5.65	5.66	5.91
	$\overline{W}_n$	Asym	5.69	5.77	5.40	5.61	5.83
	$W_{\max,n}$	MC	6.70	5.72	5.19	4.80	5.25
$b = 3$	$\overline{W}_n$	MC	6.11	7.04	6.87	6.75	7.40
	$\overline{W}_n$	Asym	6.20	6.66	6.66	6.73	7.23
	$W_{\max,n}$	MC	7.67	6.49	5.71	5.36	5.55
$b = 5$	$\overline{W}_n$	MC	6.76	9.55	11.10	11.32	13.56
	$\overline{W}_n$	Asym	6.79	9.24	10.79	11.25	13.33
	$W_{\max,n}$	MC	8.99	7.91	6.99	6.84	6.88
Normal, $\sigma = 1 + b/\sqrt{n}$, break at $0.3n/10$							
$b = 2$	$\overline{W}_n$	MC	6.26	6.49	5.80	5.63	5.76
	$\overline{W}_n$	Asym	6.37	6.17	5.63	5.63	5.68
	$W_{\max,n}$	MC	7.12	6.08	5.72	5.22	5.42
$b = 3$	$\overline{W}_n$	MC	6.91	7.37	6.74	6.41	6.95
	$\overline{W}_n$	Asym	7.09	7.10	6.51	6.39	6.80
	$W_{\max,n}$	MC	8.18	7.08	6.29	5.94	5.95
$b = 5$	$\overline{W}_n$	MC	7.89	9.40	9.92	9.91	11.13
	$\overline{W}_n$	Asym	8.09	8.99	9.65	9.79	10.98
	$W_{\max,n}$	MC	9.80	8.96	8.04	7.67	7.19

but the average involved in $\overline{W}_n$ could have weights added which are larger for c near 1 or near n. We have not tried this though the same sort of large sample theory would apply but with a different kernel replacing ξ (see (7)); note, however, that the theory would require the new kernel to be square integrable over the unit square.

It would be natural to investigate, as a referee has suggested, alternative sequences with F and G fixed and c/n (or $(n - c)/n$) getting smaller with n. We would clearly expect $W_{\max,n}$ to outperform $\overline{W}_n$ in such a limit.

The basic idea of averaging proposed here can be used in other contexts. Consider, for instance, testing for a change in mean. We describe first the unrealistic situation in which the SD is known and then how to handle estimation of that SD. Suppose $X_1, \ldots, X_n$ are independent and we wish to test the null hypothesis that they are iid with unknown mean μ and known SD σ (which we take to be 1 for notational convenience) against the alternative that the mean changes after the data point number c. The usual Z statistic is

$$T_c = \left(\frac{X_1 + \cdots + X_c}{c} - \frac{X_{c+1} + \cdots + X_n}{n-c} \right) / \sqrt{\frac{1}{c} + \frac{1}{n-c}}.$$

Our proposal would be to use the two sided test

$$\overline{T^2} = \frac{1}{n-1} \sum_{c=1}^{n} T_c^2.$$

This statistic has mean 1 under the hypothesis of no change in mean. Arguments similar to those in Section 2 show this statistic has the same limiting distribution, under the null, as the well known Anderson–Darling goodness-of-fit statistic.

In the more reasonable case where the (assumed common) SD is unknown will use the statistic

$$T_s^2 = \overline{T^2}/s^2,$$

where s^2 is some estimate of σ^2 which is consistent under the null hypothesis. The sample SD is one possibility though this can be badly biased under the alternative. An estimate which is rather less precise but still likely to be quite accurate under the alternative hypothesis is

$$s_1^2 = \frac{\sum_{i=1}^{n-1} (X_{i+1} - X_i)^2}{2(n-1)}.$$

Notice that under the alternative hypothesis all but one term in this average is an unbiased estimate of σ^2; the bias in the estimator is $\Delta_\mu^2/(2n)$ where Δ_μ denotes the change in the mean at the true change point. Under the null the estimate s_1^2 is unbiased. The statistic T_s^2 also has the same limiting distribution as the well known Anderson–Darling goodness-of-fit statistic when the null holds.

Other nonparametric goodness of fit tests can be used instead of the Cramér–von Mises test. For example a Bayesian test Al Labadi et al. (2014), likelihood tests Csörgö and Horváth (1997) or other two-sample tests Büning (2002). Sample size, the kind of alternative distribution from which we expect the data to come and the expected index of the change point should likely be used to choose the best test in any particular context. Finding the asymptotic distribution for less well-known tests can be difficult. Bootstrapping can be used instead. This deserves further research.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and two reviewers for their valuable comments and suggestions, which led to an improved paper.

ORCID

Rasmus Erlemann https://orcid.org/0000-0002-4120-2560

REFERENCES

Al Labadi, L., Masuadi, E., & Zarepour M. 2014. Two-sample Bayesian nonparametric goodness-of-fit test, *arxiv:1411.3427*.

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *Annals of Mathematical Statistics*, *33*, 1148–1159.

Blum, J. R., Kiefer, J., & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, *32*(2), 485–498. https://doi.org/10.1214/aoms/1177705055

Brodsky, E., & Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*. Springer.

Büning, H. (2002). Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramér-von Mises two-sample tests. *Journal of Applied Statistics*, *29*(6), 907–924.

Csörgö, M., & Horváth, L. (1997). *Limit theorems in change-point analysis Wiley Series in Probability and Statistics*. Wiley.

Duchesne, P., & de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, *54*, 858–862.

Guttorp, P., & Lockhart, R. A. (1988). On the asymptotic distribution of quadratic forms in uniform order statistics. *The Annals of Statistics*, *16*, 433–449.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, *48*, 419–426.

D. Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Annals of Statistics*, *7*(1):108–115, 01 1979.

Lockhart, R. A. (1991). Overweight tails are inefficient. *The Annals of Statistics*, *19*(4), 2254–2258.

McKeague, I. W., & Sun, Y. (1996). Transformations of Gaussian random fields to Brownian sheet and nonparametric change-point tests. *Statistics & Probability Letters*, *28*(4), 311–319. https://doi.org/10.1016/0167-7152(95)00140-9

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, *17*(4), 841–867.

Razali, N. M., & Yap, B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(01), 21–33.

Stephens, M. A. (1986). *Ch. 4 Tests based on EDF statistics*. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques* (pp. 97–193). Marcel Dekker.

van der Vaart, A. W., & Wellner, J. (1996). *Weak convergence and empirical processes: With applications to statistics Springer Series in Statistics*. Springer.

Zhang, T. (2014). A Kolmogorov-Smirnov type test for independence between marks and points of marked point processes. *Electronic Journal of Statistics*, *8*(2), 2557–2584. https://doi.org/10.1214/14-EJS961

How to cite this article: Erlemann, R., Lockhart, R., & Yao, R. (2021). Cramér-von Mises tests for change points. *Scandinavian Journal of Statistics*, 1–29. https://doi.org/10.1111/sjos.12544

APPENDIX

Proof of Theorems 1 and 2. We prove Theorem 2 first. The weak limit $\mathbb{Z}$ given below is discussed in Picard (1985) but we provide a few details for completeness. First, the process $\mathbb{Z}_n$ converges weakly in $\ell^\infty([0,1]^2)$ to a tight, centered, Gaussian process $\mathbb{Z}$ with covariance

$$\rho(s,t;s',t') = (s \wedge s')(t \wedge t' - tt').$$

See van der Vaart and Wellner (1996, pp. 225ff). The weak convergence results for $\mathbb{B}_n$ and $\mathbb{W}_n$ follow immediately by the continuous mapping theorem.

We now give the steps in the proof of Theorem 1. Recall the definitions of F_c at (1), G_d at (2), H_n at (3), $\mathbb{Z}_n$ at (5), and $\mathbb{W}_n$ at (6). For given c the two sample test statistic is given by

$$W_n(c) = \int_0^1 \{\mathbb{W}_n(c/n,t)\}^2 \, dH_n(t).$$

Let ν_n be the probability measure on $(0,1)$ putting mass $1/(n-1)$ on each point of the form c/n for $1 \le c \le n-1$. The statistic $\overline{W}_n$ is then

$$\overline{W}_n = \int_0^1 \int_0^1 \{\mathbb{W}_n(s,t)\}^2 \, dH_n(t) \, d\nu_n(s).$$

Each step below consists of a statement followed by a detailed proof. In each case the assertions are intended to hold under the null hypothesis and the assumption that the common distribution H is continuous.

Step 1: For any sequence c_n with $\epsilon_n \equiv c_n/n \to 0$ we have

$$\left\{ \int_0^{\epsilon_n} + \int_{1-\epsilon_n}^1 \right\} \{\mathbb{W}_n(c/n,t)\}^2 \, dH_n(t) d\nu_n(s) \\ = \frac{\sum_{i=1}^{c_n} W_n(i) + \sum_{i=n+1-c_n}^{n} W_n(i)}{n-1} \to 0, \tag{A1}$$

in probability. Under the null hypothesis the mean of $W_n(c)$ is $1/6 + 1/(6n)$; see Anderson (1962, p. 1150). The expected value of (A1) is thus

$$\frac{2c_n}{n-1}\left(\frac{1}{6} + \frac{1}{6n}\right) \to 0.$$

Step 2: The integral

$$W_\infty = \int_0^1 \int_0^1 \mathbb{W}^2(s,t) dt \, ds,$$

is almost surely finite. All the variates involved are nonnegative so

$$\mathrm{E}(W_\infty) = \mathrm{E}\left(\int_0^1 \int_0^1 \mathbb{W}^2(s,t) dt \, ds \right) = \int_0^1 \chi(s,s) \, ds \int_0^1 \psi(t,t) \, dt = 1/6 < \infty.$$

Step 3: For any sequence ϵ_n tending to 0 as $n \to \infty$ take expectations to see

$$\left\{ \int_0^{\epsilon_n} + \int_{1-\epsilon_n}^1 \right\} \int_0^1 \mathbb{W}^2(s,t) dt \, ds \to 0 \text{ in probability.}$$

Step 4: The tensor product kernel

$$\rho = \chi \otimes \psi(s,t;s',t') = \chi(s,s')\psi(t,t'),$$

is compact and has eigenvalue-eigenfunction pairs

$$\lambda_{jk} = \frac{1}{j(j+1)}\frac{1}{\pi^2 k^2}, \quad f_{jk}(s,t) = f_{\chi,j}(s) f_{\psi,k}(t),$$

indexed by j, k each running from 1 to ∞. It follows as usual that the family

$$Z_{jk} = \frac{1}{\sqrt{\lambda_{jk}}} \int_0^1 \int_0^1 \mathbb{W}(s,t) f_{jk}(s,t) dt\, ds,$$

defines a family of independent standard normal variables. Parseval's identity is

$$\int_0^1 \int_0^1 \mathbb{W}^2(s,t) dt\, ds = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{Z_{jk}^2}{j(j+1)\pi^2 k^2}.$$

Step 5: For each fixed $\epsilon > 0$ we have

$$\int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t) dH_n(t)\ d\nu_n(s) - \frac{1}{n-1} \sum_{n\epsilon < i < n(1-\epsilon)} W_n^2(i) \to 0,$$

in probability. This is an easy consequence of the fact that for $i/n \le s < (i+1)/n$ we have $\int_0^1 \mathbb{W}_n^2(s,t) dF_n(t) = \mathbb{W}_n^2(i)$.

Step 6: For each fixed $\epsilon > 0$ we have

$$\int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t) dH_n(t)\ d\nu_n(s) - \int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t)\ dt\ ds \to 0.$$

Under the null hypothesis H_n converges weakly to the uniform law on the unit interval. Moreover ν_n converges weakly to Lebesgue measure on the unit interval. The weak convergence result in Theorem 2 uses a topology of uniform convergence on compacts such as the set $[\epsilon, 1-\epsilon] \times [0,1]$ and this implies the desired result.

Step 7: For each fixed $\epsilon > 0$ we have

$$\int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t) dt\ ds \rightsquigarrow \int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}^2(s,t)\ dt\ ds.$$

This is a consequence of weak convergence and the continuous mapping theorem.

Step 8: There is a metric d on the set of probability measures on the real line for which the metric topology is the topology of weak convergence. For each fixed $\epsilon > 0$ we have

$$d\left(\mathcal{L}\left(\int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}_n^2(s,t) dt\ ds\right), \mathcal{L}\left(\int_\epsilon^{1-\epsilon} \int_0^1 \mathbb{W}^2(s,t)\ dt\ ds\right)\right) \to 0.$$

There is then a sequence $\epsilon_n \to 0$ so slowly that this convergence continues to hold with ϵ replaced by ϵ_n and so that the convergences in Steps 5 and 6 continue to hold. Notice that by Step 3

$$d\left(\mathcal{L}\left(\int_{\epsilon_n}^{1-\epsilon_n} \int_0^1 \mathbb{W}^2(s,t)\ dt\ ds\right), \mathcal{L}\left(\int_0^1 \int_0^1 \mathbb{W}^2(s,t)\ dt\ ds\right)\right) \to 0.$$

for this sequence.

Step 9: For the sequence chosen in Step 8 we therefore have

$$\frac{1}{n-1}\sum_{n\epsilon_n<i<n(1-\epsilon_n)} W_n^2(i) \rightsquigarrow \int_0^1\int_0^1 \mathbb{W}^2(s,t)\,dt\,ds.$$

In view of Step 1 we see

$$\overline{W}_n \rightsquigarrow \int_0^1\int_0^1 \mathbb{W}^2(s,t)\,dt\,ds.$$

The law of the limit is, by Step 4, that of

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\frac{Z_{jk}^2}{j(j+1)\pi^2k^2}.$$

This completes the proof of Theorem 1. ■

Proof of Theorem 3. This is standard so we present only an outline. Conditions **A1** and **A2** can be used to prove that under the null

$$\Lambda_n - S_n \to 0,$$

in probability. The Lindeberg Central limit theorem then establishes the first conclusion of the Theorem. For more detailed arguments in a similar context see Guttorp and Lockhart (1988). Thus, under the conditions of the theorem the sequence of alternatives is contiguous to a sequence for which the null holds.

Contiguity implies that tightness under the null sequence extends to tightness under the alternative sequence; see theorem 3.10.7 in van der Vaart and Wellner (1996, p. 405) which is a version of LeCam's Third Lemma. This proves tightness, under the alternative, of the sequence of processes $\mathbb{W}_n$. Thus we need only compute the limiting finite dimensional distributions under the alternative sequence. As usual we apply LeCam's Third Lemma (again similar arguments are in Guttorp & Lockhart, 1988) to reduce the problem to studying the joint law, under the null hypothesis, of Λ_n and the vector $(\mathbb{W}_n(s_1,t_1),\ldots,\mathbb{W}_n(s_k,t_k))$ for an arbitrary sequence of time points $t_1,\ldots,t_k$ all in $[0,1]$.

Theorem 2 shows that under the null hypothesis

$$(\mathbb{W}_n(s_1,t_1),\ldots,\mathbb{W}_n(s_k,t_k)) \rightsquigarrow MVN_k(0,\mathbf{R}_W)$$

where $\mathbf{R}_W$ is the $k\times k$ matrix with i,jth entry

$$R_{Wij} = \rho_W(s_i,t_i;s_j,t_j).$$

The Lindeberg Central Limit Theorem may now be used to show that the vector

$$(S_n,\mathbb{W}_n(s_1,t_1),\ldots,\mathbb{W}_n(s_k,t_k)),$$

converges in distribution to multivariate normal with mean vector $(-\tau^2/2, 0, \dots, 0)$ and variance covariance matrix of the form

$$\begin{bmatrix} \tau^2 & \mathbf{c}^\top \\ \mathbf{c} & \mathbf{R}_W \end{bmatrix}.$$

The vector $\mathbf{c}$ is the limiting covariance which is found, after some algebra, to be

$$c_i = \mu(s_i, t_i) = \mu_\chi(s_i)\mu_\psi(t_i).$$

This completes the proof of the second assertion of the Theorem. ■

The third assertion is standard; see, for example, Guttorp and Lockhart (1988).

Proof of Proposition 1. Fix $0 < \delta < 1/2$ and let A_n denote the event $\{\delta \le \hat{c}/n \le 1-\delta\}$. We will show

$$\lim_{n\to\infty} P(A_n) = 0.$$

This will prove (9) in Proposition 1. To this end fix $0 < \epsilon < \delta$. Define

$$M_n = \sup_{\delta \le s \le 1-\delta} \int_0^1 \frac{\mathbb{B}_n^2(s,t)}{s(1-s)}\, dt,$$

and

$$M_n'(\epsilon) = \sup_{\epsilon \le s \le \delta} \int_0^1 \frac{\mathbb{B}_n^2(s,t)}{s(1-s)}\, dt. \tag{A2}$$

Then

$$A_n \subset \{M_n'(\epsilon) < M_n\}.$$

Weak convergence of $\mathbb{B}_n$ to $\mathbb{B}$ guarantees that

$$\limsup_{n\to\infty} P(A_n) \le \limsup_{n\to\infty} P\{M_n'(\epsilon) < M_n\} \le P(M'(\epsilon) \le M),$$

where

$$M = \sup_{\delta \le s \le 1-\delta} \int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt \;\text{ and }\; M'(\epsilon) = \sup_{\epsilon < s \le \delta} \int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt.$$

We claim that

$$\lim_{\epsilon\to 0} P(M'(\epsilon) \le M) = 0. \tag{A3}$$

This will prove $\limsup_{n\to\infty} P(A_n) = 0$ and (9) in Proposition 1.

Assertion (A3) would follow from a law of the iterated logarithm (as $s \to 0$) for the process

$$s \mapsto \frac{\int_0^1 \mathbb{B}^2(s,t)\, dt}{s(1-s)}.$$

While we expect such a result to hold we have not tried to prove anything along those lines. We will establish instead the lower bound

$$\limsup_{s \to 0} \int_0^1 \frac{\pi^2 \mathbb{B}^2(s,t)}{2\log\{\log(1/s)\}s(1-s)}\, dt \geq 1, \tag{A4}$$

almost surely which is enough to imply (A3). We enumerate the steps needed:

Step 1: Use the Cauchy–Schwarz inequality to see that

$$\begin{aligned}\int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt &= \int_0^1 \frac{\{\mathbb{Z}(s,t) - s\mathbb{Z}(1,t)\}^2}{s(1-s)}\, dt \\ &\geq \int_0^1 \frac{\mathbb{Z}^2(s,t)}{s(1-s)}\, dt + \frac{s}{1-s}\int_0^1 \mathbb{Z}^2(1,t)\, dt \\ &\quad - \frac{2s}{1-s}\sqrt{\int_0^1 \frac{\mathbb{B}^2(s,t)}{s(1-s)}\, dt \int_0^1 \mathbb{Z}^2(1,t)\, dt}.\end{aligned}$$

From this we deduce that it is enough to show that

$$\limsup_{s \to 0} \int_0^1 \frac{\pi^2 \mathbb{Z}^2(s,t)}{2\log\{\log(1/s)\}s}\, dt \geq 1 \text{ almost surely.} \tag{A5}$$

Step 2: For each s the process $t \mapsto \mathbb{Z}(s,t)/\sqrt{s}$ is a Brownian Bridge. If we put

$$W(s) = \int_0^1 \frac{\mathbb{Z}^2(s,t)}{s}\, dt,$$

then each $W(s)$ has the same distribution as the limit law of the usual Cramér–von Mises statistic which is the law of

$$T \equiv \sum_{j=1}^{\infty} \lambda_j Z_j^2. \tag{A6}$$

In this representation the Z_j are iid standard normal and the eigenvalues λ_j are given, for $j = 1, 2, \dots,$ by $\lambda_j = \frac{1}{\pi^2 j^2}$.

Step 3: The process $\mathbb{Z}$ has independent increments in s. For $0 < s' < s$ the process

$$t \mapsto \frac{\mathbb{Z}(s,t) - \mathbb{Z}(s',t)}{\sqrt{s-s'}}.$$

has the same law as

$$t \mapsto \frac{\mathbb{Z}(s,t)}{\sqrt{s}}.$$

Step 4: Now fix $r < 1$ to be chosen later. Define $s_n = r^n$ for $n = 0, 1, \ldots$. Put

$$W_n = \int_0^1 \frac{\mathbb{Z}^2(s_n,t)}{s_n}\, dt,$$

and

$$W_n^* = \int_0^1 \frac{\{\mathbb{Z}(s_n,t) - \mathbb{Z}(s_{n+1},t)\}^2}{s_n - s_{n+1}}\, dt.$$

All of these variables have the law of $W(s)$ described above.

Step 5: Fix $\epsilon > 0$. The function

$$f(r) = \sqrt{1-r}\sqrt{1-\epsilon/2} - \sqrt{r}\sqrt{1+\epsilon/2}, \tag{A7}$$

has the property that $f^2(0) = 1 - \epsilon/2$. Choose r so small that $f^2(r) = 1 - \epsilon$. Let A_n be the event $W_n^* > 2(1-\epsilon/2)\lambda_1 \log(\log(1/s_n))$ and B_n be the event $W_{n+1} \le 2(1+\epsilon/2)\lambda_1 \log(\log(1/s_n))$. Then for this choice of r

(a) The event that A_n occurs infinitely often (i.o.) has probability 1.
(b) The event that B_n occurs for all large n has probability 1.
(c) So the event $A_n \cap B_n$ i.o. has probability 1.

Step 6: On the event $A_n \cap B_n$ we have $W_n \ge 2(1-\epsilon)\lambda_1 \log(\log(1/s_n))$ so that this event occurs infinitely often. This establishes (A5).

It remains to finish Step 5 by verifying that (a), (b), and (c) of that step all hold. By construction the events $A_n; n \ge 1$ are independent. Moreover, with T as defined in (A6) we have $T > \lambda_1 Z_1^2$. Therefore

$$\begin{aligned} P(A_n) &= P(T > 2(1-\epsilon)\lambda_1 \log(\log(1/s_n)) \\ &\ge P(Z_1^2 > 2(1-\epsilon)\log(\log(1/s_n))) \\ &= 2P\left\{Z_1 > \sqrt{2(1-\epsilon)\log(\log(1/s_n))}\right\}. \end{aligned}$$

We can then use Mill's inequalities for the tail of the normal distribution to show that for all n sufficiently large

$$P(A_n) \ge \frac{\exp\left[-(1-\epsilon/2)\left\{\log(n) + \log(\log(1/r))\right\}\right]}{\sqrt{(1-\epsilon/2)\pi\left\{\log(n) + \log(\log(1/r))\right\}}}.$$

Summing over n we get $+\infty$. The converse to Borel–Cantelli (for independent events) then proves assertion (a) of Step 5.

For event B we claim

$$\sum_n P(B_n^c) = \sum_n P\left\{T > 2(1+\epsilon)\lambda_1 \log(\log(1/s_n))\right\} < \infty,$$

which would prove statement (b) in Step 5. We use a version of Chernoff's bound. The moment generating function of T is, for all $t < 1/(2\lambda_1)$,

$$M_T(t) = \prod_{k\geq 1} \frac{1}{\sqrt{1-2\lambda_k t}}.$$

Markov's inequality then shows that for all such t we have

$$P(T > \lambda_1 x) \leq M(t)\exp(-t\lambda_1 x).$$

Take logs to get

$$\log(P(T > \lambda_1 x)) \leq -\frac{1}{2}\log(1-2\lambda_1 t) - t\lambda_1 x - \frac{1}{2}\sum_{k\geq 2}\log(1-2\lambda_k t). \tag{A8}$$

Choose t to minimize the first two terms of (A8). The minimum occurs at

$$t = t(x) = \frac{x-1}{2\lambda_1 x}.$$

Evaluating (A8) at $t(x)$ and using $\lambda_k/\lambda_1 = 1/k^2$ we get the inequality

$$\log(P(T > \lambda_1 x)) \leq \frac{\log(x)}{2} - \frac{x-1}{2} - \sum_{k\geq 2}\log\{1-(1-1/x)/k^2\}.$$

Careful analysis of the right-hand size of this inequality shows that for

$$x_n = 2(1+\epsilon/2)\left\{\log(n) + \log(\log(1/r))\right\},$$

we have

$$\sum_n P(B_n^c) < \infty.$$

Assertion (b) of Step 5 follows by the Borel–Cantelli Lemma.

Finally we check assertion (c) of step 5. Straightforward algebra shows that

$$W_n = (1-r)W_n^* + rW_{n+1} + \frac{2\sqrt{s_n(s_n - s_{n+1})}}{s_n}\int_0^1 \frac{\mathbb{Z}(s_{n+1}, t)}{\sqrt{s_{n+1}}}. \tag{A9}$$

Apply the Cauchy–Schwarz inequality to the final integral in (A9) to see that

$$W_n \geq (1-r)W_n^* + rW_{n+1} - 2\sqrt{r(1-r)W_n^* W_{n+1}} \tag{A10}$$

$$= \left(\sqrt{(1-r)W_n^*} - \sqrt{rW_{n+1}}\right)^2. \tag{A11}$$

On the event $A_n \cap B_n$ the quantity inside parentheses in (A10) is at least

$$\sqrt{(1-r)2(1-\epsilon)\log(\log(1/s_n))} - \sqrt{r \cdot 2(1+\epsilon)\log(\log(1/s_n)} \\ = f(r)\sqrt{2\log(\log(1/s_n))},$$

where $f(r)$ is defined at (A7). For the value of r chosen at the start of step 5 we then find that on $A_n \cap B_n$ we have

$$W_n > 2(1-\epsilon)\log(\log(1/s_n)).$$

This finishes the proof of Assertion (9) of Proposition 1. Assertion (10) follows by symmetry. The argument leading to (A12) shows that there is a sequence ϵ_n tending to 0 for which $M_n'(\epsilon_n) \to \infty$ in probability where $M_n'(\epsilon)$ is defined at (A2). This is enough to prove Assertion (11) of Proposition 1. ■

A.1 Evidence for Conjectures 1 and 2

For $\epsilon > 0$ we define

$$I_n(\epsilon) = \{c : 1 \le c \le n\epsilon \;\text{ or }\; 1 \le n - c \le n\epsilon\}.$$

Proposition 1 establishes that there is a sequence $\epsilon_n \searrow 0$ such

$$\lim_{n\to\infty} P(\hat{c}_n \in I_n(\epsilon_n)) = 1.$$

Thus

$$P\left[W_{\max,n} = \max\{W_n(c) : c \in I_n(\epsilon_n)\}\right] \to 1. \tag{A12}$$

We now outline the steps in our strategy for proving the conjecture before giving some evidence for each step.

Step 1: There are constants a_n and b_n and a random variable V such that

$$a_n W_{\max,n} - b_n \rightsquigarrow V,$$

and V has a continuous limit distribution.

Step 2: So

$$a_n \max\{W_n(c) : c \in I_n(\epsilon_n)\} - b_n \rightsquigarrow V.$$

Step 3: There are random variables $\tilde{W}_n(c)$ such that under the null hypothesis

$$a_n \max\{|W_n(c) - \tilde{W}_n(c)| : c \in I_n(\epsilon_n)\} \to 0,$$

and such that for each $c \in I_n(\epsilon_n)$ the variable $\tilde{W}_n(c)$ is measurable with respect to the σ field generated by $X_c, c \in I_n(\epsilon_n)$. To be specific we define, for $c < n/2$,

$$\tilde{W}_n(c) = \int_0^1 c\{F_c(u) - u\}^2 \, du,$$

and, for $c > n/2$,

$$\tilde{W}_n(c) = \int_0^1 d\{G_d(u) - u\}^2 \, du.$$

(Recall the shorthand $d = n - c$.)

Step 4: Define

$$\tilde{\Lambda}_n = \sum_{n\epsilon_n < c \le c_0} \phi_f(X_c)/\sqrt{n} - \sum_{c_0 < c < n - n\epsilon_n} \phi_g(X_c)/\sqrt{n}.$$

The log-likelihood ratio Λ_n satisfies

$$\Lambda_n - \tilde{\Lambda}_n \to 0,$$

in probability, under the null hypothesis.

Step 5: Since $\tilde{W}_{\max}$ is independent of $\tilde{\Lambda}_n$ we may apply LeCam's third lemma to show that under the sequence of contiguous alternatives we have

$$a_n W_{\max,n} - b_n \rightsquigarrow V.$$

Step 6: Since this limit law is the same as under the null we must power minus level tends to 0.

For some of these steps we can fill in partial evidence.

For Step 1 we would hope to follow the ideas in Jaeschke (1979) to show that the limit V has an extreme value distribution. In that paper the maximizer of the usual empirical process, standardized by dividing by its SD, is shown to have an extreme value limit with constants analogous to a_n and b_n involving $\sqrt{\log\log n}$ and $\log\log\log n$.

Step 2 is a consequence of Step 1 and (A12).

In Step 3 we would hope to use the closeness of H_n to the uniform distribution to convert the $dH_n(u)$ integrals to du integrals. Then we write

$$\frac{cd}{n} \int_0^1 \{F_c(u) - G_d(u)\} \, 2 \, du,$$

as a sum of three terms

$$T_1 = \frac{d}{n} \int_0^1 c\{F_c(u) - u\}^2 \, du,$$

$$T_2 = \frac{c}{n} \int_0^1 d\{G_d(u) - u\}^2 \, du,$$

and

$$T_3 = -2\frac{\sqrt{cd}}{n} \int_0^1 \sqrt{c}\,\{F_c(u) - u\} \cdot \sqrt{d}\,\{G_d(u) - u\} \, du.$$

The integrals in T_1 and T_2 are both one sample Cramér-von Mises statistics so they are on the order 1. For any sequence $c = c_n$ such that $c_n/n \to 0$ the coefficient in front of T_2 is $o(1)$. So T_2 is negligible relative to T_1. The Cauchy–Schwarz inequality then shows T_3 is negligible relative to T_2. There is a parallel argument when $c_n/m \to 1$.

Step 4 is not conjecture; its proof is straightforward from the assumptions of the Conjecture. Steps 5 and 6 are exactly parallel to the arguments in Lockhart (1991).