International Journal of Statistics and Probability; Vol. 2, No. 1; 2013
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Competitor of the Kolmogorov–Smirnov Test for the Goodness-of-fit Problem

Claudio G. Borroni[1] & Paola M. Chiodini[1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

Correspondence: C. G. Borroni, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, via Bicocca degli Arcimboldi 8, Milano 20126, Italy. Tel: 39-2-6448-3120. E-mail: claudio.borroni@unimib.it

Received: September 19, 2012 Accepted: November 25, 2012 Online Published: December 19, 2012

doi:10.5539/ijsp.v2n1p1 URL: http://dx.doi.org/10.5539/ijsp.v2n1p1

Abstract

The classical goodness-of-fit problem, in the case of a null continuous and completely specified distribution, is faced by a new version of the Girone–Cifarelli test (see Girone, 1964; Cifarelli, 1974 & 1975). This latter test was introduced for the two-sample problem and showed a substantial gain of power over other common tests based on the empirical distribution function, notably over the Kolmogorov–Smirnov test. First, the problem of the re-definition of the Girone–Cifarelli test-statistic is considered, by reviewing the literature on the subject. A classical remark by Anderson (1962) is shown to be useful to choose the integrating function in the newly defined test-statistic. The sample properties of such a test-statistic are then studied. A table of critical values is obtained by simulation; moreover, the asymptotic null distribution is considered and its accuracy as an approximation of the finite distribution is discussed. Finally, a simulation study, considering a wide set of distributions mostly used in applications, is conducted to compare the proposed test with its classical competitors. The study gives some indications to locate such situations where the Girone-Cifarelli test performs at its best, notably over the Kolmogorov–Smirnov test.

Keywords: goodness-of-fit tests, empirical distribution function, Girone–Cifarelli test, nonparametric statistical methods

1. Introduction

A random sample $x_1, \ldots, x_n$ is drawn from a population X with continuous distribution function F, to test the null hypothesis $H_0 : F(x) = F_0(x)$ against the alternative $H_1 : F(x) \neq F_0(x)$, $x \in \Re$, where F_0 is completely specified. This common goodness-of-fit problem is usually faced by three classes of tests: the chi-square test, the tests based on spacings and the tests based on the empirical distribution function (edf). In this latter class several test-statistics can be considered, usually by adapting their versions for the two-sample problem.

The most known test based on the edf F_n is surely the Kolmogorov–Smirnov test, which rejects H_0 for large values of the test-statistic

$$K_n = \sup_{t \in (-\infty, +\infty)} |F_0(t) - F_n(t)|. \tag{1}$$

As known, other test-statistics can be defined by considering the square of the difference $|F_0(t) - F_n(t)|$, like in the Cramér–Von Mises test

$$C_n = n \int_{-\infty}^{+\infty} [F_0(t) - F_n(t)]^2 \, \mathrm{d}F_0(t). \tag{2}$$

Notice that in the above considered test-statistics F_0, a continuous model, is compared with F_n, which has discontinuities at $x_1, \ldots, x_n$. However, in (1) the supremum of the difference $|F_0(t) - F_n(t)|$ is taken, while in (2) the squared difference $[F_0(t) - F_n(t)]^2$ is integrated with respect to the continuous function F_0. Because of these latter choices, no particular care is needed in the definition of the value taken by the edf at its points of discontinuity. This means that one can use the usual definition

$$F_n(x) = \frac{i}{n}, \qquad \text{for } x_{(i)} \leq x < x_{(i+1)} \quad (i = 0, \ldots, n), \tag{3}$$

(where $x_{(1)}, \ldots, x_{(n)}$ denotes the ordered sample, $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$), which makes F_n to be right-continuous, or equivalently set

$$F_n(x_{(i)}) = \frac{i-c}{n} \qquad (i = 1, \ldots, n), \tag{4}$$

(where c is chosen in [0,1]), so that F_n can take every value of its jump at $x_{(i)}$ $(i = 1, \ldots, n)$.

Turning back to (2), the function with which the squared difference $[F_0(t)-F_n(t)]^2$ is integrated could be substituted by the edf itself. This choice allows to simplify the test-statistic as

$$C_n' = n\int_{-\infty}^{+\infty} [F_0(t) - F_n(t)]^2 \, \mathrm{d}F_n(t) = \sum_{i=1}^{n} [F_0(x_{(i)}) - F_n(x_{(i)})]^2, \tag{5}$$

but the definition of $F_n(x_{(i)})$ becomes now relevant. However, Anderson (1962) pointed out that, when $c = 1/2$ is taken in (4), the test-statistics C_n and C_n' are equivalent, as the former can be also written as

$$C_n = \sum_{i=1}^{n} \left[F_0(x_{(i)}) - \frac{i-1/2}{n}\right]^2 + \frac{1}{12n}. \tag{6}$$

Besides such a latter equivalence, setting $c = 1/2$ in (4) is, as a matter of fact, a natural choice. Indeed, forcing the edf to take the mid-point of its jump at $x_{(i)}$ seems less arbitrary than choosing any other value in the jump (including the extremes i/n and $(i-1)/n$, $i = 1, \ldots, n$.

Notice again that any choice of $F_n(x_{(i)})$, made to give a final form to C_n' in (5), does not affect the usual definition of the edf in the open intervals $(x_{(i)}, x_{(i+1)})$, $i = 1, \ldots, n-1$. However, in the literature some modifications of the edf in such intervals were also proposed. For instance, Green and Hegazy (1976) pointed out that when the edf is re-defined as

$$F_n'(x) = \frac{i+1/2}{n+1} \qquad \text{for } x_{(i)} < x < x_{(i+1)} \qquad (i = 1, \ldots, n-1), \tag{7}$$

the criterion C_n in (2) reduces, up to a multiplicative constant, to

$$\sum_{i=1}^{n} \left[F_0(x_{(i)}) - \frac{i}{n+1}\right]^2, \tag{8}$$

which is shown to lead to a powerful test under some circumstances. Notice that the test-statistic in (8) can be also obtained from C_n' in (5) by re-defining accordingly the value of the edf at its discontinuities, that is by setting

$$F_n'(x_{(i)}) = \frac{i}{n+1} \qquad (i = 1, \ldots, n), \tag{9}$$

which is again the mid-point of the jump of F_n' at $x_{(i)}$. Other modifications of the definition of the edf in the open intervals $(x_{(i)}, x_{(i+1)})$ are known. By noticing that the term $i/(n+1)$ is actually the expectation of $F_0(x_{(i)})$ under the null hypothesis, Pyke (1959) proposed a new version of the Kolmogorov–Smirnov criterion (1), which in turn induces a further modification of the definition of the edf (see also Brunk, 1962).

The above remarks will be used in this paper to propose a goodness-of-fit version of the Girone–Cifarelli test, which was mainly studied for the two-sample problem. The definition of the test-statistic for goodness-of-fit purposes raises some questions which will be addressed in the next section, where the sample properties of the newly proposed test-statistic will be also analyzed. Section 3 will report some results of a simulation study, where the proposed test is compared with its most important competitors based on the edf. Section 4 will provide a real-data example and some conclusions.

2. Definition of the Test-statistic

Girone (1964) proposed a test for the equality of two populations X and Y, based on the statistic

$$(m+n)\int_{-\infty}^{+\infty} |F_n(t) - G_m(t)| \, \mathrm{d}H_{m+n}(t), \tag{10}$$

where F_n, G_m and H_{m+n} denote respectively the edf's of a n-sample from X, a m-sample from Y and the pooled $(m+n)$-sample. The test was actually originally proposed in the special case $n = m$ and its sample properties were studied by Cifarelli (1974 & 1975). Generalizations for the case $n \leq m$ were proposed by Goria (1972), by Borroni (2001) and independently by Schmid and Trede (1995). For the two-sample problem, the Girone–Cifarelli test proved to be superior to other common tests, notably the Kolmogorov–Smirnov test, under a wide set of circumstances. This fact is far from being unexpected, as in (10) the whole behavior of the difference $|F_n(t) - G_m(t)|$ is considered, while in the Kolmogorov–Smirnov test just its supremum is taken.

A goodness-of-fit version of the Girone–Cifarelli test would result useful. Using the same settings as in section 1, the edf F_n of the single n-sample is now to be compared with the null model F_0. The function with respect to which the difference $|F_n(t) - F_0(t)|$ is to be integrated could then be the null model F_0 or the edf F_n. As above remarked, this latter choice highly simplifies the structure of the test statistic, as

$$n \int_{-\infty}^{+\infty} |F_0(t) - F_n(t)| \, \mathrm{d}F_n(t) = \sum_{i=1}^{n} |F_0(x_{(i)}) - F_n(x_{(i)})|; \tag{11}$$

as a consequence, the definition of the value taken by the edf at its discontinuities becomes relevant. Following the above suggestion by Anderson (1962) for $C_n^{'}$, we can then take

$$F_n(x_{(i)}) = \frac{i - 1/2}{n} \quad (i = 1, \ldots, n), \tag{12}$$

and define

$$A_n^{'} = \sum_{i=1}^{n} \left| F_0(x_{(i)}) - \frac{i - 1/2}{n} \right|. \tag{13}$$

The sample properties of $A_n^{'}$ are easily derived from its two-sample equivalent. First of all notice that, being F_0 a continuous model, the variables $F_0(x_{(i)})$, $i = 1, ..., n$, are uniform over [0,1] and hence $A_n^{'}$ is distribution-free under H_0. For small sample sizes, the null distribution of $A_n^{'}$ can then be determined by simulation, as pointed out in the next section. Moreover, following Cifarelli (1975), $n^{(-1/2)} A_n^{'}$ is asymptotically distributed as the r.v.

$$\int_0^1 |w(\tau)| \, \mathrm{d}\tau \Big| w(1) = 0, \tag{14}$$

where $\{w(\tau), t \in [0, 1]\}$ denotes the Brownian motion in [0,1]. A tabulation of the quantiles of (14) is found in Johnson and Killeen (1983); see also Shepp (1982 & 1991) and Takács (1993).

Differently from $C_n^{'}$, $A_n^{'}$ is not equivalent to the statistic obtained by using F_0 as an integrating function in (11). This is shown by considering that

$$A_n = \int_{-\infty}^{+\infty} |F_0(t) - F_n(t)| \, \mathrm{d}F_0(t) = \frac{1}{2} \sum_{i=1}^{n} \left[\left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \left(F_0(x_{(i)}) - \frac{i-1}{n} \right) - \left| F_0(x_{(i)}) - \frac{i}{n} \right| \left(F_0(x_{(i)}) - \frac{i}{n} \right) \right]. \tag{15}$$

Schmid and Trede (1996) considered $\sqrt{n} A_n$ as a test-statistic and reported a small simulation study to evaluate its performance. They concluded that the power of A_n is quite close to the one of the Cramér–Von Mises test, without reporting situations where A_n performs definitely nor uniformly better than C_n. It should be pointed out that $A_n^{'}$, which has a rather simpler form, is not equivalent to A_n, even if the two tests have often similar powers. Consequently, the next section will first present some results of a simulation study without distinguishing between $A_n^{'}$ and A_n. In the following, some insights about the situations where the two tests are likely to perform differently will then be given. In a sense, the reported simulation study can be considered as an extension of the one by Schmid and Trade (1996), because it will be able to locate some alternatives where the test based on $A_n^{'}$, along with the one based on A_n, performs definitely better than the Cramér–Von Mises test.

3. Simulation Study

The first task to develop a goodness-of-fit test based on $A_n^{'}$ is to determine its critical values. As above mentioned, being F_0 completely specified and continuous, the transformation $F_0(X)$ gives a Uniform distribution over (0,1). Hence the null distribution of $A_n^{'}$ can be simulated by randomly generating a large number of samples from such a distribution, with a fixed size n. The critical values of the test can then be determined by computing the value taken by $A_n^{'}$ for each simulated sample as long as the related frequency distribution. For a selected range of sample sizes and some common significance levels, Table 1 reports the critical values of $n^{(-1/2)} A_n^{'}$ based on 10^6 simulated samples.

Table 1. Simulated critical values of $n^{(-1/2)}A_n^{'}$

α	$n = 5$	$n = 10$	$n = 20$	$n = 30$	$n = 50$	∞
0.01	0.7142	0.7364	0.7436	0.7465	0.7478	0.7518
0.05	0.5670	0.5747	0.5783	0.5791	0.5807	0.5821
0.10	0.4893	0.4942	0.4966	0.4972	0.4982	0.4993
0.15	0.4398	0.4439	0.4459	0.4462	0.4470	0.4480
0.20	0.4029	0.4064	0.4081	0.4088	0.4092	0.4103

As a term of comparison, the last column of Table 1 reports the critical values of the asymptotic distribution of $n^{(-1/2)}A_n^{'}$ (see section 2). The fast convergence to the asymptotic approximation can be easily appreciated. In order to get further indications about the accuracy of the asymptotic distribution and the sample sizes needed to use it, the simulated cdf's obtained for fixed values of n were compared with the asymptotic cdf, whose expression is found in Johnson and Killeen (1983). Figure 1 reports the results obtained for $n = 10$. As seen, the asymptotic cdf is very close to the "real" one, even if a certain difference is observed, especially for small values of the variable. However, one can claim that, to develop a goodness-of-fit test based on $A_n^{'}$, just the right tail of its null distribution is relevant. In effect, when the last part of the distribution is considered (say for such x so that $\Pr\{A_n^{'} \leq x\} > 0.8$) the finite cdf is rather close to the asymptotic cdf. To get into further details, Table 2 reports, for some selected sample sizes, the greatest absolute difference of the two cdf's and the same difference referred to the right tail of the distribution. From such a table, a minimum value of $n = 50$ is to be advised to get a correct approximation of the null distribution.

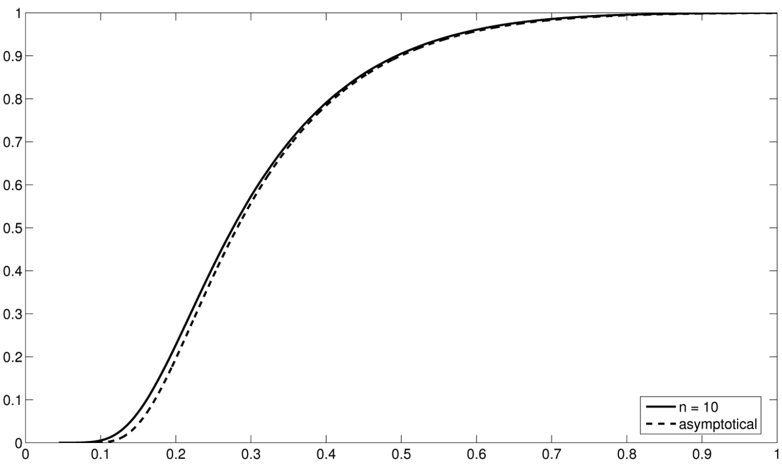

Figure 1. Comparison of the "real" cdf and the asymptotic cdf of $A_n^{'}$ under H_0 ($n = 10$)

Table 2. Greatest absolute difference between the "real" cdf and the asymptotic cdf of $A_n^{'}$ under H_0 (whole distribution and right tail)

n	whole distrib.	right tail
5	0.0626	0.0112
10	0.0330	0.0065
20	0.0173	0.0035
30	0.0123	0.0023
50	0.0068	0.0018
100	0.0042	0.0005

After computing the critical values of the test based on $A_n^{'}$, its power can be estimated by simulation as well. This section reports some results of a wide simulation study conducted at this aim. Notice that the power of a goodness-of-fit test will depend on the model F_0 chosen under H_0 as long as on the real cdf of X under H_1, which will be denoted as F_1. Generally F_1 will belong to a family of distributions containing F_0 itself, which is hence obtained by an appropriate choice of the parameter(s) of the family. In this paper we will focus on three models for H_0, mostly used in applications: the standard Normal distribution, the unit exponential distribution and the uniform distribution on the unit interval. For each null model, F_1 will belong to three different families of distributions containing F_0.

Consider first the standard Normal as a null model. A suitable family for F_1 could be the skew-normal (SN) distribution (see Azzalini, 1985) with density:

$$f(x) = 2\,\phi(x)\,\Phi(ax), \qquad x \in \Re, \tag{16}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cdf of a standard Normal respectively. The parameter $a \in \Re$ regulates the skewness of the distribution, thus giving a standard Normal if set to zero. To this end, using family (16) for F_1 in a simulation study, can result in an useful analysis of such situations where the researcher needs to test normality against possible asymmetries of data. It is known, however, that data may depart from normality due to heavy-tailedness. To simulate such latter situations, the Student's T density can be used as a family for F_1 :

$$f(x) = \frac{\Gamma\left(\frac{a+1}{2}\right)}{\sqrt{a\pi}\,\Gamma\left(\frac{a}{2}\right)}\left(1 + \frac{x^2}{a}\right)^{-\frac{a+1}{2}}, \quad x \in \Re, \tag{17}$$

($\Gamma(\cdot)$ denotes the gamma function). The family (17) gives only symmetric distributions with heavy tails, such phenomenon being reduced by increasing the parameter $a > 0$; as known, the family converges to the normal distribution when $a \to \infty$. Finally, to simulate such cases where the normality of data depends on the application of the central limit theorem, one can choose for F_1 the gamma (GA) density with unit scale:

$$f(x) = \frac{1}{\Gamma(a)}\, x^{a-1} e^{-x}, \quad x > 0. \tag{18}$$

As an effect of the above theorem, when $a \to \infty$, family (18) gives a normal density (which can be then standardized to be consistent with the null model F_0). However, in applications, a may not be large enough to guarantee a good convergence; the researcher may then need a powerful test to detect such a failed convergence.

Table 3. Simulated powers when the null model is a standard Normal distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	A_n'	K_n	C_n	D_n
		.0101	.0099	.0099	.0103
10	null	.0500	.0506	.0502	.0506
		.1007	.1010	.1013	.1002
		.1989	.1575	.1925	.1707
10	SN(1)	**.4503**	.3849	.4392	.4088
		.5948	.5308	.5841	.5555
		.3320	.2743	.3291	.2755
10	SN(1.5)	**.6460**	.5739	.6399	.5908
		.7850	.7243	.7807	.7431
		.0312	.0243	.0284	**.4673**
10	T(1.5)	.1117	.0933	.1044	**.6088**
		.1948	.1772	.1843	**.6868**
		.0369	.0302	.0339	**.5883**
10	T(1.25)	.1272	.1062	.1197	**.7080**
		.2185	.2007	.2076	**.7703**
		.0119	**.0203**	.0158	.0146
10	GA(2)	.0687	**.0841**	.0748	.0768
		.1363	.1485	.1415	**.1525**

Table 3 (continued). Simulated powers when the null model is a standard Normal distribution ($\alpha = 0.01, 0.05, 0.1$).

n	H_1	A_n'	K_n	C_n	D_n
		.0132	**.0240**	.0179	.0169
10	GA(1.5)	.0769	**.0946**	.0839	.0879
		.1508	.1653	.1556	**.1716**
		.2022	.1491	.1904	.1904
25	SN(0.5)	**.4215**	.3494	.4067	.4054
		.5501	.4809	.5385	.5350
		.6637	.5441	.6471	.6302
25	SN(1)	**.8716**	.7989	.8627	.8585
		.9317	.8894	.9272	.9251
		.0335	.0274	.0313	**.6016**
25	T(1.75)	.1299	.1107	.1219	**.7532**
		.2321	.1986	.2160	**.8246**
		.0392	.0347	.0374	**.7334**
25	T(1.5)	.1528	.1329	.1434	**.8468**
		.2717	.2355	.2543	**.8957**
		.0296	**.0455**	.0363	.0316
25	GA(2)	.1230	**.1403**	.1314	.1395
		.2171	.2295	.2201	**.2520**
		.0398	**.0592**	.0483	.0427
25	GA(1.5)	.1585	.1745	.1754	**.1817**
		.2703	.2692	.2698	**.3192**
		.2427	.1775	.2304	.2400
100	SN(0.25)	**.4684**	.3886	.4523	.4626
		.5904	.5198	.5777	.5859
		.8476	.7331	.8315	.8439
100	SN(0.5)	**.9514**	.9040	.9453	.9507
		.9762	.9505	.9730	.9760
		.0728	.0677	.0675	**.9398**
100	T(2)	.2909	.2545	.2708	**.9839**
		.4779	.4245	.4563	**.9931**
		.1005	.1005	.0953	**.9811**
100	T(1.75)	.3682	.3421	.3225	**.9959**
		.5697	.5291	.5558	**.9984**
		.1959	.2275	.2179	**.2678**
100	GA(2)	.4719	.4442	.4711	**.6476**
		.6408	.5777	.6278	**.8400**
		.3151	.3245	.3351	**.4569**
100	GA(1.5)	.6350	.5839	.6327	**.8505**
		.7863	.8208	.7805	**.9635**

Table 3 reports the results of a set of simulations, each based on 10^5 replications, for the null standard normal model. Some selected alternative distributions, all belonging to the above described families, are chosen. Table 3 reports the powers of the tests based on A_n', K_n and C_n. Another classical goodness-of-fit test is also considered:

the Anderson–Darling test,

$$D_n = \int_{-\infty}^{+\infty} [F_n(t) - F_0(t)]^2 \, \frac{1}{F_0(t)[1 - F_0(t)]} \, \mathrm{d}F_0(t); \tag{19}$$

here the squared difference $[F_n(t) - F_0(t)]^2$ is weighted to get more sensibility in the tails of the distributions. The powers reported in Table 3 were obtained by fixing three different values of the significance level α: 0.01, 0.05 and 0.1 (for each entry, the corresponding powers are listed in the latter order; the best power is highlighted too). It seems, however, that the performance of none of the considered tests is really affected by the choice of α. Moreover, the first row of Table 3 reports the estimated actual significance level, which is always very close to the nominal one, even for a small sample size as $n = 10$ (similar results, not reported here for the sake of brevity, were obtained for larger sample sizes). Notice that, for each considered alternative distribution, the values of the related parameter were set to allow a relevant comparison of the estimated powers; this need implies, incidentally, that the same value of the parameter cannot be chosen for all sample sizes in most cases. However, Table 3 (along with the following tables) was built so that at least one same value of the parameter is chosen for adjacent sample sizes. Table 3 emphasizes that, when used as a test of normality, the Girone–Cifarelli test has a good power against some kinds of alternatives. More specifically, the test outperforms all the other considered tests (and notably the Kolmogorov–Smirnov and the Cramér–Von Mises test) when the alternative distribution belongs to the skew-normal class. The superiority of the Girone–Cifarelli test for skewed alternatives seems indeed to be a general rule, at least among the considered tests, as further evidenced in the following simulations. When normality is to be tested against heavy tailedness, like for the Student's T alternatives considered in Table 3, the performance of the Girone–Cifarelli test gets worse, notably over the Anderson–Darling test. This result is far from being unexpected, but it has to be underlined that $A_n^{'}$ still keeps its superiority over the Cramér–Von Mises test (and the Kolmogorov–Smirnov test). The superiority of the Anderson–Darling test still characterizes Gamma alternatives. In this chance, however, the Girone–Cifarelli test gets worse even over the Cramér–Von Mises test and the Kolmogorov–Smirnov test. A global evaluation of Table 3 shows that, as expected, the performances of the considered tests become similar when the sample size increases, even if all the above conclusions still hold. Notably, $A_n^{'}$ outperforms the other considered tests for skew-normal alternatives, as D_n does for Student's T and Gamma alternatives. However, in this latter case, the Girone–Cifarelli test seems to grow better over its competitors as the sample size increases.

Another set of simulations was conducted by setting the unit exponential distribution as a null model. This assumption is typical for many datasets in reliability analyses. In this kind of applications, exponentiality is often to be tested against some more complicate distributional assumptions. To this end, a natural choice for F_1 is again the gamma (GA) density with unit scale. When $a = 1$, (18) reduces to the unit exponential. Another family of distributions, mostly used in reliability analyses as well, is the Weibull (W) density with unit scale:

$$f(x) = a\, x^{a-1}\, \mathrm{e}^{-x^a}, \qquad x > 0, \tag{20}$$

which was used as a family for F_1, after noticing that it reduces to the unit exponential when $a = 1$. Finally, a third family was used to shape the alternative hypothesis:

$$f(x) = (1 + a\, x)^{-\left(1+\frac{1}{k}\right)}, \qquad x > 0, \tag{21}$$

that is the generalized-Pareto (GP) density with zero location and unit shape. (21) gives the unit exponential density as $a \to 0$. The best results for the Girone–Cifarelli test were obtained for the Gamma alternatives, as shown by Table 4, which has the same settings as Table 3. $A_n^{'}$ outperforms all the other considered tests, notably the Cramér–Von Mises test. The Anderson–Darling test has generally a worse power than $A_n^{'}$, even if it becomes its main competitor as n increases. It has to be emphasized that the simulated powers reported for Gamma alternatives in Table 4 cannot be compared with the ones reported in Table 3, as the null distribution is quite different in the two sets of simulations. More specifically, when the null model is the Normal distribution, the power of each considered test is a decreasing function of the parameter a in (18); conversely, when the null model is the unit exponential distribution, the power is an increasing function, at least if $a > 1$. This fact explains why, even if the sample size and the value of a may coincide, Table 4 and Table 3 report quite different values of the estimated powers. Turning to other distributions considered in Table 4, one can notice that the above conclusions are reversed for alternatives of the generalized-Pareto kind, as D_n outperforms here all the other tests; $A_n^{'}$ has a similar power to the one of the Cramér–Von Mises test, but it seems to worsen as n increases. The Weibull alternatives evidence a problem of bias for some tests under some circumstances; apart from this fact, this case resembles the Student's T alternatives of Table 3: except for $n = 5$, the test based on D_n has definitely the best power, but A_n' clearly outperforms the Cramér–Von Mises and the Kolmogorov–Smirnov tests.

Table 4. Simulated powers when the null model is a unit exponential distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	A_n'	K_n	C_n	D_n
		.0094	.0100	.0097	.0095
10	null	.0494	.0495	.0494	.0492
		.1001	.0992	.1001	.0989
		.1515	.1152	.1437	.1444
10	GA(1.5)	**.3560**	.2891	.3428	.3420
		.4827	.4229	.4710	.4684
		.6191	.4863	.5928	.6048
10	GA(2)	**.8388**	.7511	.8222	.8269
		.9088	.8507	.8988	.9004
		.0021	**.0144**	.0049	.0010
10	W(2)	.0618	**.0984**	.0712	.0360
		.1951	**.2036**	.1903	.1315
		.0038	**.0501**	.0105	.0007
10	W(3)	.2587	**.2889**	.2579	.1469
		.5906	.4925	.5440	.4554
		.0269	.0215	.0257	**.0871**
10	GP(0.35)	.0946	.0829	.0908	**.2048**
		.1659	.1486	.1585	**.3019**
		.0411	.0323	.0337	.1469
10	GP(0.45)	.1179	.1018	.1119	**.2830**
		.1892	.1964	.1830	**.3821**
		.1098	.0818	.1018	.1039
25	GA(1.25)	**.2772**	.2249	.2645	.2649
		.3882	.3334	.3764	.3765
		.4973	.3752	.4711	.4811
25	GA(1.5)	**.7339**	.6351	.7149	.7226
		.8281	.7538	.8140	.8201
		.0224	**.0554**	.0326	.0235
25	W(1.75)	.2296	.2205	.2199	**.2317**
		.4582	.3806	.4291	**.4618**
		.0708	**.1160**	.0818	.0689
25	W(2)	.4462	.3758	.4174	**.4549**
		.7093	.5743	.6680	**.7231**
		.0422	.0344	.0395	**.1456**
25	GP(0.35)	.1339	.1153	.1271	**.3028**
		.2168	.2007	.2152	**.4181**
		.0647	.0531	.0616	**.2599**
25	GP(0.45)	.1743	.1564	.1696	**.4433**
		.2705	.2531	.2672	**.5603**

Table 4 (continued). Simulated powers when the null model is a unit exponential distribution ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
		.1843	.1332	.1737	.1829
100	GA(1.15)	**.3945**	.3228	.3812	.3916
		.5202	.4504	.5064	.5144
		.5705	.4317	.5476	.5675
100	GA(1.25)	**.7873**	.6902	.7700	.7868
		.8705	.7996	.8578	.8696
		.2994	.2756	.2920	**.4657**
100	W(1.5)	.7534	.6165	.7201	**.8719**
		.9046	.7945	.8842	**.9605**
		.8391	.7054	.8199	**.9463**
100	W(1.75)	.9880	.9462	.9832	**.9986**
		.9982	.9850	.9977	**.9999**
		.1441	.1424	.1504	**.5048**
100	GP(0.35)	.3378	.3432	.3488	**.7313**
		.4754	.4844	.4865	**.8288**
		.2359	.2712	.2604	**.7616**
100	GP(0.45)	.4807	.5287	.5124	**.9028**
		.6251	.6670	.6529	**.9461**

In the simulations reported in Table 4, both the null and the alternative distributions are skewed. To consider other cases where a null symmetric model is to be tested against skewed alternatives, like for the above standard Normal case, a third set of simulations is finally reported. The uniform distribution on the unit interval $(0, 1)$ is used as a null model. Some "modifications" of the uniform density are considered as alternatives. The first has density

$$f(x) = \begin{cases} a\ (2x)^{a-1} & 0 \leq x \leq 0.5, \\ a\ (2-2x)^{a-1} & 0.5 \leq x \leq 1, \end{cases} \tag{22}$$

(where $a > 0$) and it is labeled as MU; the second,

$$f(x) = (1-2a)^{-1}, \qquad a < x < 1-a, \tag{23}$$

is essentially a "compressed" uniform (CU) distribution over the interval $(a, 1-a)$, where $0 \leq a \leq 1/2$. Both densities reduces to the uniform distribution on the unit interval when $a = 0$. They were drawn from the study conducted by Schmid and Trede (1996). To complete such an investigation, a third alternative family is here considered for F_1:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} x^{a-1}(1-x)^{b-1}, \qquad 0 < x < 1. \tag{24}$$

The Beta (B) density in (24) reduces to the uniform distribution on $(0, 1)$ when $a = b = 1$. In the reported simulation study, b is then set to 1 and $a > 0$ is left to vary. Notice that, as a grows over 1, the distribution becomes more skewed, thus giving exactly the needed kinds of alternatives.

Table 5. Simulated powers when the null model is a uniform distribution on the unit interval ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
		.0100	.0097	.0101	.0098
10	null	.0507	.0501	.0501	.0503
		.1000	.1001	.0998	.1004
		.0001	.0047	.0001	.0000
10	MU(4.5)	.2705	**.2974**	.2263	.1042
		.6944	.4963	.6357	.5301
		.0001	.0051	.0001	.0000
10	MU(5)	**.3862**	.2588	.3115	.1569
		.7963	.5793	.7432	.6469
		.0001	.0032	.0003	.0000
10	CU(0.3)	**.1621**	.0890	.1219	.0480
		.6449	.3176	.5732	.4564
		.0002	.0039	.0005	.0000
10	CU(0.35)	**.5598**	.2214	.4939	.2308
		.9814	.8019	.9931	.9715
		.1951	.1556	.1898	.1686
10	B(2)	**.4501**	.3830	.4378	.4085
		.5941	.5289	.5829	.5548
		.6371	.5259	.6251	.5755
10	B(3)	**.8780**	.8048	.8689	.8451
		.9408	.8998	.9372	.9243
		.3870	.2725	.3411	.3740
25	MU(3.5)	.9392	.7778	.9174	**.9474**
		.9922	.9349	.9873	**.9940**
		.6321	.4205	.5783	.6197
25	MU(4)	**.9874**	.8985	.9812	.9906
		.9992	.9811	.9982	**.9994**
		.0242	.0234	.0202	.0232
25	CU(0.2)	.4149	.2214	.3596	**.5457**
		.7733	.5318	.7540	**.9305**
		.0574	.0432	.0466	.0620
25	CU(0.225)	**.6697**	.4173	.6478	.8581
		.9374	.8782	.9554	.9985
		.1833	.1386	.1740	.1661
25	B(1.5)	**.4043**	.3390	.3935	.3831
		.5395	.4766	.5280	.5210
		.6650	.5449	.6490	.6319
25	B(2)	**.8694**	.7972	.8621	.8561
		.9314	.8894	.9265	.9244
		.0648	.0901	.0668	**.1444**
100	MU(1.5)	.4222	.3506	.3963	**.5674**
		.6748	.5539	.6434	**.7775**

Table 5 (continued). Simulated powers when the null model is a uniform distribution on the unit interval ($\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	K_n	C_n	D_n
		.3958	.3005	.3642	**.6086**
100	MU(1.75)	.8715	.7177	.8446	**.9441**
		.9659	.8822	.9549	**.9871**
		.0805	.0474	.0635	**.3127**
100	CU(0.1)	.4913	.2932	.4456	**.9256**
		.7670	.5883	.7563	**.9996**
		.2246	.1208	.1959	**.7489**
100	CU(0.12)	.7880	.6817	.8031	**.9996**
		.9500	.9990	.9712	**.9999**
		.2458	.1851	.2351	.2413
100	B(1.25)	**.4830**	.4078	.4704	.4801
		.6123	.5431	.6008	.6099
		.8400	.7374	.8276	**.8410**
100	B(1.5)	.9523	.9100	.9474	**.9544**
		.9777	.9563	.9755	**.9796**

Table 5 reports some results of this last set of simulations. The alternatives of kind (22) and (23) evidence that all the considered tests suffer from a problem of bias, to which the Anderson–Darling test seems to be the most exposed. A second remark is that the performance of $A_n^{'}$ is similar to the one of C_n, even if the former has almost everywhere a higher estimated power. Both the Girone–Cifarelli and the Cramér–Von Mises test are outperformed by the Anderson–Darling test for as large sample sizes as $n = 100$ (for alternatives (23) even from $n = 25$). These conclusions add few extra details to the ones obtained by Schmid and Trede (1996) for the test based on A_n in (15), which has actually a performance similar to the Girone–Cifarelli test. However, the alternatives of the Beta class represent a considerable addition in the evaluation of such tests of uniformity: the power of $A_n^{'}$, as long as the one of A_n (unreported), is here steadily over the one of the other considered tests, notably over the one of C_n, with some minor exceptions for D_n. Even if Table 5 reports just the results for selected values of the parameter a in (24), the simulation study showed the Girone–Cifarelli test to be uniformly more powerful than the Cramér–Von Mises test for $a > 1$.

The discussion of Table 5 raises an important issue to be considered before giving some general conclusions in the next section. As stated from the very beginning, the Girone–Cifarelli test performs often similarly to the test based on A_n in (15); this fact resulted clearly from the conducted simulation study and it is essentially the reason why no separate results about A_n are reported in the above discussion. However, the two test-statistics $A_n^{'}$ and A_n are not equivalent, as evidenced by the following simple decomposition:

$$A_n = \frac{2}{n} \sum_{i \in \overline{\mathcal{A}}} \left| F_0(x_{(i)}) - \frac{i-1/2}{n} \right| + 2 \sum_{i \in \mathcal{A}} \left[\left(F_0(x_{(i)}) - \frac{i}{n} \right) \left(F_0(x_{(i)}) - \frac{i-1}{n} \right) + \frac{1}{2n^2} \right], \tag{25}$$

where $\mathcal{A} \equiv \left\{ i : \frac{i-1}{n} < F_0(x_{(i)}) < \frac{i}{n} \right\}$. Notice that the set $\mathcal{A}$ is not empty (and thus $A_n^{'}$ and A_n are not equivalent) as long as the empirical distribution function is not dominated by the null model F_0 (or conversely). Hence the possible differences in the powers of $A_n^{'}$ and A_n are likely to be observed when the alternative distribution does not dominate the null model (or conversely), a fact that can be partially guaranteed by letting the two distributions have the same location. A last set of simulations was then conducted where the alternative distribution was forced to have the same mean of the null model. In effect, some of the above-reported alternative distributions do not guarantee such a requirement. In addition, small values of the sample size were chosen, as the effect of the second summand in (25) is likely to decrease with n. Table 6 reports some results when $A_n^{'}$ and A_n are used to test unit exponentiality against other skewed alternatives, a situation which proved to be good for both tests against their classical competitors. On the average, $A_n^{'}$ turns out to perform still similarly to A_n, even if there are cases where the difference in their powers becomes relevant. Notice that, with some minor exceptions, the Girone–Cifarelli test has never a lower power with respect to A_n. $A_n^{'}$ outperforms A_n for Gamma and notably for generalized-Pareto alternatives. The Weibull case is less definite, as $A_n^{'}$ has only a minor advantage over A_n and not for very small same sizes.

Table 6. Comparison between the powers of $A_n^{'}$ and A_n when the null and the alternative distributions have the same location (H_0 = unit exponential, $\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	A_n
		.0625	.0615
5	GA(2)	.1534	.1504
		.2464	.2411
		.0885	.0874
5	GA(3)	.1909	.1866
		.3120	.3043
		.1058	.1045
5	GA(4)	.2167	.2106
		.3601	.3503
		.0000	.0000
5	W(3)	.0354	.0382
		.2207	.2278
		.0000	.0000
5	W(4)	.0353	.0459
		.3403	.3526
		.0000	.0000
5	W(5)	.0401	.0572
		.4746	.4870
		.2889	.2866
5	GP(0.45)	.4406	.4323
		.5734	.5635
		.3162	.3138
5	GP(0.47)	.4659	.4574
		.6032	.5937
		.3428	.3409
5	GP(0.49)	.4953	.4864
		.6322	.6226
		.0624	.0618
7	GA(2)	.1665	.1645
		.2778	.2749
		.0849	.0839
7	GA(3)	.2275	.2242
		.3745	.3709
		.1030	.1019
7	GA(4)	.2745	.2698
		.4475	.4418
		.0014	.0015
7	W(3)	.1873	.1876
		.5175	.5172

Table 6 (continued). Comparison between the powers of $A_n^{'}$ and A_n when the null and the alternative distributions have the same location (H_0 = unit exponential, $\alpha = 0.01, 0.05, 0.1$)

n	H_1	$A_n^{'}$	A_n
		.0007	.0007
7	W(4)	.3404	.3422
		.7563	.7557
		.0005	.0004
7	W(5)	.5029	.5108
		.8934	.8941
		.3406	.3386
7	GP(0.45)	.5568	.5522
		.6969	.6937
		.3689	.3666
7	GP(0.47)	.5898	.5853
		.7280	.7246
		.3999	.3971
7	GP(0.49)	.6216	.6167
		.7578	.7550
		.0643	.0638
9	GA(2)	.1834	.1815
		.3042	.3014
		.0933	.0927
9	GA(3)	.2655	.2623
		.4319	.4276
		.1147	.1140
9	GA(4)	.3358	.3314
		.5240	.5189
		.0163	.0168
9	W(3)	.4135	.4128
		.7501	.7486
		.0297	.0305
9	W(4)	.6895	.6878
		.9370	.9365
		.0546	.0539
9	W(5)	.8668	.8649
		.9874	.9871
		.4183	.4166
9	GP(0.45)	.6565	.6532
		.7814	.7778
		.4593	.4579
9	GP(0.47)	.6923	.6891
		.8124	.8097
		.4943	.4927
9	GP(0.49)	.7261	.7232
		.8381	.8355

4. A Real-Data Example and Some Conclusions

Before drawing some conclusions, a simple example where the test proposed in this paper is applied to real data is reported. The "warp breaks" dataset in Pearson (1963) has fast become a term of comparison of the results of various goodness-of-fit tests where the null distribution F_0 is completely specified. In this case, one has to test if the places where some warp breaks occur on a loom can be considered as uniformly distributed on the whole length of the warp. More specifically, the following distances of $n = 20$ breaks from the beginning of the warp are recorded: 30, 36, 104, 286, 291, 658, 893, 955, 1149, 1195, 1208, 1240, 1277, 1282, 1363, 1384, 1421, 1477, 1504, 1510 (see Pearson, 1963 for further details) and the ratios of these distances with respect to the total length (1520) are considered; a goodness-of-fit test is then applied to verify if such a sample of ratios comes from a population with unit uniform distribution. The observed value of $A_n^{'}$ is 2.9964, so that $n^{(-1/2)}A_n^{'} = 0.6633$ and, by looking at Table 1, the null hypothesis is to be rejected at the 5%-level but not at the 1%-level. More specifically, by using the simulated null distribution of $A_n^{'}$, a p-value 0.0213 is obtained. As a term of comparison, the p-values of the other considered tests are: 0.0090 for the Kolmogorov–Smirnov test, 0.0156 for the Cramér–Von Mises test and 0.0110 for the Anderson–Darling test. The results of all tests are then consistent, even if some differences can be appreciated.

This paper presents a simulations study which gives some new insights about goodness-of-fit tests based on the empirical distribution function. The main conclusion is that a good analysis should never neglect tests based on the averaged absolute difference $|F_n(t) - F_0(t)|$. The tests based on $A_n^{'}$ and A_n will both serve at this aim, even if the former can give some slight advantages over the latter, at least for small sample sizes. Moreover, the test-statistic $A_n^{'}$ has a rather simple form and it can be computed very easily. A second important conclusion is that $A_n^{'}$ (and A_n) has very often a different performance from the one of C_n, which is based on the averaged squared difference $[F_n(t) - F_0(t)]^2$. The reported simulations give a good evidence of such alternatives where $A_n^{'}$ outperforms C_n. It seems, specifically, that this happens more frequently for skewed alternatives. Concerning the Kolmogorov–Smirnov test K_n, which takes into consideration the supremum and not an average of the difference $|F_n(t)-F_0(t)|$, the reported study shows that there are few practical situations where it performs better than the other considered tests, and notably than $A_n^{'}$. The superiority, under some circumstances, of the Girone–Cifarelli test over the Kolmogorov–Smirnov test has been evidenced, in effect, in other studies concerning the two-samples problem. As a last issue, one can claim that the real competitor of $A_n^{'}$ (and similarly for A_n) is the Anderson–Darling test D_n, rather than K_n or C_n. The discussion in this paper shows that there are cases where D_n outperforms all other considered tests and that it leaves $A_n^{'}$ as a second best. These are mainly cases of alternatives with heavy tails, probably thanks to the weighting function in the definition of D_n. An important element of a future research could then be to evaluate the effect of the introduction of suitable weighting functions in the definition of $A_n^{'}$ as well.

References

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *Ann. Math. Statist., 33*, 1148-1159. http://dx.doi.org/10.1214/aoms/1177704477

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist., 12*, 171-178.

Borroni, C. G. (2001). Some notes about nonparametric tests for the equality of two populations. *Test, 10*, 147-159. http://dx.doi.org/10.1007/BF02595829

Brunk, H. D. (1962). On the range of the difference between hypothetical distribution function and Pyke's modified empirical distribution function. *Ann. Math. Statist., 33*, 525-532. http://dx.doi.org/10.1214/aoms/1177704578

Cifarelli, D. M. (1974). Intorno alla distribuzione del test di G. Girone. *Giornale degli Economisti e Annali di Economia, 33*, 283-308.

Cifarelli, D. M. (1975). Contributi intorno ad un test per l'omogeneità tra due campioni. *Giornale degli Economisti e Annali di Economia, 34*, 233-249.

Girone, G. (1964). Su un indice di omogeneità di due distribuzioni del tipo dell'indice semplice di dissomiglianza. *Atti della XXIV Riunione Scientifica della SIS*, 53-78.

Goria, M. (1972). Some generalizations of Girone's test statistic. *Annali dell'Istituto di Statistica - Università di Bari*, 53-72.

Green, J. R., & Hegazy, Y. A. S. (1976). Powerful modified-EDF goodness-of-fit tests. *J. Amer. Statist. Assoc., 71*, 204-209. http://dx.doi.org/10.1080/01621459.1976.10481516

Johnson, B., & Killeen, T. (1983). An explicit formula for the c.d.f. of the L1 norm of the Brownian bridge. *Ann. Prob., 11*, 807-808. http://dx.doi.org/10.1214/aop/1176993528

Pearson, E. S. (1963). Comparison of tests for randomness of points on a line. *Biometrika, 50*, 315-325.

Pyke, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Statist., 30*, 568-576. http://dx.doi.org/10.1214/aoms/1177706269

Schmid, F., & Trede, M. (1995). A distribution free test for the two sample problem for general alternatives. *Comput. Stat. Data Anal., 20*, 409-419. http://dx.doi.org/10.1016/0167-9473(95)92844-N

Schmid, F., & Trede, M. (1996). An L1-variant of the Cramér-von Mises test. *Stat. Probabil. Lett., 26*, 91-96. http://dx.doi.org/10.1016/0167-7152(95)00256-1

Shepp, L. A. (1982). On the integral of the absolute value of the pinned Wiener process. *Ann. Prob., 10*, 234-239. http://dx.doi.org/10.1214/aop/1176993926

Shepp, L. A. (1991). Acknowledgement of priority. *Ann. Prob., 19*, 1397. http://dx.doi.org/10.1214/aop/1176990351

Takács, L. (1993). On the distribution of the integral of the absolute value of the Brownian motion. *Ann. Appl. Probab., 3*, 186-197.