J. R. Statist. Soc. B (2003)
65, *Part* 1, *pp.* 115–126

Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions

J. Fortiana and A. Grané

Universitat de Barcelona, Spain

[Received February 2001. Revised July 2002]

Summary. We propose a goodness-of-fit statistic Q_n based on the Hoeffding maximum correlation for testing uniformity and we show its relationship to Gini's mean difference. We compute exact and asymptotic critical values and study the power of the test proposed against a representative set of alternatives.

Keywords: Goodness of fit; *L*-statistics; Maximum correlation; Orthogonal decomposition of statistics

1. Introduction

Let F_1 and F_2 be two cumulative distribution functions (CDFs) having second-order moments. The *maximum correlation* (Hoeffding) of (F_1, F_2) is defined as the correlation coefficient $\rho^+(F_1, F_2)$ corresponding to the bivariate CDF $H^+(x, y) = \min\{F_1(x), F_2(y)\}$, the *upper Fréchet bound* of F_1 and F_2, i.e. the upper bound of the Fréchet class $\mathcal{F}(F_1, F_2)$ of bivariate CDFs with marginals F_1 and F_2, ordered according to their correlation coefficients. The CDF H^+ is a singular distribution, having support on the one-dimensional set $\{(x, y) \in \mathbb{R}^2 : F_1(x) = F_2(y)\}$, and its correlation coefficient is given by

$$\rho^+(F_1, F_2) = \left\{ \int_0^1 F_1^-(p)\, F_2^-(p)\, \mathrm{d}p - \mu_1\mu_2 \right\} \Big/ \sigma_1\sigma_2, \tag{1}$$

where F_i^- is the left continuous pseudoinverse of F_i, $\mu_i = E(F_i)$ and $\sigma_i^2 = \mathrm{var}(F_i)$, $i = 1, 2$ (see, for example, Cambanis *et al.* (1976)).

The maximum correlation $\rho^+(F_1, F_2)$ is a measure of agreement between F_1 and F_2, since $\rho^+ = 1$ if and only if $F_1 = F_2$ (almost everywhere) up to a scale and location change. In particular, Cuadras and Fortiana (1993) proposed the statistic $\rho^+(F_n, F_0)$ as a qualitative measure of goodness of fit of an independent and identically distributed (IID) sample $x_1, \ldots, x_n$, with empirical distribution function F_n, to a given distribution F_0. This paper is devoted to testing uniformity, i.e. $F_0 = F_{\mathrm{U}}$, a [0, 1] uniform distribution, a test which extends to case 0 in Stephens (1986), testing for a fully specified continuous distribution.

In Section 2 we define the geometric variability of a probability distribution with respect to a metric d, showing that, for a given CDF F, $\rho^+(F, F_{\mathrm{U}})$ is an instance of this quantity. In

Address for correspondence: J. Fortiana, Departament d'Estadística, Facultat de Matemàtiques, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007-Barcelona, Spain.
E-mail: fortiana@mat.ub.es

 1369–7412/03/65115

particular, it follows that

$$\rho^+(F_n, F_U) = \frac{\sqrt{3}}{2s_n}\frac{n-1}{n}G_n,$$

where s_n^2 is the empirical variance and G_n is the well-known Gini mean difference statistic. This property, together with other technical considerations, leads us to define as our test statistic

$$Q_n = \frac{s_n}{\sqrt{(1/12)}}\rho^+(F_n, F_U) = 3\frac{n-1}{n}G_n,$$

which is asymptotically equivalent to $\rho^+(F_n, F_U)$.

In Section 3 we obtain a decomposition of Q_n in terms of a sequence of statistics $\{\beta_{nj}\}_{j\in\mathbb{N}}$. This construction is analogous to those studied by Anderson and Darling (1952, 1954), Durbin and Knott (1972), Durbin *et al.* (1975) and Stephens (1974).

In Section 4 we show that both Q_n and the $\{\beta_{nj}\}_{j\in\mathbb{N}}$ are L-statistics, and we study small sample and asymptotic properties of their distributions under the null hypothesis. In particular we give an explicit algorithm to compute critical regions.

In Section 5 we study the power of the test based on Q_n against a wide set of alternatives. We also depict some power functions to illustrate the results.

Tables 1 and 2 are tables of critical values, obtained from the exact probability distribution (for $n \leqslant 50$) and from the asymptotic probability distribution. For the latter, the relative error is shown in parentheses.

A Mathematica implementation of the algorithm that was used to compute the critical regions and a Matlab–Octave program performing the test proposed are available from

`http://www.blackwellpublishers.co.uk/rss/`

2. Geometric variability and Gini's mean difference

The *geometric variability* of a CDF F with respect to a metric $d(\cdot,\cdot)$ on $\mathbb{R}$ is defined by

$$V_F(d) = \frac{1}{2}\int_{\mathbb{R}^2} d^2(x, y)\,\mathrm{d}F(x)\,\mathrm{d}F(y).$$

This measure of dispersion, introduced by Cuadras and Fortiana (1995), extends the concept of variance, which appears as a particular case, when d is the usual l^2 Euclidean metric. It turns out that the maximum correlation $\rho^+(F, F_U)$ is closely related to $V_F(d)$, for a particular metric, as shown in the following proposition.

Proposition 1. Let F be a CDF on $\mathbb{R}$, with second-order moment. Then

$$\rho^+(F, F_U) = \frac{\sqrt{3}}{\sigma}V_F(d), \tag{2}$$

where $d(x, y) = \sqrt{|x-y|}$ and $\sigma^2 = \mathrm{var}(F)$.

To prove this statement, we shall need the following result, which is an immediate application of the change of variables theorem, taking into account the inverse probability transformation property of one-dimensional distributions.

Lemma 1. Let F be a CDF on $\mathbb{R}$, and let F^- be its left continuous pseudoinverse. A measurable function $g : \mathbb{R} \to \mathbb{R}$ is integrable with respect to F if and only if $g \circ F^-$ is integrable with respect to the Lebesgue measure on [0, 1], and in such cases

$$\int_{\mathbb{R}} g(x)\,\mathrm{d}F(x) = \int_{[0,1]} (g \circ F^-)(t)\,\mathrm{d}t.$$

Table 1. Two-tail exact critical values of Q_n at the 5% significance level

n	*Lower tail*	*Upper tail*	*n*	*Lower tail*	*Upper tail*	*n*	*Lower tail*	*Upper tail*
2	0.01887	1.26283	19	0.72416	1.14345	35	0.81208	1.11562
3	0.12573	1.20760	20	0.73310	1.14103	36	0.81530	1.11441
4	0.24049	1.20177	21	0.74131	1.13875	37	0.81836	1.11323
5	0.33397	1.20121	22	0.74887	1.13657	38	0.82129	1.11209
6	0.40709	1.19460	23	0.75586	1.13451	39	0.82409	1.11099
7	0.46459	1.18852	24	0.76235	1.13254	40	0.82678	1.10992
8	0.51056	1.18321	25	0.76839	1.13066	41	0.82935	1.10888
9	0.54796	1.17819	26	0.77403	1.12887	42	0.83182	1.10788
10	0.57894	1.17350	27	0.77931	1.12715	43	0.83420	1.10690
11	0.60502	1.16915	28	0.78427	1.12550	44	0.83648	1.10595
12	0.62730	1.16513	29	0.78894	1.12392	45	0.83867	1.10503
13	0.64658	1.16139	30	0.79334	1.12240	46	0.84079	1.10413
14	0.66346	1.15790	31	0.79750	1.12095	47	0.84283	1.10326
15	0.67837	1.15463	32	0.80143	1.11954	48	0.84480	1.10241
16	0.69166	1.15158	33	0.80516	1.11819	49	0.84670	1.10158
17	0.70359	1.14870	34	0.80871	1.11688	50	0.84854	1.10077
18	0.71437	1.14600						

Table 2. Two-tail asymptotic critical values of Q_n at the 5% significance level†

n	*Lower tail*	*Upper tail*	*n*	*Lower tail*	*Upper tail*
5	0.35651 (6.3)	1.24349 (3.4)	30	0.80165 (1.0)	1.13169 (0.8)
10	0.60018 (3.5)	1.19981 (2.1)	35	0.81927 (0.9)	1.12356 (0.7)
15	0.69382 (2.2)	1.17287 (1.6)	40	0.83311 (0.8)	1.11689 (0.6)
20	0.74514 (1.6)	1.15486 (1.2)	45	0.84434 (0.7)	1.11122 (0.6)
25	0.77822 (1.3)	1.14177 (1.0)	50	0.85366 (0.6)	1.10634 (0.5)

†The relative error (as a percentage) is shown in parentheses.

We observe that, in particular,

$$E(F) = \int_{[0,1]} F^-(t)\,\mathrm{d}t.$$

Proof of proposition 1.

$$V_F(d) = \frac{1}{2}\int_{[0,1]^2} |F^-(s) - F^-(t)|\,\mathrm{d}s\,\mathrm{d}t = \int_{[0,1]^2 \cap \{t<s\}} \{F^-(s) - F^-(t)\}\,\mathrm{d}s\,\mathrm{d}t,$$

since the integrand is a symmetric function, null on the diagonal $\{t = s\}$, and F^- is non-decreasing. Integration by parts on s in

$$V_F(d) = \int_0^1 \left\{ s\,F^-(s) - \int_0^s F^-(t)\,\mathrm{d}t \right\} \mathrm{d}s$$

gives

$$V_F(d) = \int_0^1 (2s-1)\,F^-(s)\,\mathrm{d}s.$$

Comparing this expression with

$$\rho^+(F, F_U) = \left\{ \int_0^1 t\, F^-(t)\, \mathrm{d}t - \frac{1}{2} E(F) \right\} \Big/ \sqrt{\left(\frac{1}{12}\right)}\sigma$$

we obtain equation (2). □

In the following, $x_1, \ldots, x_n$ will be an IID sample, $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$ the ordered sample, F_n its empirical CDF and $\bar{x}_n = E(F_n)$ and ${s_n}^2 = \mathrm{var}(F_n)$ the empirical mean and variance respectively. The statistic $\rho^+(F_n, F_U)$ has been used as a qualitative measure of goodness of fit (Cuadras and Fortiana, 1993, 1994). However, the behaviour of $\rho^+(F_n, F_U)$ leaves much to be desired: for finite n, its exact probability density function (PDF) under the null hypothesis is unknown, requiring simulation methods to compute critical regions which, in turn, renders quite impracticable any comprehensive evaluation of power. The PDF is a source of additional, practical, computational difficulties: it is wedge or spike shaped, strongly concentrated near 1. Furthermore, it is asymmetric, truncated at 1. An asymptotic approximation is available, but convergence to its limiting law is rather slow. These difficulties can be circumvented by considering the statistic

$$Q_n = \frac{s_n}{\sqrt{(1/12)}} \rho^+(F_n, F_U).$$

Both statistics have the same limiting law under the null hypothesis, since

$$\lim_{n\to\infty} \{s_n/\sqrt{(1/12)}\} = 1$$

(almost surely). Since proposition 3 in Section 4 shows that Q_n is an L-statistic of the [0, 1] uniform distribution it is possible to compute the exact PDF and critical regions applying the Dwass (1961), Matsunawa (1985) and Ramallingam (1989) algorithm (see Section 4.1).

Additionally, from proposition 1 it follows that $Q_n = 3(n-1)/nG_n$, where

$$G_n = \frac{1}{2} n(n-1) \sum_{1 \leqslant i < j \leqslant n} |x_i - x_j|$$

is the well-known *Gini mean difference* statistic.

3. Principal components

The principal axes of a uniform [0, 1] random variable U, with respect to the metric $d(s, t) = \sqrt{|s-t|}$, were defined by Cuadras and Fortiana (1994) as a sequence of random variables $\{C_j\}_{j\in\mathbb{N}}$ such that

$$\sum_{j=1}^{\infty} \{C_j(t) - C_j(s)\}^2 = d^2(s, t) = |s - t|.$$

The standardized principal axes are given by $C_j^* = -\sqrt{2}\, \cos(j\pi U)$, $j \geqslant 1$.

The coefficient of the projection of a second-order random variable X on the jth standardized principal axis is

$$\beta_j = E(XC_j^*) = -\int x\sqrt{2}\, \cos(j\pi u)\, \mathrm{d}H(u, x), \qquad j \geqslant 1, \tag{3}$$

where H is the bivariate distribution function of (U, X).

Given a CDF F with second-order moments, $\{\beta_j(F)\}_{j\in\mathbb{N}}$ will denote the sequence of coefficients (3), with $H = H^+$, i.e.

$$\beta_j(F) = -\int_0^1 F^-(u)\sqrt{2}\,\cos(j\pi u)\,\mathrm{d}u, \qquad j \geqslant 1. \tag{4}$$

$F^- \in L^2([0,1])$, and $-\beta_j(F)$ is the jth Fourier coefficient in its expansion with respect to the orthonormal basis $\{\sqrt{2}\,\cos(j\pi u)\}_{j\in\mathbb{N}}$, $u \in [0,1]$.

Proposition 2. Let F be a CDF on $\mathbb{R}$, with second-order moments. We have the expansion

$$\rho^+(F, F_{\mathrm{U}}) = \frac{4\sqrt{6}}{\pi^2\sigma}\sum_{j=0}^{\infty}\frac{\beta_{2j+1}(F)}{(2j+1)^2} \tag{5}$$

where $\sigma^2 = \mathrm{var}(F)$ and F_{U} is the $[0,1]$ uniform distribution.

Proof. The standardized inverse of F_{U} (i.e. the inverse function of a uniform distribution with zero mean and unit variance) is $\varphi(t) = 2\sqrt{3}(t - \frac{1}{2})$, $t \in (0,1)$. From equation (1),

$$\rho^+(F, F_{\mathrm{U}}) = \frac{1}{\sigma}\int_0^1 2F^-(t)\,\sqrt{3}\left(t - \frac{1}{2}\right)\mathrm{d}t. \tag{6}$$

The jth Fourier coefficient of $\varphi(t)$, relative to the complete orthonormal set $\{\sqrt{2}\,\cos(j\pi t)\}_{j\in\mathbb{N}}$, is

$$\gamma_j = \int_0^1 2\sqrt{3}(t-\tfrac{1}{2})\sqrt{2}\,\cos(j\pi t)\,\mathrm{d}t = \begin{cases} 0 & \text{if } j \text{ is even,} \\ -4\sqrt{6}/(j\pi)^2 & \text{if } j \text{ is odd.} \end{cases}$$

Substituting in equation (6),

$$\rho^+(F, F_{\mathrm{U}}) = \frac{1}{\sigma}\int_0^1 F^-(t)\sum_{j\geqslant 0}\gamma_j\sqrt{2}\,\cos(j\pi t)\,\mathrm{d}t = -\frac{1}{\sigma}\sum_{j\geqslant 0}\gamma_j\,\beta_j(F). \qquad \square$$

Substituting F_n for F in equation (4) we obtain a sequence $\{\beta_{nj} = \beta_j(F_n)\}_{j\geqslant 1}$ of statistics, in terms of which Q_n admits the expansion

$$Q_n = \frac{24\sqrt{2}}{\pi^2}\sum_{j\geqslant 0}\frac{\beta_{n,2j+1}}{(2j+1)^2}.$$

4. Distribution under the null hypothesis

Proposition 3. Q_n and β_{nj} are L-statistics, $\Sigma_{i=1}^n a_i x_{(i)}$, with coefficients

$$a_i = \begin{cases} \dfrac{6}{n^2}(2i - n - 1), & 1 \leqslant i \leqslant n, \text{ for } Q_n, \\ \dfrac{\sqrt{2}}{j\pi}\left[\sin\left\{\dfrac{(i-1)j\pi}{n}\right\} - \sin\left(\dfrac{ij\pi}{n}\right)\right], & 1 \leqslant i \leqslant n, \text{ for } \beta_{nj}. \end{cases} \tag{7}$$

Proof. The representation of Gini's mean difference coefficient as an L-statistic is well known (see Serfling (1980), page 263). Alternatively, from equation (1),

$$\rho^+(F_n, F_{\mathrm{U}}) = \frac{1}{\sqrt{(1/12)s_n}}\left\{\int_0^1 F_n^-(p)\,F_{\mathrm{U}}^-(p)\,\mathrm{d}p - \frac{1}{2}\bar{x}_n\right\},$$

where F_n^- is the left continuous pseudoinverse of F_n. The numerator equals

$$\sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} p x_{(i+1)} \, \mathrm{d}p - \frac{1}{2n} \sum_{i=1}^{n} x_{(i)} = \sum_{i=0}^{n-1} \frac{x_{(i+1)}}{2} \frac{(i+1)^2 - i^2}{n^2} - \frac{1}{2n} \sum_{i=1}^{n} x_{(i)};$$

hence

$$\rho^+(F_n, F_{\mathrm{U}}) = \frac{\sqrt{(1/12)}}{s_n} \frac{6}{n^2} \sum_{i=1}^{n} (2i - n - 1) x_{(i)}.$$

Similarly,

$$\begin{aligned} \beta_{nj} = \beta_j(F_n) &= -\int_0^1 F_n^-(p) \sqrt{2} \cos(j\pi p) \, \mathrm{d}p \\ &= -\frac{\sqrt{2}}{j\pi} \sum_{i=0}^{n-1} x_{(i+1)} \left[\sin\left\{ \frac{(i+1)j\pi}{n} \right\} - \sin\left(\frac{ij\pi}{n} \right) \right]. \end{aligned}$$ □

It is worth noting that expression (7) is equivalent to

$$a_i = \frac{-2\sqrt{2}}{j\pi} \sin\left(\frac{j\pi}{2n} \right) \cos\left(\frac{i - \frac{1}{2}}{n} j\pi \right)$$

which, for a given j, reminds us of the ith term of the trigonometric expansion of a function. Each such term can be seen as a contribution to the goodness-of-fit statistic. This analogy is further explored in Fortiana and Grané (2000).

Using expression (7) it can be shown that the statistics β_{nj} have the periodicity

$$j\beta_{nj} = (j + 2n)\beta_{n,\, j+2n}, \qquad j \geqslant 1.$$

4.1. Small sample properties

The supports of the PDFs of Q_n and β_{nj} depend on the coefficients $b_i = \Sigma_{l=i}^n a_l$, $1 \leqslant i \leqslant n$:

$$\mathrm{supp}(Q_n) = \begin{cases} [0, 3/2], & \text{even } n, \\ [0, 3/2(1 - 1/n^2)], & \text{odd } n, \end{cases}$$

and

$$\mathrm{supp}(\beta_{nj}) = \left[\min_{1 \leqslant i \leqslant n} (b_i), \max_{1 \leqslant i \leqslant n} (b_i) \right],$$

where

$$b_i = \frac{\sqrt{2}}{j\pi} \sin\left\{ \frac{(i-1)j\pi}{n} \right\}.$$

The expected values and variances of Q_n and β_{nj} under the null hypothesis are given in the following proposition.

Proposition 4.

$$\begin{aligned} E(Q_n) &= \frac{n-1}{n}, \\ \mathrm{var}(Q_n) &= \frac{n^3 + 4n^2 + n - 6}{5n^3(n+2)}, \end{aligned} \tag{8}$$

$$E(\beta_{nj}) = \begin{cases} 0, & j \text{ even}, \\ \dfrac{\sqrt{2}}{j\pi}\dfrac{1}{n+1}\cot\left(\dfrac{j\pi}{2n}\right), & j \text{ odd}. \end{cases} \tag{9}$$

$$\operatorname{var}(\beta_{nj}) = \begin{cases} \dfrac{1}{j^2\pi^2}\dfrac{1}{n+2}\dfrac{n}{n+1}, & j \text{ even}, \\ \dfrac{2}{j^2\pi^2}\dfrac{1}{n+2}\dfrac{2}{(n+1)^2}\left\{\dfrac{1}{2}n(n+1) - \cot^2\left(\dfrac{j\pi}{2n}\right)\right\}, & j \text{ odd}. \end{cases} \tag{10}$$

The sequence $\{\beta_{nj}\}_{j\in\mathbb{N}}$ has non-zero covariances for small n. However, these covariances decrease quickly with n.

Proposition 5.

$$\operatorname{cov}(\beta_{nj}, \beta_{nk}) = \begin{cases} 0, & \text{if either } j \text{ or } k \text{ are even}, \\ -\dfrac{2}{jk\pi^2}\dfrac{1}{n+2}\dfrac{1}{(n+1)^2}\cot\left(\dfrac{j\pi}{2n}\right)\cot\left(\dfrac{k\pi}{2n}\right), & \text{if both } j \text{ and } k \text{ are odd}. \end{cases} \tag{11}$$

Proof of propositions 4 and 5. To obtain the expected value of Q_n just change $x_{(i)}$ for its expected value $i/(n+1)$ in the expression for Q_n. Its variance can be obtained from the following product of matrices $\operatorname{var}(Q_n) = \mathbf{a}'\mathbf{C}\mathbf{a}$, where

$$\mathbf{a} = \frac{6}{n^2}\begin{pmatrix} 1-n \\ 3-n \\ \vdots \\ n-1 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} n & n-1 & n-2 & \dots & 1 \\ n-1 & 2(n-1) & 2(n-2) & \dots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \dots & n \end{pmatrix} = (C_{rs}).$$

Here

$$C_{rs} = \operatorname{cov}(x_{(r)}, x_{(s)}) = \frac{1}{n+2}\frac{1}{(n+1)^2}\{(n+1)\ \min(r,s) - rs\}.$$

To prove expression (9) just change $x_{(i)}$ for its expected value $i/(n+1)$, in the expression of β_{nj}. To prove expressions (10) and (11) it is necessary to introduce the following notation. Let $\mathbf{C}$ be the same matrix as before and $a_j = -(\sqrt{2}/j\pi)(\mathbf{I}-\mathbf{N})d_j$, where

$$d_j = \begin{pmatrix} \sin(j\pi/n) \\ \sin(2j\pi/n) \\ \vdots \\ \sin(j\pi) \end{pmatrix} \quad \text{and} \quad \mathbf{I}-\mathbf{N} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{pmatrix}.$$

Then

$$\operatorname{var}(\beta_{nj}) = a_j'\mathbf{C}a_j = \frac{2}{(j\pi)^2}d_j'(\mathbf{I}-\mathbf{N})'\mathbf{C}(\mathbf{I}-\mathbf{N})d_j = \frac{2}{(j\pi)^2}d_j'\mathbf{M}d_j, \tag{12}$$

where $\mathbf{M}$ is the product $(\mathbf{I}-\mathbf{N})'\mathbf{C}(\mathbf{I}-\mathbf{N})$. Let $\mathbf{W}$ be the permutations matrix

$$\mathbf{W} = \begin{pmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 1 & 0 \\ \vdots & \iddots & \vdots & \vdots \\ 1 & \dots & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ -1 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

As $\mathbf{WR} = \mathbf{I} - \mathbf{N}$, we have that

$$\mathbf{M} = (\mathbf{WR})'\mathbf{CWR} = \mathbf{R}'\mathbf{CR} = \begin{pmatrix} n & -1 & \dots & -1 \\ -1 & n & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n \end{pmatrix}.$$

To obtain the variance, compute expression (12) using matrix $\mathbf{M}$. Finally, to prove expression (11) use the expression of matrix $\mathbf{M}$ in

$$\operatorname{cov}(\beta_{nj}, \beta_{nk}) = a_j'\mathbf{C}a_k = \frac{2}{jk\pi^2} d_j'\mathbf{M}d_k. \qquad \square$$

The exact PDFs of Q_n and β_{nj}, under the null hypothesis that the distribution of the IID sample $x_1, \dots, x_n$ is uniform on [0, 1], can be computed with the following algorithm, proposed by Dwass (1961), Matsunawa (1985) and Ramallingam (1989).
Let

$$b_i = \sum_{l=i}^{n} a_l = \frac{6}{n^2}(n - i + 1)(i - 1) \qquad \text{for } i = 1, 2, \dots, n,$$

let k be the number of distinct non-zero b_is and $(\nu_1, \dots, \nu_k)$ be the corresponding multiplicities of $(b_1, b_2, \dots, b_k)$. Defining on $\mathbb{C}$ the functions

$$G(s) = \left\{ \prod_{j=1}^{k} \left(s + \frac{1}{b_j} \right)^{\nu_j} \right\}^{-1},$$

$$G_l(s) = \left(s + \frac{1}{b_l} \right)^{\nu_l} G(s), \qquad l = 1, 2, \dots, k,$$

the exact PDF of Q_n, under hypothesis H_0, is given by

$$f_{Q_n}(t) = \sum_{l=1}^{k} \sum_{m=1}^{\nu_l} \operatorname{sgn}(b_l) C_{l,m}^{\#} \chi\left(\frac{t}{b_l}\right) \chi\left(1 - \frac{t}{b_l}\right) t^{m-1} \left(1 - \frac{t}{b_l}\right)^{n-m} \Big/ B(m, n - m + 1)$$

where $\chi(x)$ is the indicator of the interval $[x > 0]$,

$$C_{l,\,m} = \frac{G_l^{(\nu_l - m)}(-1/b_l)}{(\nu_l - m)!},$$

$$C_{l,m}^{\#} = \left\{ \prod_{j=1}^{k} (b_j)^{-\nu_j} \right\} C_{l,m}$$

and $G_l^{(j)}$ denotes the jth derivative of G_l.

In an analogous way to that for Q_n we can find the PDF of β_j, taking into account that

$$b_i = \sum_{l=i}^{n} a_l = \frac{\sqrt{2}}{j\pi} \sin\left\{ \frac{(i-1)j\pi}{n} \right\}, \qquad i = 1, \dots, n.$$

We have written a Mathematica program which implements these algorithms. As an illustration of their application, we have computed critical values, for $\alpha = 0.05$, to test the uniform null hypothesis. A listing for sample sizes $n \leqslant 50$ is reproduced in Tables 1 and 2.

As a general comment on these programs, the non-numerical steps that are needed to yield the exact probability densities are not highly demanding on computational resources. Just the opposite is true for the final step of evaluating the critical values, since it involves solving a

polynomial equation of degree n. Even for moderate sample sizes, it is unavoidable to perform the computations with a large number of exact digits.

4.2. Asymptotic properties

The asymptotic distribution of Q_n and β_{nj}, $j \geqslant 1$, under the null hypothesis follows from the general theory of L-statistics (see, for example, Stigler (1974)).

Let $\{x_{(i)}, 1 \leqslant i \leqslant n\}$ be the order statistics obtained from an IID sample with CDF F. Stigler (1974) computed the asymptotic distribution of

$$S_n = \frac{1}{n} \sum_{i=1}^{n} a_{ni} x_{(i)},$$

with coefficients given by

$$a_{ni} = J\left(\frac{i}{n+1}\right), \qquad i = 1, \ldots, n,$$

where $J : [0, 1] \to \mathbb{R}$ is a continuous, bounded function (almost everywhere with respect to the measure given by F^-). His result is that S_n is asymptotically normal,

$$\sqrt{n}\,\{S_n - E(S_n)\} \xrightarrow[n\to\infty]{\mathcal{L}} N(0, \sigma^2),$$

where

$$\sigma^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} J\{F(x)\} J\{F(y)\} [F\{\min(x, y)\} - F(x)\, F(y)]\, \mathrm{d}x\, \mathrm{d}y.$$

Proposition 6. The asymptotic variances of Q_n and β_{nj}, $j \geqslant 1$, are

$$\sigma_Q^2 = \frac{1}{5},$$

$$\sigma_\beta^2 = \begin{cases} \dfrac{1}{j^2\pi^2}, & \text{if } j \text{ is even,} \\ \dfrac{1}{j^2\pi^2}\left(1 - \dfrac{8}{j^2\pi^2}\right), & \text{if } j \text{ is odd} \end{cases}$$

respectively.

Proof. To prove this proposition just find the function J and then apply Stigler's (1974) result. The statistic Q_n can be expressed as a function of S_n:

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \frac{6}{n}(2i - n - 1) x_{(i)} = 6\frac{n+1}{n} S_n$$

where $J(u) = 2u - 1$. In the same way, the statistics β_{nj}, $j \geqslant 1$, can be written as multiples of a statistic similar to S_n:

$$\begin{aligned} \beta_{nj} &= \sum_{i=1}^{n} \frac{\sqrt{2}}{j\pi} \left\{\sin\left(\frac{i-1}{n} j\pi\right) - \sin\left(\frac{i}{n} j\pi\right)\right\} x_{(i)} \\ &= \frac{-2\sqrt{2}}{j\pi} \sin\left(\frac{j\pi}{2n}\right) \sum_{i=1}^{n} \cos\left(\frac{i - \frac{1}{2}}{n} j\pi\right) x_{(i)}. \end{aligned}$$

Considering the statistic

$$\tilde{S}_n = \frac{1}{n} \sum_{i=1}^{n} \tilde{J}\left(\frac{i - \frac{1}{2}}{n}\right) x_{(i)},$$

$$\beta_{nj} = \frac{-2\sqrt{2}}{j\pi} \sin\left(\frac{j\pi}{2n}\right) \tilde{S}_n$$

where $\tilde{J}(u) = \cos(u\, j\pi)$.

5. Power study

The test based on Q_n is two sided since Q_n concentrates in the lower tail if the variance of the alternative distribution is less than the variance of the uniform distribution. Otherwise, Q_n takes values in the upper tail.

We consider five parametric families of alternative probability distributions with support on $[0, 1]$. We have chosen them so that either the mean or the variance differs from those of the null hypothesis, the uniform distribution, which in each case is obtained for a particular value of the parameter. They are defined by the following CDFs:

(a) Lehmann alternatives,

$$F_\alpha(x) = x^\alpha, \qquad 0 \leqslant x \leqslant 1, \alpha > 0;$$

(b) centred distributions having a U-shaped PDF, for $\beta \in (0, 1)$, or wedge-shaped PDF, for $\beta > 1$,

$$F_\beta(x) = \begin{cases} \frac{1}{2}(2x)^\beta, & 0 \leqslant x \leqslant \frac{1}{2}, \\ 1 - \frac{1}{2}\{2(1-x)\}^\beta, & \frac{1}{2} \leqslant x \leqslant 1; \end{cases}$$

(c) compressed uniform alternatives,

$$F_\gamma(x) = \frac{x - \gamma}{1 - 2\gamma}, \qquad \gamma \leqslant x \leqslant 1 - \gamma,$$

where $0 \leqslant \gamma \leqslant \frac{1}{2}$;

(d) a bimodal locally uniform distribution, with probability mass concentrated near both extremes, 0 and 1,

$$F_\delta(x) = \begin{cases} x/2\delta, & 0 \leqslant x \leqslant \delta, \\ \frac{1}{2}, & \delta \leqslant x \leqslant 1 - \delta, \\ 1 + (x-1)/2\delta, & 1 - \delta \leqslant x \leqslant 1, \end{cases}$$

where $0 \leqslant \delta \leqslant \frac{1}{2}$;

(e) a locally uniform distribution, mixture of family (c) and (d) type of distributions,

$$F_{\varepsilon,\eta}(x) = 2\eta\varepsilon \left\{ \frac{1}{2\varepsilon} \mathbf{1}_{[0,\varepsilon]\cup[1-\varepsilon,1]}(x) \right\} + (1 - 2\eta\varepsilon) \left\{ \frac{1}{1-2\varepsilon} \mathbf{1}_{[\varepsilon,1-\varepsilon]}(x) \right\}.$$

In particular, this family reduces to family (c) for $\eta = 0$, and to family (d) for $\eta = 1/2\varepsilon$.

As an illustration of the test we depict only the power functions at the 5% significance level for a sample size of $n = 10$. We take the critical regions computed and listed in Table 1 from the exact distribution. The points computed in each power curve are estimated by the relative frequency of Q_n in the critical region for $N = 1000$ simulated n-samples of the alternative distribution.

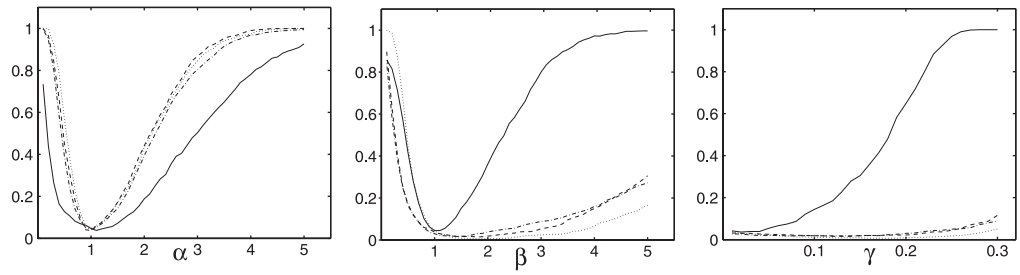

Fig. 1. Power functions for $n = 10$ for families (a)–(c): ———, Q_n; - - - - - - -, D_n; - - - - - -, W_n^2; · · · · · · ·, A_n^2

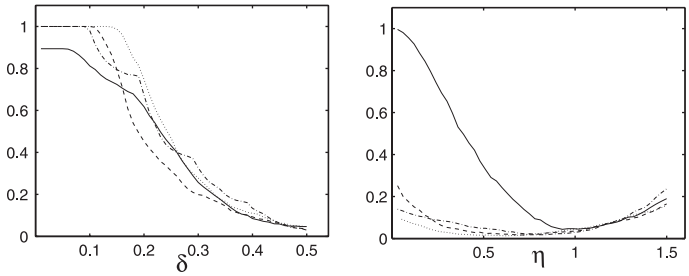

Fig. 2. Power functions for $n = 10$ for families (d) and (e) ($\varepsilon = \frac{1}{3}$): ———, Q_n; - - - - - - -, D_n; - - - - - -, W_n^2; · · · · · · ·, A_n^2

We compare the resulting power functions with the corresponding functions computed for the tests based on the Kolmogorov–Smirnov statistic D_n, Cramér–von Mises statistic W_n^2 and Anderson–Darling statistic A_n^2 in Figs 1 and 2.

6. Discussion

The test based on Q_n is consistent for all the families of alternatives studied. For families (b), (c) and (e) the tests based on D_n, W_n^2 and A_n^2 are biased. Q_n is the most powerful statistic for families (b), (c) and (e), whereas for family (a) it is the least powerful. For family (d) there is no dominant statistic in the set studied.

The Q_n-statistic can be improved for a fixed alternative or family of alternatives by using its expansion in terms of the sequence of L-statistics β_{nj}. In Fortiana and Grané (2000) we used the components β_{nj} to build the statistic

$$T_p = \sum_{j=1}^{p} \lambda_j \beta_{nj},$$

where the coefficients $\lambda_1, \ldots, \lambda_p$ were determined under the constraint that the test based on T_p had maximum power for testing uniformity against a fixed alternative or a fixed family of alternatives. We followed the Neyman–Pearson principle, but restricting the domain to a class of statistics, namely linear combinations of $\beta_{n1}, \ldots, \beta_{np}$, for some (small) positive integer p, whose distributions, either exact or asymptotic, were reasonably easy to compute.

Acknowledgement

This work was supported in part by grants MCT BFM 2000 0801 and 2001SGR00067.

References

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

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

Cambanis, S., Simons, G. and Stout, W. (1976) Inequalities for $Ek(x, y)$ when the marginals are fixed. *Z. Wahrsch. Ver. Geb.*, **36**, 285–294.

Cuadras, C. M. and Fortiana, J. (1993) Continuous metric scaling and prediction. In *Multivariate Analysis, Future Directions*, vol. 2 (eds C. M. Cuadras and C. R. Rao), pp. 47–66. Amsterdam: North-Holland.

Cuadras, C. M. and Fortiana, J. (1994) Ascertaining the underlying distribution of a data set. In *Selected Topics on Stochastic Modelling* (eds R. Gutiérrez and M. J. Valderrama), pp. 223–230. Singapore: World Scientific.

Cuadras, C. M. and Fortiana, J. (1995) A continuous metric scaling solution for a random variable. *J. Multiv. Anal.*, **52**, 1–14.

Durbin, J. and Knott, M. (1972) Components of Cramér–von Mises statistics: I. *J. R. Statist. Soc.* B, **34**, 290–307.

Durbin, J., Knott, M. and Taylor, C. C. (1975) Components of Cramér–von Mises statistics: II. *J. R. Statist. Soc.* B, **37**, 216–237; correction, **39** (1977), 394.

Dwass, M. (1961) The distribution of linear combinations of random divisions of an interval. *Trab. Estadist. Invest. Oper.*, **12**, 11–17.

Fortiana, J. and Grané, A. (2000) *5th Wrld Congr. Bernoulli Society for Probability and Mathematical Statistics and 63rd A. Meet. Institute of Mathematical Statistics.*

Matsunawa, T. (1985) The exact and approximate distributions of linear combinations of selected order statistics from a uniform distribution. *Ann. Inst. Statist. Math.*, **37**, 1–16.

Ramallingam, T. (1989) Symbolic computing the exact distribution of L-statistics from a uniform distribution. *Ann. Inst. Statist. Math.*, **41**, 677–681.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Stephens, M. A. (1974) Components of goodness-of-fit statistics. *Ann. Inst. H. Poincaré* B, **10**, 37–54.

Stephens, M. A. (1986) Tests based on EDF statistics. In *Goodness-of-fit Techniques* (eds R. B. D'Agostino and M. A. Stephens), pp. 97–193. New York: Dekker.

Stigler, S. M. (1974) Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, **2**, 676–693; correction, **7** (1979), 466.