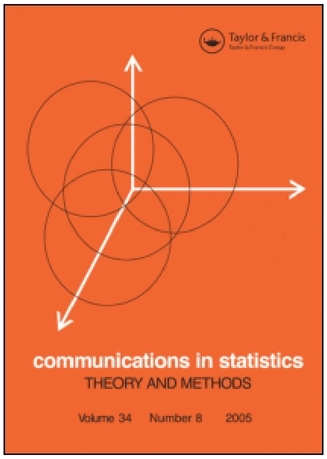

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

An Adaptive Goodness-of-Fit Test

Aurea Grane [a]; Josep Fortiana [b]
[a] Dpto. Estadística, Universidad Carlos III Madrid. Madrid. Spain
[b] Department d'Estadística, Universitat de Barcelona. Barcelona. Spain

To link to this article: DOI: 10.1080/03610920600672286
URL: http://dx.doi.org/10.1080/03610920600672286

Communications in Statistics—Theory and Methods, 35: 1141–1155, 2006
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920600672286

Taylor & Francis
Taylor & Francis Group

An Adaptive Goodness-of-Fit Test

AUREA GRANE[1] AND JOSEP FORTIANA[2]

[1]Dpto. Estadística, Universidad Carlos III Madrid, Madrid, Spain
[2]Department d'Estadística, Universitat de Barcelona, Barcelona, Spain

In Fortiana and Grané (2003), we introduced the statistic Q_n, based on Hoeffding's maximum correlation, as a general-purpose goodness-of-fit test of uniformity. It admits an expansion along a countable set of orthogonal axes, originating a sequence of statistics. Linear combinations of a given number p of terms in this sequence have easy-to-compute probability distributions, either the exact ones for a finite sample or their normal asymptotic approximations for a large sample. In this article we develop an algorithm for tailoring a statistic within this class of linear combinations to test uniformity with optimal power against a specific alternative or family of alternatives.

Keywords Decomposition of test statistics; Goodness-of-fit; Maximum correlation; Tests of uniformity.

Mathematics Subject Classification 62G10; 62G30; 62E15.

1. Introduction

Most goodness-of-fit statistics can be regarded as measures of proximity between two distributions: empirical and hypothesized. The family of tests we are concerned with in this article is based on Hoeffding's maximum correlation between cumulative distribution functions (cdf's) F_1 and F_2, henceforth denoted by $\rho^+(F_1, F_2)$, defined as the maximum of the correlation coefficients of bivariate distributions having F_1 and F_2 as marginals. Since $\rho^+(F_1, F_2)$ equals 1 if and only if $F_1 = F_2$ (almost everywhere) up to a scale and location change, it is adequate to our purpose. In Fortiana and Grané (2003) we defined and studied the properties of

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

where F_n is the empirical cdf for n iid real-valued random variables, s_n is the empirical standard deviation, and F_U is the probability distribution function of a

Address correspondence to Aurea Grane, Dpto. Estadística, Universidad Carlos III Madrid, C/Madrid 126, 28903 Getafe, Madrid, Spain; E-mail: agrane@est-econ.uc3m.es

$[0, 1]$ uniform random variable, as a test statistic for $H_0 : F = F_U$ (also for $H_0 : F = F_0$, where F_0 is a fully specified continuous probability distribution function, case 0 in Stephens, 1986). We found that Q_n has good properties as a goodness-of-fit test: since it is an L-statistic, it is possible to find its exact distribution under the null hypothesis for small samples and, additionally, it is asymptotically normally distributed. The test based on Q_n can advantageously replace those of Kolmogorov–Smirnov, Cramér-von Mises, and Anderson–Darling for a wide range of alternatives. We also proved the following identity:

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

where the sequence $\{\beta_{nj}\}_{j\geq 0}$ appears as a decomposition of this test, analogous to those studied by Anderson and Darling (1952, 1954), Durbin and Knott (1972, 1975), and Stephens (1974). Also, β_{nj} is the jth Fourier coeffcient of the pseudoinverse of the empirical distribution function F_n^- for an orthonormal sequence, $\{\beta_j(t)\}_{j\geq 0}$, in $L^2[0, 1]$ (see Cuadras and Fortiana, 1993; Fortiana and Cuadras, 1997; Fortiana and Grané, 2003 for details). In the present article, we seek to improve the performance of Q_n for an alternative or family of alternatives by choosing a linear combination of a finite number p of terms in the sequence $\{\beta_{nj}\}_{j\geq 0}$ that maximizes power against the given alternative. Even if the best Neyman–Pearson power is not attained, the resulting test keeps the computational advantages of Q_n.

In Sec. 2 we define and list properties of the class of statistics we are interested in. In Sec. 3 we express the problem of power optimization in terms of certain quadratic forms depending on moments of the order statistic. As an example of family of alternatives we use the $[0, \theta]$-uniform distributions. Section 4 contains a method for finding the statistic in a generic case, including an algorithm to perform the actual computation, with some numerical illustrations. Finally, Sec. 5 includes power computations, as well as a general discussion of results.

2. Definition of the Statistic

Let F be a probability distribution function with finite second order moment and let F_n be the empirical cdf of n i.i.d. random variables with distribution F, $x_1, \ldots, x_n$. With the orthonormal (in $L^2[0, 1]$) system $\{\beta_0 = 1; \beta_j(t) = \sqrt{2}\cos(j\pi t), j \geq 1\}$, we define the statistics:

$$\beta_{nj} \equiv \beta_{nj}(F) = \int_0^1 F_n^-(t)\beta_j(t)dt, \quad j \geq 0, \tag{1}$$

where F_n^- is the pseudoinverse of F_n, defined as $F_n^-(t) = \inf\{x \in \mathbb{R} : F_n(x) \geq t\}$. This sequence was thoroughly studied in Fortiana and Grané (2003). Explicitly, in terms of the order statistic $\mathbf{x} = (x_{(1)}, \ldots, x_{(n)})'$,

$$\begin{aligned} \beta_{n0} &= \bar{x}_n, \\ \beta_{nj} &= \frac{\sqrt{2}}{j\pi} \sum_{i=1}^{n} \left(\sin\frac{ij\pi}{n} - \sin\frac{(i-1)j\pi}{n} \right) x_{(i)}, \quad j \geq 1, \end{aligned} \tag{2}$$

and their expectations under H_0 are given by

$$E(\beta_{n0}) = \frac{1}{2},$$
$$E(\beta_{nj}) = \begin{cases} 0, & \text{if } j \text{ is even,} \\ -\dfrac{\sqrt{2}}{j\pi}\dfrac{1}{n+1}\cot\left(\dfrac{j\pi}{2n}\right), & \text{if } j \text{ is odd,} \end{cases}, \quad j \geq 1. \tag{3}$$

In Fortiana and Grané (2003, Eq. (3), p. 118), the β_{nj} statistics were defined in a slightly different way. Here we have included β_{n0} and changed the sign of β_{nj}, for $j \geq 1$. We will consider all statistics of the form

$$T = T(\tilde{\lambda}) = \sum_{j \geq 0} \tilde{\lambda}\beta_{nj}, \tag{4}$$

where $\{\tilde{\lambda}_j\} \in \ell^1_{\mathbb{R}}$ is a sequence of real numbers. T is an L-statistic,

$$T = \mathbf{w}'\mathbf{x} = \sum_{i=1}^{n} w_i x_{(i)},$$

with coefficients

$$w_i = \frac{\tilde{\lambda}_0}{n} + \frac{\sqrt{2}}{\pi}\sum_{j\geq 1}\frac{\tilde{\lambda}_j}{j}\left(\sin\frac{ij\pi}{n} - \sin\frac{(i-1)j\pi}{n}\right).$$

The ℓ^1 condition imposed on $\{\tilde{\lambda}_j\}$ ensures that each w_i, $0 \leq i \leq n$, is the sum of an absolutely convergent sequence. Moreover, due to the periodicity $j\beta_{nj} = (j+2n)\beta_{n,j+2n}$, $j \geq 1$, definition (4) is equivalent to

$$T = T(\lambda) = \sum_{j=0}^{2n} \lambda_j \beta_{nj}, \tag{5}$$

where

$$\lambda_0 = \tilde{\lambda}_0, \quad \lambda_j \sum_{k=0}^{\infty} \tilde{\lambda}_{2nk+j}\frac{j}{j+2nk}, \quad \text{for } 1 \leq j \leq 2n.$$

Note that this is an absolutely convergent sum. We will use $T_p \equiv T_p(L)$, the result of truncating (5) at $j = p$, where $L = (\lambda_0, \ldots, \lambda_p)'$. Since in practical situations p will be much smaller than n, henceforth we assume $p < n$.

In matrix notation

$$T_p(L) = \mathbf{w}'\mathbf{x} = L'\boldsymbol{\Gamma}\mathbf{S}_p'(\mathbf{I} - \mathbf{N})'\mathbf{x}, \tag{6}$$

where $L = (\lambda_1, \dots, \lambda_p)'$, $\mathbf{\Gamma} = diag(1/n, \sqrt{2}/\pi, \sqrt{2}/(2\pi), \dots, \sqrt{2}/(p\pi))$,

$$\mathbf{I} - \mathbf{N} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ & \ddots & \ddots & \\ 0 & \cdots & -1 & 1 \end{pmatrix}, \quad \mathbf{S}_p = \begin{pmatrix} 1 & s_{11} & s_{12} & \cdots & s_{1p} \\ 2 & s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & s_{n1} & s_{n2} & \cdots & s_{np} \end{pmatrix},$$

where $s_{ij} = \sin(ij\pi/n)$, $1 \le i \le n$, $1 \le j \le p$.

Given an alternative cdf F_1, we select L in order to maximize power for testing $H_0 : F = F_U$, vs. $H_1 : F = F_1$. Clearly, the resulting test is less powerful than the optimal (Neyman–Pearson), but on the other hand, its distribution under the null hypothesis is easily computed, both for large samples, applying the asymptotic theory of L-statistics, and for small samples, with the exact distribution, as described in Fortiana and Grané (2003).

From formula (2.9) in Ramallingam (1989), it is also possible to get explicit expressions for the moments under H_0:

$$\mu_r = E(T^r) = \frac{r!n!}{(n+r)!} \prod_{i=2}^{n} \frac{1}{b_i} \sum_{i=2}^{n} b_i^{r+1},$$

in particular,

$$\mu_1 = \frac{1}{n+1} \frac{\sum_{i=2}^{n} b_i^2}{\prod_{i=2}^{n} b_i},$$

where

$$b_i = \sum_{k=i}^{n} w_k = \frac{n-i+1}{n}\lambda_0 - \frac{\sqrt{2}}{\pi} \sum_{j=1}^{p} \frac{\lambda_j}{j} \sin\frac{(i-1)j\pi}{n}.$$

3. Computation and Optimization of the Power Function

It is possible to use a one-sided test for a fixed alternative F_1, since it may be proved that when H_1 is true T_p tends to its upper tail if $var(F_1) > 1/12$ and to its lower tail otherwise. In general, however, for a family of alternatives we must consider the two-sided test. Henceforth this will be our assumption.

Let F_U be the $[0, 1]$-uniform distribution and F_1, also with support on $[0, 1]$, belonging to a family of alternatives. From the general theory of L-statistics (see, e.g., Stigler, 1974, or Ch. 19 of Shorack and Wellner, 1986) the asymptotic distribution of $T_p(L)$ is normal. For a given significance level $\varepsilon \in (0, 1)$, we are looking for critical values $c_1, c_2 \in \mathbb{R}$, such that

$$P(T_p(L) > c_1 \mid H_0) = \varepsilon/2, \quad P(T_p(L) < c_2 \mid H_0) = \varepsilon/2.$$

Since the distribution of $T_p(L)$ under H_0 is centered at $\mu_0 = E(T_p(L) \mid H_0)$, we take c_1, c_2 symmetric with respect to μ_0, that is, $c_1 = \mu_0 + c_{\varepsilon/2}\sigma_0$, $c_2 = \mu_0 - c_{\varepsilon/2}\sigma_0$, where

$\sigma_0^2 = var(T_p(L)\,|\,H_0)$ and $c_{\varepsilon/2}$ is the $(1-\varepsilon/2)$-quantile of the $N(0,1)$ distribution. The power,

$$P(T_p(L) > c_1\,|\,H_1) + P(T_p(L) < c_2\,|\,H_1),$$

is asymptotically approximated by

$$\Psi(L) = 1 - P_Z\left[\left(\frac{\mu_0-\mu_1}{\sigma_1} - c_{\varepsilon/2}\frac{\sigma_0}{\sigma_1}, \frac{\mu_0-\mu_1}{\sigma_1} + c_{\varepsilon/2}\frac{\sigma_0}{\sigma_1}\right)\right],$$

where $\mu_1 = E(T_p(L)\,|\,H_1)$, $\sigma_1^2 = var(T_p(L)\,|\,H_1)$, and $Z \sim N(0,1)$. Due to the symmetry of the normal distribution, $\mu_0 - \mu_1$ can be replaced with $|\mu_0 - \mu_1|$, hence

$$\Psi(L) = 1 - P_Z\left\{\left(\left[\frac{a(L)}{c(L)}\right]^{1/2} - \left[\frac{b(L)}{c(L)}\right]^{1/2}, \left[\frac{a(L)}{c(L)}\right]^{1/2} + \left[\frac{b(L)}{c(L)}\right]^{1/2}\right)\right\},$$

in terms of the following quadratic forms:

$$\begin{aligned} a(L) &= (\mu_0-\mu_1)^2 = L'\mathbf{A}L, \quad \text{where } \mathbf{A} = \mathbf{D}'(\mathbf{M}_0-\mathbf{M}_1)(\mathbf{M}_0-\mathbf{M}_1)'\mathbf{D},\\ b(L) &= c_{\varepsilon/2}^2\sigma_0^2 = L'\mathbf{B}L, \quad \text{where } \mathbf{B} = c_{\varepsilon/2}^2\mathbf{D}'\Sigma_0\mathbf{D},\\ c(L) &= \sigma_1^2 = L'\mathbf{C}L, \quad \text{where } \mathbf{C} = \mathbf{D}'\Sigma_1\mathbf{D}, \end{aligned} \tag{7}$$

where $\mathbf{M}_i = E(\mathbf{x}\,|\,H_i)$, $\Sigma_i = Var(\mathbf{x}\,|\,H_i)$, $i = 0, 1$, and $\mathbf{D} = (\mathbf{I}-\mathbf{N})\mathbf{S}_p\boldsymbol{\Gamma}$. The expectation and variance of the order statistic under the null hypothesis are given by formula (10) below. See Sec. 4 to compute them under the alternative hypothesis.

As stated in the previous section, we want to find L such that the power is maximized. Since $\Psi(L)$ is invariant when L is multiplied by an arbitrary constant, we assume $c(L) = 1$, thus we have to compute the extremes of

$$\Upsilon(L) = 1 - \Phi\big(a(L)^{1/2} + b(L)^{1/2}\big) + \Phi\big(a(L)^{1/2} - b(L)^{1/2}\big) + \lambda(c(L)-1), \tag{8}$$

where Φ is the standard normal probability distribution function and λ is a Lagrange multiplier.

Degenerate case. If $a(L) = 0$, the expectation of $T_p(L)$ is the same under both hypotheses, then the asymptotically approximated power function is

$$\Psi(L) - 1 - P_Z\left\{\left(-\left[\frac{b(L)}{c(L)}\right]^{1/2}, \left[\frac{b(L)}{c(L)}\right]^{1/2}\right)\right\},$$

with the restriction $c(L) = 1$, and (8) can be written as

$$\Upsilon(L) = 2 - 2\Phi\big(b(L)^{1/2}\big) + \lambda(c(L)-1).$$

Differentiating with respect to L and equating to zero, we obtain an eigenvalue-type problem:

$$\beta(L)\mathbf{B}L = \lambda\mathbf{C}L,$$

where $\beta(L) = b(L)^{-1/2}\phi\big(b(L)^{1/2}\big)$, and ϕ is the standard normal probability density function.

General case. If $a(L) \neq 0$, then differentiating (8) with respect to L and equating to zero, we obtain the following eigenvalue-type problem:

$$[\alpha(L)\mathbf{A} + \beta(L)\mathbf{B}]L = \lambda\mathbf{C}L, \tag{9}$$

where

$$\begin{aligned}
\alpha(L) &= a(L)^{-1/2}(\phi_+(L) - \phi_-(L)),\\
\beta(L) &= b(L)^{-1/2}(\phi_+(L) + \phi_-(L)),\\
\phi_+(L) &= \phi\big(a(L)^{1/2} + b(L)^{1/2}\big),\\
\phi_-(L) &= \phi\big(a(L)^{1/2} - b(L)^{1/2}\big).
\end{aligned}$$

The degenerate case appears for $\alpha(L) = 0$.

In order to compute L we change to the new variable $u = \mathbf{C}^{1/2}L$, and let $\mathbf{E} = \mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}$, $\mathbf{F} = \mathbf{C}^{-1/2}\mathbf{B}\mathbf{C}^{-1/2}$, and $\mathbf{G}(u) = \alpha(u)\mathbf{E} + \beta(u)\mathbf{F}$, where $\alpha(u)$, $\beta(u)$ denote the quantities $\alpha(L)$, $\beta(L)$ as defined in (9) in terms of the new variable u. Given an initial u, we compute the set of eigenvectors/eigenvalues of $\mathbf{G}(u)$. The new u will be the eigenvector such that $\Psi(u)$ is maximum. This process is iterated to stationarity. The last step is to recover and normalize L. The result is rather robust, leading to a single maximum with a small number of iterations for a widely diverse choice of the initial u. A Matlab program implementing this computation may be requested from the authors.

3.1. *Example: Scale Alternatives*

We consider an alternative distribution belonging to the family $U[0, \theta]$, uniform on $[0, \theta]$, with $\theta > 0$. The vector of expectations $\mathbf{M}_0$ and the matrix of covariances Σ_0 of the order statistic $x_{(1)}, \ldots, x_{(n)}$ obtained from n random variables iid $\sim U[0, 1]$ are

$$\mathbf{M}_0 = \frac{1}{n+1}(1, 2, \ldots, n)', \quad \Sigma_0 = (v_{ij})_{1\le i,j\le n}, \tag{10}$$

where

$$v_{ij} = \frac{1}{(n+2)(n+1)^2}[(n+1)\min\{i, j\} - ij],$$

(see, e.g., David, 1981) and those of the order statistic obtained from n iid $\sim U[0, \theta]$ random variables are $\mathbf{M}_1 = \theta\mathbf{M}_0$, $\Sigma_1 = \theta^2\Sigma_0$. Hence (7) becomes

$$\begin{aligned}
a(L) &= (1-\theta)^2 L'\mathbf{A}L, \quad \text{where } \mathbf{A} = \mathbf{D}'\mathbf{M}_0\mathbf{M}_0'\mathbf{D},\\
b(L) &= c_{\varepsilon/2}^2 L'\mathbf{B}L, \quad \text{where } \mathbf{B} = \mathbf{D}'\Sigma_0\mathbf{D},\\
c(L) &= \theta^2 L'\mathbf{B}L.
\end{aligned}$$

We have to maximize

$$\Psi(L) = 1 - P_Z\left(\frac{1-\theta}{\theta}\left(\frac{L'\mathbf{A}L}{L'\mathbf{B}L}\right)^{1/2} - \frac{c_{\varepsilon/2}}{\theta}, \frac{1-\theta}{\theta}\left(\frac{L'\mathbf{A}L}{L'\mathbf{B}L}\right)^{1/2} + \frac{c_{\varepsilon/2}}{\theta}\right),$$

under the restriction $L'\mathbf{B}L = 1$, which is equivalent to maximizing the quotient $L'\mathbf{A}L/L'\mathbf{B}L$, and also to finding the eigenvector of maximum eigenvalue in $\mathbf{A}L = \lambda\mathbf{B}L$ with the restriction $L'\mathbf{B}L = 1$, i.e.,

$$\mathbf{D}'\mathbf{M}_0\mathbf{M}_0'\mathbf{D}L = \lambda\mathbf{D}'\Sigma_0\mathbf{D}L$$

with the restriction $L'\mathbf{D}'\Sigma_0\mathbf{D}L = 1$. Note that the left-hand side matrix has unit rank, hence there is only one eigenvector with a non null eigenvalue. Additionally, this solution does not depend on the parameter θ.

As a numerical example, for a sample size $n = 20$, a significance level $\varepsilon = 0.05$ and $p = 4$,

$$T_p = 0.3554\beta_{n0} - 0.4447\beta_{n1} + 0.4985\beta_{n2} - 0.4373\beta_{n3} + 0.4860\beta_{n4}.$$

In a practical situation, T_p should be expressed directly in terms of the observed order statistic using (6). The critical values of the test based on T_p are

$$c_1 = 0.33974, \quad c_2 = 0.26608.$$

We have compared T_p with the statistic Q_n obtained in Fortiana and Grané (2003), with the Kolmogorov–Smirnov statistic D_n and the Cramér-von Mises statistic W_n^2. Figure 1 shows the power curves for the tests based on these statistics. These curves are plotted from 20 computed points, for each of which we have generated $N = 1000$

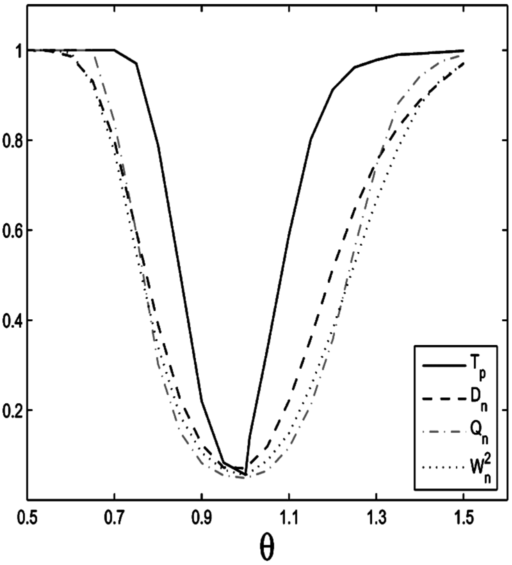

Figure 1. Power functions for scale alternatives.

samples of size $n = 20$, and we have estimated the power as the relative frequency of values in the critical region. We allowed θ to take values below and above 1, thus obtaining a two-sided power curve.

4. Generic Alternatives

If for a given alternative distribution we know how to compute $\mathbf{M}_1$ and Σ_1 in Eq. (7), we can proceed as described in Sec. 3. Here we develop an algorithm to compute the T_p statistic for an alternative cdf F such that its pseudoinverse F^- admits the expression:

$$F^-(t) = \gamma_0 + \sqrt{2}\sum_{k=1}^{q}\gamma_k\cos(k\pi t), \tag{11}$$

where γ_k are real numbers. An arbitrary cdf can be approximated by taking the first q terms of the corresponding Fourier series. The pseudoinverse F^- appears in the formulas of the moments of L-statistics, (18) and (19) see below, hence it seems natural to expand F^- instead of F itself or its density. $\mathbf{M}_0$ and Σ_0 in (7) are the same as in (10). The entries in $\mathbf{M}_1$ are given by

$$\begin{aligned} E(x_{(i)}\,|\,H_1) &= i\binom{n}{i}\int_0^1 F^-(t)t^{i-1}(1-t)^{n-i}dt \\ &= i\binom{n}{i}\gamma_0\mathrm{Beta}(i, n-i+1) + i\binom{n}{i}\sqrt{2}\sum_{k=1}^{q}\left(\gamma_k\int_0^1\cos(k\pi t)t^{i-1}(1-t)^{n-i}\,dt\right) \\ &= \gamma_0 + \sqrt{2}\sum_{k=1}^{q}\gamma_k F_{2,3}\left(\left\{\frac{i+1}{2}, \frac{i}{2}\right\}; \left\{\frac{1}{2}, \frac{n+1}{2}, \frac{n+2}{2}\right\}; \frac{-k^2\pi^2}{4}\right), \quad 1 \le i \le n, \end{aligned}$$

where

$$F_{p,q}(\mathbf{a}; \mathbf{b}; z) = \sum_{k=0}^{\infty}\frac{(a_1)_k\cdots(a_p)_k}{(b_1)_k\cdots(b_q)_k}\frac{z^k}{k!}$$

is the generalized hypergeometric function with parameters $a = \{a_1, \ldots, a_p\}$, $\mathbf{b} = \{b_1, \ldots, b_q\}$, for $p \ge 0$, $q \ge 1$, and $(r)_k$ denotes the Pochhammer symbol, that is

$$(r)_k = r(r+1)\cdots(r+k-1) = \Gamma(r+k)/\Gamma(r),$$

see Wolfram (1996, Sec. 3.2.10). In general, an exact formula to compute Σ_1 will not be available, instead we can determine matrix $\mathbf{C}$ from the following asymptotic approximation.

Proposition 4.1. *Let $T_p = T_p(L)$ be the statistic defined in* (6), *in which the order statistic has been obtained from n iid random variables with distribution* (11). *We have the following convergences in law*

$$\sqrt{n}[T_p - \mu_n] \xrightarrow[n\to\infty]{\mathscr{L}} N(0, \sigma_1^2), \tag{12}$$

$$\sqrt{n}\frac{[T_p - \mu_n]}{\sigma_n} \xrightarrow[n\to\infty]{\mathscr{L}} N(0, 1), \tag{13}$$

where

$$\mu_n = \lambda_0\gamma_0 + \sum_{j=1}^{\min(p,q)} \lambda_j\gamma_j \frac{2n}{j\pi}\sin\left(\frac{j\pi}{2n}\right) \tag{14}$$

$$\sigma_1^2 = \lim_{n\to\infty}\sigma_n^2, \quad \sigma_n^2 = \sum_{j=0}^{p}\sum_{l=0}^{p}\lambda_j\lambda_l\sigma_{n,jl}, \quad \sigma_{n,jl} = 2\pi^2\xi_{nj}\xi_{nl}\sum_{k=1}^{q}\sum_{m=1}^{q} km\gamma_k\gamma_m I_{jklm}, \tag{15}$$

where $\xi_{n0} = 1$, *and* $\xi_{nj} = \sqrt{2}(2n/(j\pi))\sin(j\pi/(2n))$, *for* $1 \le j \le p$,

$$I_{jklm} = \frac{1}{8\pi^2}\left\{\frac{1}{(k+j)^2}[\delta_{m-l,k+j} + \delta_{m+l,k+j}]\right\}, \quad \textit{for } k = j,$$

$$I_{jklm} = \frac{1}{8\pi^2}\left\{\frac{1}{(k-j)^2}[\delta_{m-l,k-j} + \delta_{m+l,k-j}] + \frac{1}{(k+j)^2}[\delta_{m-l,k+j} + \delta_{m+l,k+j}]\right\},$$

for $k \neq j$, *and* δ *is Kronecker's delta.*

Proof. From (2) and (5),

$$T_p = \frac{1}{n}\sum_{i=1}^{n}\left[\lambda_0 + \sum_{j=1}^{p}\lambda_j\frac{2n}{j\pi}\sin\left(\frac{j\pi}{2n}\right)\sqrt{2}\cos\frac{(2i-1)j\pi}{2n}\right]x_{(i)}$$

and, defining,

$$J_{n\lambda}(t) = \sum_{j=0}^{p}\lambda_j\xi_{nj}\cos(j\pi t), \tag{16}$$

where

$$\xi_{n0} = 1, \quad \xi_{nj} = \sqrt{2}\frac{2n}{j\pi}\sin\left(\frac{j\pi}{2n}\right), \ \ 1 \le j \le p, \tag{17}$$

the statistic, T_p can be written as

$$T_p = \frac{1}{n}\sum_{i=1}^{n} J_{n\lambda}\left(\frac{1-1/2}{n}\right)x_{(i)},$$

where $J_{n\lambda}$ is a continuous and bounded a.e. (F^-) function. The natural centering constant μ_n, (see p. 661 of Shorack and Wellner, 1986), defined as

$$\mu_n = \int_0^1 J_{n\lambda}(t)F^-(t)dt, \tag{18}$$

is equal to

$$\lambda_0\gamma_0 + \lambda_0\sqrt{2}\sum_{k=1}^{q}\gamma_k\int_0^1\cos(k\pi t)dt + \gamma_0\sum_{j=1}^{p}\lambda_j\xi_{nj}\int_0^1\cos(j\pi t)dt$$
$$+\sqrt{2}\sum_{j=1}^{p}\lambda_j\xi_{nj}\sum_{k=1}^{q}\gamma_k\int_0^1\cos(j\pi t)\cos(k\pi t)dt,$$

which is (14). Similarly,

$$\sigma_n^2 = \int_0^1 \int_0^1 J_{n\lambda}(s) J_{n\lambda}(t)[\min(s,t) - st] dF^-(s) dF^-(t)$$
$$= 2\pi^2 \sum_{j=0}^{p} \sum_{k=1}^{q} \sum_{l=0}^{p} \sum_{m=1}^{q} \lambda_j \lambda_l \xi_{nj} \xi_{nl} km \gamma_k \gamma_m I_{jklm}, \tag{19}$$

where

$$I_{jklm} = \int_0^1 \int_0^1 K(s,t) \cos(j\pi s) \sin(k\pi s) \cos(l\pi t) \sin(m\pi t) dt\, ds,$$

where $K(s,t) = \min(s,t) - st$, $0 \le s, t \le 1$. Defining

$$\sigma_{n,jl} = 2\pi^2 \xi_{nj} \xi_{nl} \sum_{k=1}^{q} \sum_{m=1}^{q} km \gamma_k \gamma_m I_{jklm},$$

we have that

$$\sigma_n^2 = \sum_{j=0}^{p} \sum_{l=0}^{p} \lambda_j \lambda_l \sigma_{n,jl}.$$

Observing that the eigenvalues and orthonormalized eigenfunctions of the integral operator with kernel $K(s,t)$ are, respectively,

$$\mu_j = \frac{1}{(j\pi)^2}, \quad f_j(t) = \sqrt{2}\sin(j\pi t), \quad 0 \le t \le 1, \ \ j \in \mathbb{N},$$

(see pp. 213–214 of Shorack and Wellner, 1986), we obtain

$$I_{jklm} = \int_0^1 \int_0^1 K(s,t) \frac{1}{2}[\sin((k-j)\pi s) + \sin((k+j)\pi s)]$$
$$\times \frac{1}{2}[\sin((m-l)\pi s) + \sin((m+l)\pi s)] dt\, ds$$
$$= \frac{1}{4} \int_0^1 \int_0^1 K(s,t) \frac{1}{\sqrt{2}}[f_{k-j}(s) + f_{k+j}(s)] \frac{1}{\sqrt{2}}[f_{m-l}(t) + f_{m+l}(t)] dt\, ds$$
$$= \frac{1}{8} \int_0^1 [f_{m-l}(t) + f_{m+l}(t)] \int_0^1 K(s,t)[f_{k-j}(s) + f_{k+j}(s)] ds\, dt.$$

For $k \neq j$,

$$I_{jklm} = \frac{1}{8} \int_0^1 [f_{m-l}(t) + f_{m+l}(t)] \left[\frac{1}{(k-j)^2\pi^2} f_{k-j}(t) + \frac{1}{(k+j)^2\pi^2} f_{k+j}(t) \right] dt$$
$$= \frac{1}{8\pi^2} \left\{ \frac{1}{(k-j)^2} [\delta_{m-l,k-j} + \delta_{m+l,k-j}] + \frac{1}{(k+j)^2} [\delta_{m-l,k+j} + \delta_{m+l,k+j}] \right\},$$

and for $k = j$,

$$I_{jklm} = \frac{1}{8\pi^2}\left\{\frac{1}{(k+j)^2}[\delta_{m-l,k+j} + \delta_{m+l,k+j}]\right\},$$

where δ is Kronecker's delta. To compute $\sigma_1^2 = \lim_{n\to\infty} \sigma_n^2$, substitute

$$J_\lambda(t) = \lim_{n\to\infty} J_{n\lambda}(t) = \lambda_0 + \sqrt{2}\sum_{j=1}^{p} \lambda_j \cos(j\pi t),$$

for $J_{n\lambda}$ inside the integral in (19), which can be done because $J_{n\lambda}$ is a continuous and bounded a.e. (F^-) function. Then,

$$\sigma_1^2 = \sum_{j=0}^{p}\sum_{l=0}^{p} \lambda_j \lambda_l \sigma_{jl},$$

where

$$\sigma_{jl} = \lim_{n\to\infty} \sigma_{n,jl} = 4\pi^2 \sum_{k=1}^{q}\sum_{m=1}^{q} km\gamma_k\gamma_m I_{jklm}.$$

The convergence of (12) is obtained applying Theorem 1, pp. 664–665 of Shorack and Wellner (1986). The convergence of (13) is immediate from (12) and the fact that $\sigma_1^2 = \lim_{n\to\infty} \sigma_n^2$. □

Comparing the expression for $c(L) = \sigma_1^2 = L'\mathbf{C}L$ in (7) with (15), we see that the entries in **C** are either $\sigma_{n,jl}$ or the limit $\sigma_{jl} = \lim_{n\to\infty} \sigma_{n,jl}$. Some computational examples suggest that a better approximation is obtained with $\sigma_{n,jl}$.

4.1. *Some Examples*

To illustrate the method we have choosen four parametric families of alternative distributions with support on [0, 1]. We have choosen 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 probability distribution function.

A1. Lehmann alternatives,

$$F_\alpha(x) = x^\alpha, \quad 0 \le x \le 1, \ \alpha > 0.$$

A2. Centered distributions having U-shaped probability density function, for $\beta \in (0, 1)$, or wedge-shaped probability density function, for $\beta > 1$,

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

A3. Compressed uniform alternatives,

$$F_\gamma(x) = \frac{x-\gamma}{1-2\gamma}, \quad \gamma \le x \le 1-\gamma, \ 0 \le \gamma \le \frac{1}{2}.$$

A4. A bimodal locally uniform distribution, with probability mass concentrated near both extremes, 0 and 1,

$$F_\delta(x) = \begin{cases} x/(2\delta), & 0 \le x \le \delta, \\ \dfrac{1}{2}, & \delta \le x \le 1-\delta, \quad 0 < \delta \le 1/2, \\ 1-(x-1)/(2\delta), & 1-\delta \le x \le 1. \end{cases}$$

As examples of construction of the test for generic alternatives, we have considered the families above for some values of the parameters. For each alternative we determine coefficients γ_k of (11), for $0 \le k \le q$. Applying the algorithm for a sample size of $n = 20$ and a significance level $\varepsilon = 0.05$ we determine

Table 1
Computations for statistic T_p for families A1, A2, and A3

Family	Fourier coeff.	Weights	Critical values
A1		−0.801406	
$\alpha = 1/2$	$\gamma_0 = \frac{1}{3}$	−0.426894	$c_1 = -0.159319$
	$\gamma_k = (-1)^k \frac{2\sqrt{2}}{(k\pi)^2}$	0.353300	
	$1 \le k \le q$	−0.036017	$c_2 = -0.407395$
		0.222244	
A2	$\gamma_0 = 1/2$		
$\beta = 2$	$\gamma_1 = -0.197286$	0	
	$\gamma_2 = 0$	−0.971767	$c_1 = 0.327609$
	$\gamma_3 = -0.0448157$	0	
	$\gamma_4 = 0$	−0.235944	$c_2 = 0.215799$
	$\gamma_5 = -0.0197851$	0	
A3	$\gamma_0 = 1/2$	0	
$\gamma = 0.15$	$\gamma_k = 0,$	0.837951	$c_1 = -0.200292$
	$1 \le k \le q, k$ even,	0	
	$\gamma_k = -\frac{2\sqrt{2}}{(k\pi)^2}(1-2\gamma),$	0.545746	$c_2 = -0.288662$
	$1 \le k \le q, k$ odd.	0	
A4	$\gamma_0 = 1/2$		
$\delta = 0.05$	$\gamma_k = 0,$	0	
	$1 \le k \le q, k$ even,	0.998779	$c_1 = -0.207086$
	$\gamma_k = -\frac{4\delta\sqrt{2}}{(k\pi)^2} + \frac{(2\delta-1)\sqrt{2}}{k\pi}\sin(k\pi/2),$	−0.049396	$c_2 = -0.334052$
	$1 \le k \le q, k$ odd.	0	

Table 2
Power of the test based on T_p, Q_n, D_n, W_n^2, and the UMP test for the A1 family

α	T_p	Q_n	D_n	W_n^2	UMP
0.25	0.9980	0.4411	0.9970	0.9973	1.0000
0.5	0.7492	0.1203	0.6550	0.7211	0.9259
0.75	0.1973*	0.0764	0.1830	0.1987	0.3918
2	0.8347	0.3984	0.6730	0.7708	0.9185
3	0.9872	0.8779	0.9910	0.9955	0.9998
4	0.9991	0.9900	1.0000	1.0000	1.0000

Table 3
Power of the tests based on T_p, Q_n, D_n, and W_n^2 for the A2 family

β	T_p	Q_n	D_n	W_n^2
0.25	0.9447	0.9651	0.8071	0.8597
0.5	0.6524	0.7238	0.2879	0.2840
0.75	0.1893*	0.2203	0.0916	0.0905
2	0.8045	0.7523	0.1288	0.1013
3	0.9929	0.9955	0.4029	0.5107
4	1.0000	1.0000	0.7361	0.8951

Table 4
Power of the test based on T_p, Q_n, D_n, and W_n^2 for the A3 family

γ	T_p	Q_n	D_n	W_n^2
0.05	0.1761*	0.1016	0.0453	0.0387
0.10	0.6049	0.3609	0.0451	0.0426
0.15	0.9894	0.8244	0.0677	0.0669
0.25	1.0000	1.0000	0.3775	0.6195
0.35	1.0000	1.0000	1.0000	1.0000

Table 5
Power of the test based on T_p, Q_n, D_n, and W_n^2 for the A4 family

δ	T_p	Q_n	D_n	W_n^2
0.05	0.9619	0.9585	1.0000	1.0000
0.15	0.8905	0.9309	1.0000	1.0000
0.25	0.7951	0.7736	0.8817	0.7533
0.35	0.3934*	0.3097	0.3321	0.1964
0.45	0.0745*	0.0697	0.0931	0.0752

the vector of coefficients L, which are the components weights, for the test based on T_p, for $p = 4$ and $q = 5$. We also compute the critical regions for each case. The results are shown in Table 1.

5. Discussion and Concluding Remarks

Tables 2–5 contain the power comparisons of the test based on T_p with the tests based on Q_n, D_n, and W_n^2. These powers have been estimated from $N = 10,000$ samples of size $n = 20$ as the relative frequency of values of the statistic in the critical region. Since the UMP test is easy to compute for the A1 family, we have included these results in Table 2 for comparison.

For some values of the parameter in the A2 and A4 families, the power of the test based on Q_n exceeds that of T_p. According to theory, this should not happen, Q_n belongs to the class of statistics from which T_p has been extracted, via an optimization. After checking several possible causes of inaccuracy, we can discard both a too small p and a too small q, since doubling them does not improve substantially final results. Our best guess as to the cause of this anomaly is that T_p is obtained as the *exact* solution to an *approximate* problem, namely the optimization of an asymptotic power. It appears that these phenomenon appears if the expectation of the alternative distribution coincides with that of the null hypothesis and it becomes more significant for those parameter values for which the variance of the alternative distribution is close to that of the null hypothesis. An asterisk denotes entries in the tables having this problem.

Acknowledgment

The work is supported in part by 2001SGR-00067 (Generalitat de Catalunya).

References

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

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

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

David, H. A. (1981). *Order Statistics*. 2nd ed. New York: John Willey & Sons, Inc.

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

Durbin, J., Knott, M. (1975). Components of Cramér-Von Mises statistics. II. *J. Roy. Statist. Soc. B* 37:216–237.

Fortiana, J., Cuadras, C. M. (1997). A family of matrices, the discretized brownian bridge, and distance-based regression. *Linear Alge. Applic.* 264:173–188.

Fortiana, J., Grané, A. (2003). Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions. *J. Roy. Statist. Soc. B* 65(1):115–126.

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

Shorack, G. R., Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons.

Stephens, M. A. (1974). Components of goodness-of-fit statistics. *Annales de l'Institut Henri Poincaré, Section B* 10:37–54.

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

Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. *Ann. Statist.* 2:676–693. Correction in: Vol. 7, (1979), p. 466.

Wolfram, S. (1996). *The Mathematica Book*. 3rd ed. Cambridge University Press.