

*International Statistical Review* (1993), **61**, 3, pp. 419–433. Printed in Great Britain


# A Consistent Test for Bivariate Dependence

**Andrey Feuerverger**

*Department of Statistics, University of Toronto, Ontario, M5S 1A1, Canada*

## Summary

**A new and consistent rank test for bivariate dependence is developed. Let $X_i'$ and $Y_i'$ denote the (approximate) normal scores associated with the iid vectors $(X_i, Y_i)$, $i = 1, \ldots, n$. Then the proposed test statistic may be obtained by removing the first Hájek projection from the quantity $\xi \equiv n^{-2} \sum \sum |X_j' - X_k'| \cdot |Y_j' - Y_k'|$. Empirical characteristic function considerations are used in our development and some related graphical methods are proposed. Some difficulties that arise in extensions to dimension $k > 2$ are noted. A small simulation study provides evidence of the effectiveness of the new procedure.**

*Key words:* Test for dependence; nonparameteric test; consistency; graphical methods; empirical characteristic function.

## 1 Introduction

This paper is concerned with the general problem of testing for dependence in the bivariate case. In this context we consider the independent, identically distributed random vectors $(X_j, Y_j)$, $j = 1, \ldots, n$, but with unknown joint distribution, and we seek to test the null hypothesis $H_0$ that $X$ and $Y$ are independent. Our emphasis will be on procedures consistent against nonparametric alternative classes, and hence appropriate for data from potentially complex distributional structures. In simpler situations, where the dependency has what can loosely be described as monotone character, the product moment correlation is generally used, and when robustness against nonnormality is important, tests such as the Spearman rank-correlation, Kendall-tau, and Fisher–Yates normal scores tests are often used instead. But these tests are not consistent in general, and in fact will not be appropriate even in the simple situation where $Y$ has a non-monotone regression on $X$, and $X$ is sampled randomly. More interesting situations can arise in engineering applications where signals are modeled and simulated using expansions such as those of Karhunen–Loève and it is of interest to test for dependency amongst the uncorrelated random coefficients that appear in such models. Other potential applications involve testing the adequacy of random number generators. Although our emphasis here is on the bivariate context, it is applications such as these, wherein the forms of dependency can be of a complex nature, that motivate the study of consistent tests for dependence.

In part because simple forms of dependence are the most common, there is only a modest literature on the general problem of testing for dependence. The main exception is the $H_0$-distribution-free rank test first proposed by Hoeffding (1948) and given later in a more transparent form by Blum, Kiefer & Rosenblatt (1961). This test is based upon

$$B_n = \int\int [F_n(x, y) - F_n^X(x) F_n^Y(y)]^2 \, dF_n(x, y) \tag{1.1}$$

where $F_n$ is the sample cdf, and $F_n^X$, $F_n^Y$ are the sample marginal cdf's for the $(X_j, Y_j)$. An alternative test based upon

$$R_n = \int\int [f_n(x, y) - f_n^X(x) f_n^Y(y)]^2 a(x, y)\, dx\, dy \tag{1.2}$$

where the $f_n$ are kernel density estimates and $a(x, y)$ is a weight function was studied by Rosenblatt (1975). This test is not $H_0$-distribution-free, although in itself this fact does not in any important way affect the applicability of (1.2) as a practical testing procedure. In contrast however, Rosenblatt reported that tests based on density estimates typically are less powerful than those based on sample distribution functions; this fact is of significant concern for applications. Also surprising are the very different asymptotic distributional characters of (1.1) and (1.2) which, under $H_0$, are a weighted sum of independent $\chi_1^2$ variates and normally distributed respectively, while under (fixed) alternatives are both normally distributed. In our discussions below, we shall uncover a reason for these irreconcilable differences.

Although we shall use Fourier considerations, and in particular empirical characteristic functions (ecf's), to develop our proposal for a test statistic, the resulting procedures are more general than the context from which they arise. In fact our main proposal may be introduced and motivated in a suggestive alternative manner. To this end let $X_i'$ and $Y_i'$ denote the (approximate) normal scores associated with the $X_i$ and $Y_i$ respectively. (Our definition for the approximate scores is given in (8.1) below.) Making use of the well known $U$-statistics identity

$$\frac{1}{n}\sum U_j V_j - \left(\frac{1}{n}\sum U_j\right)\left(\frac{1}{n}\sum V_j\right) = \frac{1}{2n^2}\sum_j\sum_k (U_j - U_k)(V_j - V_k), \tag{1.3}$$

we start here with the Fisher–Yates normal scores test statistic which we shall write in the form

$$\frac{1}{2n^2}\sum_j\sum_k (X_j' - X_k')(Y_j' - Y_k'). \tag{1.4}$$

We next make an adjustment to the statistic (1.4); interjecting absolute value signs (and removing the factor of 2 from the denominator for convenience) we arrive at the form

$$\xi \equiv \frac{1}{n^2}\sum_j\sum_k |X_j' - X_k'| \cdot |Y_j' - Y_k'|. \tag{1.5}$$

Lastly, we shall now remove from (1.5) its first order Hájek projection. To do this we note that under independence

$$\mu \equiv E_0\xi = \frac{1}{n^3(n-1)}\left(\sum_j\sum_k |X_j' - X_k'|\right)\left(\sum_j\sum_k |Y_j' - Y_k'|\right), \tag{1.6}$$

so that the Hájek projection under $H_0$ is

$$\begin{aligned}\sum_j E_0[\xi - \mu \mid (X_j', Y_j')] = {} & \frac{2}{n(n-1)(n-2)}\sum_j\sum_k\sum_l |X_j' - X_k'| \cdot |Y_j' - Y_l'| \\ & - \frac{2}{n^2(n-1)(n-2)}\sum_j\sum_k\sum_l\sum_m |X_j' - X_k'| \cdot |Y_l' - Y_m'|. \end{aligned} \tag{1.7}$$

The calculations in (1.6) and (1.7) involve algebra based on the null distribution of random assignment of the $X'$ scores to $Y'$ scores. Consequently the projection corrected

form of (1.5), which is the difference between (1.5) and (1.7) is given by

$$\frac{1}{n^2}\sum_j\sum_k |X'_j - X'_k| \cdot |Y'_j - Y'_k| - \frac{2}{n(n-1)(n-2)}\sum_j\sum_l\sum_m |X'_j - X'_l| \cdot |Y'_j - Y'_m|$$
$$+\frac{2}{n^2(n-1)(n-2)}\sum_j\sum_k\sum_l\sum_m |X'_j - X'_k| \cdot |Y'_l - Y'_m|. \quad (1.8)$$

and this is equivalent to the test statistic we shall obtain using ecf methods below.

Our main proposal for a test statistic for dependence is an ecf version of the statistics (1.1) and (1.2) and is presented in Section 4, but first, some related graphical methods are presented in Section 2, and some properties of dependence-related ecf quantities are discussed in Section 3. In Section 5, the relation to Rosenblatt's tests is discussed, and in Section 6, the asymptotic distributions are derived. Some analytical difficulties occur in extensions to dimension $k > 2$; this is described in Section 7. Finally, in Section 8, a brief numerical study is presented.

## 2 Some Graphical Tools—The Correlographs

New graphical tools of interest arise from the observation that $X$ and $Y$ are independent if and only if $\text{cov}[f(X), g(Y)] = 0$ for all $f$ and $g$ ranging over a separating class of functions (see for example Breiman, 1968, p. 165ff). Using a statistical computing package such as *S*, plots of covariances or correlations over such classes are easily obtained and may provide useful and interesting information concerning the structure of dependence. For example, consideration of the well-known separating class consisting of the functions $\cos(tx)$, $\sin(tx)$, $t \geqslant 0$ leads us to define the functions

$$\rho_{cc}(s, t) = \text{cov}[\cos(sX), \cos(tY)] \quad (2.1)$$

$$\rho_{cs}(s, t) = \text{cov}[\cos(sX), \sin(tY)] \quad (2.2)$$

$$\rho_{sc}(s, t) = \text{cov}[\sin(sX), \cos(tY)] \quad (2.3)$$

$$\rho_{ss}(s, t) = \text{cov}[\sin(sX), \sin(tY)]. \quad (2.4)$$

Figures 1a–d show a typical empirical realization of these functions for a standard bivariate normal sample of size $n = 200$ generated under independence, except that in these figures, sample correlations rather than covariances have been plotted. The domains in these perspective plots are $0 < s \leqslant 3{\cdot}0$, $0 < t \leqslant 3{\cdot}0$; the $s$-axis is at 45 degrees and the $t$-axis is at 135 degrees as shown in Fig. 1a. To obtain these plots, each function was first rescaled by its maximum absolute value in order to obtain satisfactory output from the *S*-routine 'persp' that was used to produce the plots. Hence in order to interpret any plot, its scale must be known; to this end, the maximum and minimum values are provided below each graph. In Figs. 1a–d we see that the correlations are all quite modest except possibly for the maximum of 0·167 in the sine–sine correlation plot. However this value is well within the null hypothesis, as a little experience with such plots shows, and occurs essentially because the sample size $n = 200$ is rather modest, and many correlations are being computed.

Figures 2a–d were produced in a like manner, except now are based on a sample of size $n = 200$ of bivariate normals of means 0, variances 1, and correlation $\frac{1}{2}$. The essential significant feature in these plots is the high level plane emanating from the origin in the sine–sine correlation plot. This behaviour is quite typical of situations in which dependence is essentially of a simple linear kind, and data are centred at the origin.

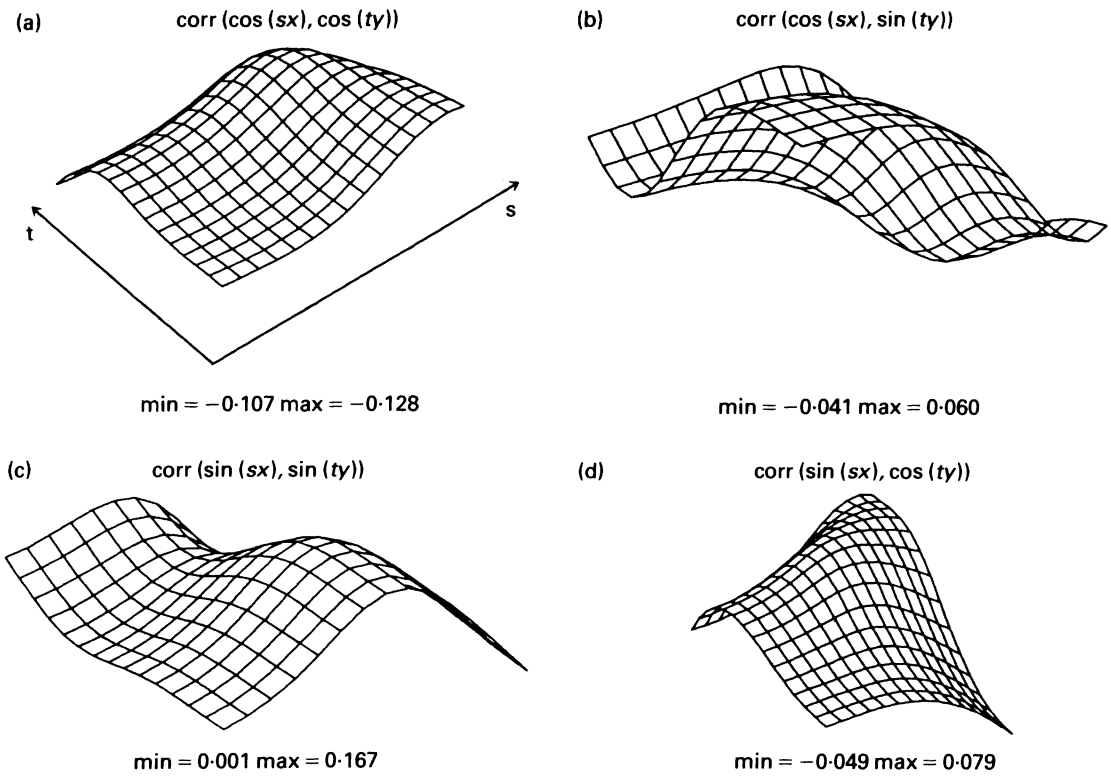


**Figure 1.** *Correlographs for a sample of size* $n = 200$ *from the standard bivariate normal distribution (independent).*

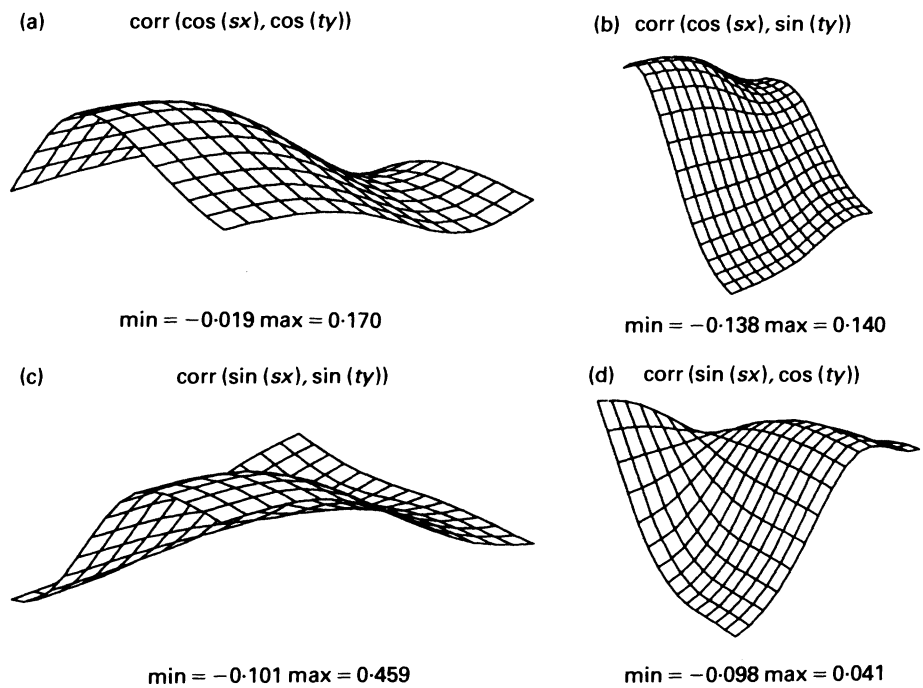


**Figure 2.** *Correlographs for a sample of size* $n = 200$ *from the bivariate normal distribution with means* 0, *variances* 1, *and correlation* $\frac{1}{2}$.

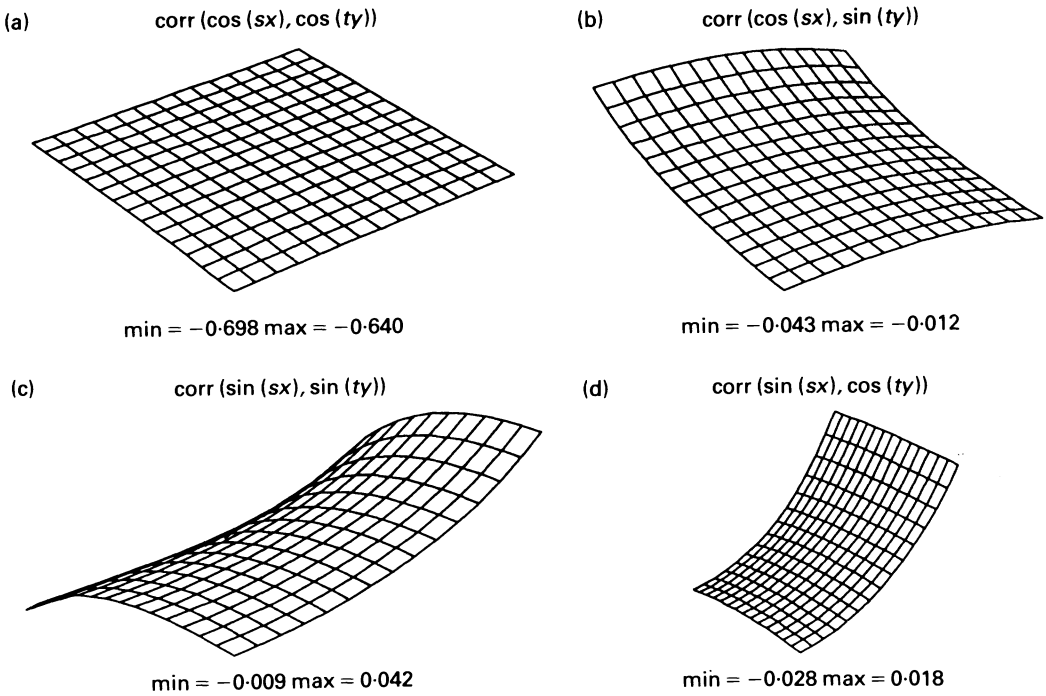


**Figure 3.** *Correlographs for a sample of size* $n = 200$ *from an annular distribution.*

Finally, Figs. 3a–d were produced from a sample of $n = 200$ $(X, Y)$ points in the plane each chosen by selecting a radius uniformly in the interval $[0{\cdot}5, 1]$ and then an angle uniformly on $[0, 2\pi)$. In these plots the significant feature is the very low cosine–cosine correlation plane which ranges over large negative values. This kind of behaviour is typical of annular type distributions when centred at the origin.

Of course, there are many possible variants of these plots to experiment with. For example, the functions shown could be standardized by dividing them by their estimated standard error functions which may readily be derived; alternatively data can be replaced, say, by their normal scores prior to the computations. We will not pursue these variants here.

The quantities (2.1)–(2.4) are related to certain empirical characteristic function (ecf) quantities which we shall make use of below. Specifically let

$$c(s, t) = Ee^{i(sX+tY)}, \quad c^X(s) = Ee^{isX}, \quad c^Y(t) = Ee^{itY}$$

and define the dependence measure

$$\Gamma(s, t) = c(s, t) - c^X(s)c^Y(t);$$

define also the empirical quantities

$$c_n(s, t) = \frac{1}{n}\sum_{j=1}^{n} e^{i(sX_j+tY_j)}, \quad c_n^X(s) = \frac{1}{n}\sum_j e^{isX_j}, \quad c_n^Y(t) = \frac{1}{n}\sum_j e^{itY_j}$$

and

$$\Gamma_n(s, t) = c_n(s, t) - c_n^X(s)c_n^Y(t).$$

Then the following relation (which holds also in the empirical version) is easily verified:

$$\Gamma(s, t) = [\rho_{cc}(s, t) - \rho_{ss}(s, t)] + i \cdot [\rho_{cs}(s, t) + \rho_{sc}(s, t)].$$

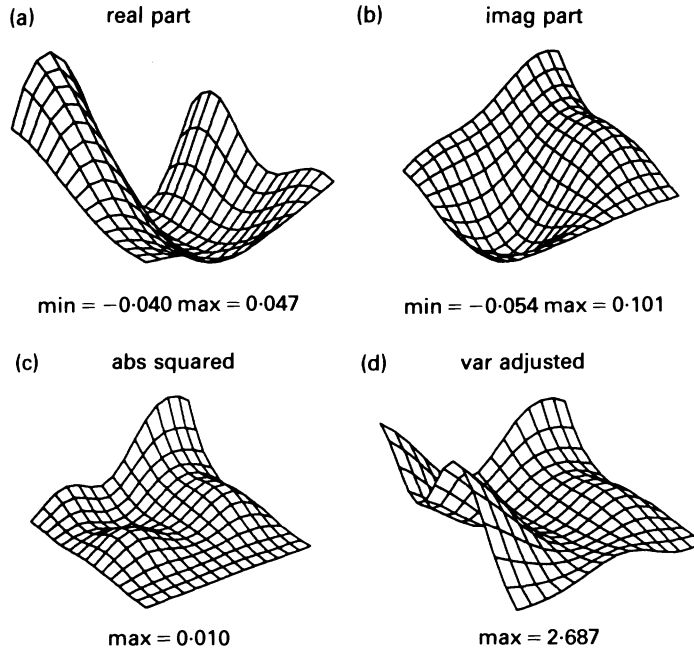


**Figure 4.** *Correlographs for a sample of size* $n = 200$ *from the standard bivariate normal distribution* (*independent*).

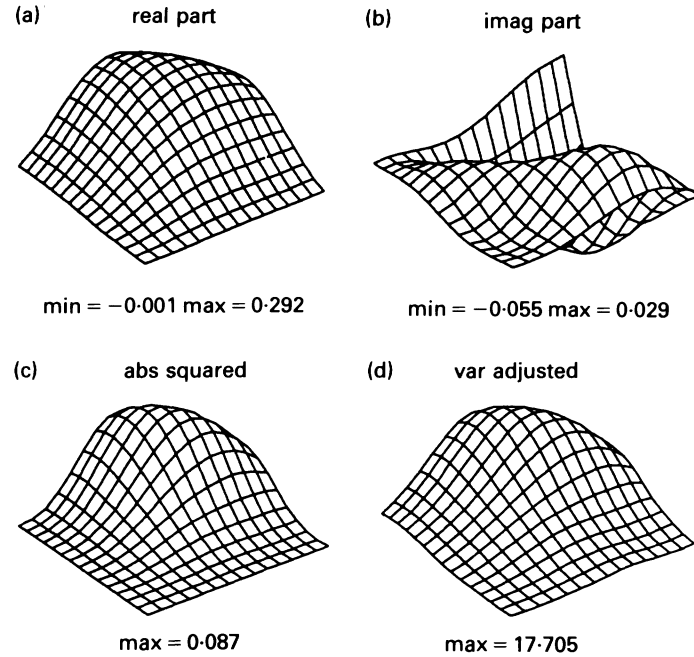


**Figure 5.** *Correlographs for a sample of size* $n = 200$ *from the* $(UZ, UZ')$ *distribution* (*see text*).

Inverse relations, based on natural symmetries, are also easily obtained:

$$\rho_{cc}(s, t) = \tfrac{1}{2} \cdot \text{Re}\,\{\Gamma(s, t) + \Gamma(s, -t)\}$$
$$\rho_{ss}(s, t) = -\tfrac{1}{2} \cdot \text{Re}\,\{\Gamma(s, t) - \Gamma(s, -t)\}$$
$$\rho_{cs}(s, t) = \tfrac{1}{2} \cdot \text{Im}\,\{\Gamma(s, t) + \Gamma(-s, t)\}$$
$$\rho_{sc}(s, t) = \tfrac{1}{2} \cdot \text{Im}\,\{\Gamma(s, t) + \Gamma(s, -t)\}.$$

Figures 4a–d show a typical empirical realization of the functions $\text{Re}\,\Gamma_n(s, t)$, $\text{Im}\,\Gamma_n(s, t)$, $|\Gamma_n(s, t)|^2$, and $n \cdot (1 - e^{-s^2})^{-1} \cdot (1 - e^{-t^2})^{-1} \cdot |\Gamma_n(s, t)|^2$ for a standard bivariate normal sample of size $n = 200$ generated under independence, except that it is the approximate normal scores of the data that were actually used. Note that the fourth plot is just that obtained from the third on dividing it by the asymptotic variance function of the scored version of $\Gamma_n(s, t)$; see equation (3.4) below.

Figures 5a–d were produced in like manner except now based on a sample of $n = 200$ observations of $X_i = U_i Z_i$, $Y_i = U_i Z_i'$, where the $Z_i$, $Z_i'$ are independent $N(0, 1)$ variables and the $U_i$ are independent uniform variables on $[0, 2]$. Because $X_i$, $Y_i$ share a common value of $U_i$, they are not independent, but this fact is quite difficult to discern from an $(X, Y)$-scatterplot. On the other hand, the variance-scaled Fig. 5d provides very clear evidence of the presence of dependence.

For data sets of moderate size or larger, all of these displays carry significant information concerning the structure of dependency. We refer to displays such as those shown here collectively as correlographs.

## 3 Some Dependence Processes

The properties of the ecf are discussed, for example, in Feuerverger & Mureika (1977), Feuerverger & McDunnough (1981) and references appearing therein. See also Csörgő (1985) and Feuerverger (1987). By straightforward computations, the ecf quantities are found to be unbiased and consistent estimators of the corresponding cf's. The properties of $\Gamma_n(s, t)$ are readily established using similar methods and we find that $\Gamma_n(s, t)$ is consistent for $\Gamma(s, t)$ and has mean

$$E\Gamma_n(s, t) = \frac{n-1}{n}\Gamma(s, t).$$

Since independence of $X, Y$ may be characterized by $\Gamma(s, t) = 0$ for all $(s, t) \in R^2$, consistent tests of $H_0$ may be based on the idea of comparing $\Gamma_n(s, t)$ to zero.

Now, the covariance structure

$$\begin{aligned}\Sigma_n(s_1, t_1, s_2, t_2) &\equiv \text{cov}\,(\Gamma_n(s_1, t_1), \Gamma_n(s_2, t_2)) \\ &\equiv E(\Gamma_n(s_1, t_1) - E\Gamma_n(s_1, t_1))\overline{(\Gamma_n(s_2, t_2) - E\Gamma_n(s_2, t_2))}\end{aligned}$$

as well as the limiting covariance form

$$\Sigma(s_1, t_1, s_2, t_2) = \lim_{n \to \infty} n \cdot \Sigma_n(s_1, t_1, s_2, t_2) \tag{3.1}$$

may be determined in a straightforward manner, but the resulting expressions are very lengthy and thus we omit them here. Under $H_0$ however the expressions simplify and we find

$$\text{cov}_0\,(\Gamma_n(s_1, t_1), \Gamma_n(s_2, t_2)) = \frac{n-1}{n^2}\Sigma_0(s_1, t_1, s_2, t_2)$$

where

$$\Sigma_0(s_1, t_1, s_2, t_2) = (c^X(s_1 - s_2) - c^X(s_1)\overline{c^X(s_2)}) \cdot (c^Y(t_1 - t_2) - c^Y(t_1)\overline{c^Y(t_2)}). \quad (3.2)$$

(Remark: this simple form of factoring does not extend to dimension higher than two.) It is also readily established that at finite collections of points $(s_k, t_k)$, $k = 1, \ldots, K$ the process $n^{\frac{1}{2}}(\Gamma_n(s, t) - \Gamma(s, t))$ is asymptotically normal with mean zero and covariance function $\Sigma$ referred to in (3.1) above; under $H_0$ the covariance function is given by $\Sigma_0$ as in (3.2). Further, weak convergence (in the space of continuous functions under the supremum norm) may be shown to hold under mild conditions (Csörgő, 1985).

Concerning the shape of $\Gamma(s, t)$, note that $\Gamma(s, t)$ is bounded and that $\Gamma(s, t) = 0$ on the axes $s = 0$ or $t = 0$. Also, since $\Gamma(s, -t) = \overline{\Gamma(-s, -t)}$, the natural domain of definition for $\Gamma(s, t)$ can be taken as any half plane with boundary passing through the origin. Related to this, since $\overline{\Gamma_n(s, t)} = \Gamma_n(-s, -t)$, the covariance structure for the real and imaginary components $\operatorname{Re}\Gamma_n(s, t)$, $\operatorname{Im}\Gamma_n(s, t)$ may readily be computed from the complex-valued covariance forms. Note also that if $(X, Y)$ are absolutely continuous (i.e. possess density relative to the Lebesque measure on $R^2$) then $\Gamma(s, t) \to 0$, as $\|(s, t)\| \to \infty$. This follows from the Reimann–Lebesque Lemma (see for example Kawata 1972, p. 43) as a consequence of which both $c(s, t)$ and $c^X(s) \cdot c^Y(t) \to 0$ as $\|(s, t)\| \to \infty$. (It does not hold in general: consider $X \equiv Y$ taking on values 0 or 1 with probability $\frac{1}{2}$ each.)

To develop nonparametric (i.e. $H_0$-distribution-free) tests we shall need to replace the $X_i$ and $Y_i$ by approximate normal score quantities $X_i'$ and $Y_i'$. This leads us to consider the associated quantities $c_n^{X'}(s)$, $c_n^{Y'}(s)$, $c_n'(s, t)$ and

$$\Gamma_n'(s, t) = c_n'(s, t) - c_n^{X'}(s) \cdot c_n^{Y'}(t)$$

calculated from the scored data. In this context, the marginal quantities $c_n^{X'}(\cdot)$ and $c_n^{Y'}(\cdot)$ will approximate the function $e^{-\frac{1}{2}t^2}$ and will no longer be random. Consequently, under $H_0$ we will have

$$\begin{aligned}\operatorname{cov}_0(\Gamma_n'(s_1, t_1), \Gamma_n'(s_2, t_2)) &= \operatorname{cov}_0(c_n'(s_1, t_1), c_n'(s_2, t_2)) \\ &= \frac{1}{n-1}[\phi_n(s_1 - s_2) - \phi_n(s_1)\overline{\phi_n(s_2)}][\phi_n(t_1 - t_2) - \phi_n(t_1)\overline{\phi_n(t_2)}],\end{aligned} \quad (3.3)$$

where $\phi_n$ is the ecf corresponding to the approximate normal scores. This computation involves the straightforward combinatorics of uniformly random assignments of $X_i'$ to $Y_i'$ scores, as appropriate to the $H_0$ distribution. It follows from (3.3) that

$$n \cdot \operatorname{Var}(\Gamma_n'(s, t)) \to (1 - e^{-s^2}) \cdot (1 - e^{-t^2}). \quad (3.4)$$

Finally, we need to consider the stochastic process

$$\xi_n(s, t) = n^{\frac{1}{2}} \cdot (c_n'(s, t) - Ec_n'(s, t)).$$

By applying the Wald–Wolfowitz Theorem (e.g. Serfling 1980, p. 297) in conjunction with the Cramér–Wold device, we may show that, under $H_0$, the finite dimensional distributions of $\xi_n$ converge in distribution to those of a zero mean Gaussian process $\xi(s, t)$ such that

$$E\xi(s_2, t_2)\overline{\xi(s_1, t_1)} = [\phi(s_2 - s_1) - \phi(s_2)\overline{\phi(s_1)}][\phi(t_2 - t_1) - \phi(t_2)\overline{\phi(t_1)}],$$

where $\phi(t) = e^{-t^2/2}$. Next, again under $H_0$, we may see that

$$\begin{aligned}
&\frac{n-1}{n} \cdot E\,|\xi_n(s_2, t_2) - \xi_n(s_1, t_1)|^2 = [1 - |\phi_n(s_2)|^2][1 - |\phi_n(t_2)|^2] + [1 - |\phi_n(s_1)|^2][1 - |\phi_n(t_1)|^2] \\
&\quad - 2 \cdot \mathrm{Re}\,\{[\phi_n(s_2 - s_1) - \phi_n(s_2)\overline{\phi_n(s_1)}][\phi_n(t_2 - t_1) - \phi_n(t_2)\overline{\phi_n(t_1)}]\} \\
&\leq 2[1 - \mathrm{Re}\,\{\phi_n(s_2 - s_1)\phi_n(t_2 - t_1)\}] \\
&= 2\int\int [1 - \cos(x(s_2 - s_1) + y(t_2 - t_1)]\, dF_n'(x)\, dF_n'(y) \\
&\leq 2\int\int (x(s_2 - s_1) + y(t_2 - t_1))^2\, dF_n'(x)\, dF_n'(y) \\
&\leq K \cdot [(s_2 - s_1)^2 + (t_2 - t_1)^2].
\end{aligned}$$

In these calculations $\phi_n$ and $F_n'$ are the ecf and the distribution function corresponding to the scores, and $K$ is a constant. It therefore follows from Theorem 8 of Whitt (1970) and Theorem 12.3 of Billingsley (1969) that the process $\xi_n(s, t)$ converges weakly to the process $\xi(s, t)$ on compact regions. A similar argument is applicable in the case of the alternatives.

## 4 Development of the Test Statistics

We are now in a position to make a specific proposal for a useful class of nonparametric test statistics for dependence, namely

$$T_n = \int\int \frac{|\Gamma_n'(s, t)|^2}{(1 - e^{-s^2})(1 - e^{-t^2})} W(s, t)\, ds\, dt \tag{4.1}$$

where $W(s, t) > 0$ is an appropriate weight function. The denominator in (4.1), which equals $n$ times the limiting variance function (3.4) of $\Gamma_n'(s, t)$, tends to 0 as $s$ or $t$ tends to 0. Nevertheless, the integrand may be defined by continuity at these limiting values (as in the following paragraph). The presence of the (variance) divisor is an important aspect of (4.1); without it, the effects measured at $s = 0$ and/or $t = 0$ would be largely eliminated from the overall test statistic and substantial power would be lost for many typical alternatives.

Now one suitable and particularly convenient bell-shaped choice for the weight in (4.1) is the function

$$W(s, t) = \left(\frac{1 - e^{-s^2}}{s^2}\right) \cdot \left(\frac{1 - e^{-t^2}}{t^2}\right) \tag{4.2}$$

which has Cauchy-like tails and results in the test statistic

$$T_n = \int\int \frac{|c_n'(s, t) - c_n^{X'}(s) \cdot c_n^{Y'}(t)|^2}{s^2 \cdot t^2}\, ds\, dt. \tag{4.3}$$

Making use of the notation

$$\mathrm{cov}_{j=1}^n (U_j, V_j) \equiv \frac{1}{n}\sum U_j V_j - \left(\frac{1}{n}\sum U_j\right)\left(\frac{1}{n}\sum V_j\right)$$

here affords the easiest way to see that the quantity

$$\frac{c_n'(s, t) - c_n^{X'}(s) \cdot c_n^{Y'}(t)}{s \cdot t} = \frac{1}{s \cdot t}\,\mathrm{cov}_{j=1}^n (e^{isX_j}, e^{itY_j})$$

can be defined by continuity as

$$\frac{i}{t}\mathrm{cov}_{j=1}^{n}(X_j, e^{itY_j}) \quad \text{when} \quad s=0,\ t\neq 0,$$

$$\frac{i}{s}\mathrm{cov}_{j=1}^{n}(e^{isX_j}, Y_j) \quad \text{when} \quad t=0,\ s\neq 0,$$

and as

$$-\mathrm{cov}_{j=1}^{n}(X_j, Y_j) \quad \text{when} \quad s=t=0.$$

Now use the $U$-statistics identity (1.3) within the modulus of the numerator of (4.3) to get

$$T_n = \int\int \left|\frac{1}{2n^2}\sum_j\sum_k \left(\frac{e^{isX_j'} - s^{isX_k'}}{s}\right)\left(\frac{e^{itY_j'} - e^{itY_k'}}{t}\right)\right|^2 ds\, dt \tag{4.4}$$

$$= \left(\frac{1}{2n^2}\right)^2 \sum_j\sum_k\sum_l\sum_m (\text{term 1})\cdot(\text{term 2}) \tag{4.5}$$

where

$$\text{term 1} = \int_{-\infty}^{\infty} \frac{(e^{isX_j'} - e^{isX_k'})(e^{-isX_l'} - e^{-isX_m'})}{s^2}\, ds \tag{4.6}$$

$$= -\pi\cdot[|X_j' - X_l'| - |X_j' - X_m'| - |X_k' - X_l'| + |X_k' - X_m'|] \tag{4.7}$$

and term 2 is identical except with $t$ and $Y$ in place of $s$ and $X$. The step from (4.6) to (4.7) is obtained by multiplying out the exponentials in the numerator of (4.6), dropping the sines which are odd functions, and then using the identity

$$\int_{-\infty}^{\infty} \frac{\cos(a\cdot x) - 1}{x^2}\, dx = -\pi\cdot|a|.$$

See for example Gradshteyn & Ryzhik (1980, item 3.784.3). We are thus led to the test statistic

$$T_n = \frac{\pi^2}{4n^4}\sum_j\sum_k\sum_l\sum_m [|X_j' - X_l'| - |X_j' - X_m'| - |X_k' - X_l'| + |X_k' - X_m'|]$$
$$\cdot[|Y_j' - Y_l'| - |Y_j' - Y_m'| - |Y_k' - Y_l'| + |Y_k' - Y_m'|].$$

Multiplying out and resubscripting the sixteen terms that arise we find

$$T_n = \frac{\pi^2}{n^2}\sum_j\sum_k |X_j' - X_k'|\cdot|Y_j' - Y_k'| - \frac{2\pi^2}{n^3}\sum_j\sum_l\sum_m |X_j' - X_l'|\cdot|Y_j' - Y_m'|$$
$$+\frac{\pi^2}{n^4}\sum_j\sum_k\sum_l\sum_m |X_j' - X_k'|\cdot|Y_l' - Y_m'|. \tag{4.8}$$

Note that the three terms on the right in (4.8) can each be computed in $O(n^2)$ operations, and that the last term is a constant which depends only on $n$.

The choice (4.2) is not the only tractable possibility for a weight function in (4.1). Thus consider choices of the form

$$W(s,t) = \left(\frac{1-e^{-s^2}}{s^2}\right)\cdot\left(\frac{1-e^{-t^2}}{t^2}\right)\cdot V(s)\cdot V(t.$$

Following the same computational steps as before, we again arrive at (4.5) except that now

$$\text{term 1} = \int_{-\infty}^{\infty} \frac{(e^{isX_j'} - e^{isX_k'})(e^{-isX_l'} - e^{-isX_m'})}{s^2}\cdot V(s)\, ds,$$

and similarly for term 2. We may progress from here by first taking note of the identity

$$\int_{-\infty}^{\infty} \frac{[\cos(a \cdot x) - 1]\cos(b \cdot x)}{x^2}\, dx = -\pi \cdot (|b| - |a|),$$

when $|b| \leq |a|$, and equals 0 when $|a| < |b|$ (Gradshteyn and Ryzhik, 1980, item 3.786.3). Therefore if $V(s)$ can be represented in the form

$$V(s) = \int_0^{\infty} \cos(s \cdot u)\, dG(u) \tag{4.9}$$

we will then obtain

$$\text{term } 1 = g(X_j' - X_l') - g(X_j' - X_m') - g(X_k' - X_l') + g(X_k' - X_m')$$

where

$$g(x) = \pi \cdot \left( \int_0^{|x|} u\, dG(u) - |x| \cdot [G(|x|) - G(0)] \right), \tag{4.10}$$

and hence the test statistic

$$\begin{aligned} T_n &= \frac{\pi^2}{4n^4} \sum_j \sum_k \sum_l \sum_m [g(X_j' - X_l') - g(X_j' - X_m') - g(X_k' - X_l') + g(X_k' - X_m')] \\ &\quad \cdot [g(Y_j' - Y_l') - g(Y_j' - Y_m') - g(Y_k' - Y_l') + g(Y_k' - Y_m')], \\ &= \frac{\pi^2}{n^2} \sum_j \sum_k g(X_j' - X_k') \cdot g(Y_j' - Y_k') - \frac{2\pi^2}{n^3} \sum_j \sum_l \sum_m g(X_j' - X_l') \cdot g(Y_j' - Y_m') \\ &\quad + \frac{\pi^2}{n^4} \sum_j \sum_k \sum_l \sum_m g(X_j' - X_k') \cdot g(Y_l' - Y_m'). \end{aligned} \tag{4.11}$$

This will be tractable for any appropriate choice of $G$ in (4.9) that results in a tractable $g$ in (4.10).

Let us now return to the statistic (4.8). Firstly, on closer examination we see that, to an adequate degree of approximation, (4.8) is essentially a bivariate, rank-based, $U$ statistic (based on a kernel of order 2) and is of degree 2 in the sense that, except for constant terms, its first order Hájek projection under $H_0$ is essentially null. The calculations to back up this assertion were provided in Section 1 where the test statistic (1.8) was obtained by removing the projection from (1.5). Now, the final terms in (1.8) and (4.8) are constants and thus may be ignored. We may also ignore the irrelevant $\pi^2$ factor in (4.8). We thus see the essential difference between (1.8) and (4.8) is just

$$\left[ \frac{2}{n(n-1)(n-2)} - \frac{2}{n^3} \right] \sum_j \sum_l \sum_m |X_j' - X_l'| \cdot |Y_j' - Y_m'|$$

which is $O_P(n^{-1})$ so that tests based on (4.8) and on the projection corrected form of (1.5) are asymptotically equivalent.

Secondly, there is an approximation to (4.8) which is computationally convenient. Note that the second term in (4.8) involves repetitive approximation to the function

$$q(x) = E|x - Z| = (2/\pi)^{\frac{1}{2}} \cdot e^{-x^2/2} + 2x\Phi(x) - x$$

where $Z \sim N(0, 1)$ and $\Phi$ is the $N(0, 1)$ distribution function, while the third term involves approximating the square of

$$Eq(Z) = \frac{2}{\pi^{\frac{1}{2}}}.$$

Dropping the $\pi^2$ factor in (4.8) we thus arrive at the following modification of the test statistic:

$$T_n^{\text{mod}} = \frac{1}{n^2}\sum_j \sum_k |X_j' - X_k'|\,|Y_j' - Y_k'| - \frac{2}{n}\sum_j q(X_j')\cdot q(Y_j') + \frac{4}{\pi}. \tag{4.12}$$

The difference between (4.8) and (4.12) depends on the specific approximate scores used, but here was found numerically to be approximately equal to $3.1/n$ over a very broad range of conditions that were tested.

## 5 Relation to Rosenblatt's Tests

Suppose now that we disregard the initial variance rescaling, and also that we use unscored data, so that instead of (4.1) we choose to work with

$$T_n = \int\int |\Gamma_n(s, t)|^2\, W(s, t)\, ds\, dt. \tag{5.1}$$

Then apply the Parseval Theorem to write (5.1) alternatively as

$$T_n = \int\int \left| \int\int V(x-u, y-v)\, d^2[F_n(u, v) - F_n^X(u)F_n^Y(v)] \right|^2 dx\, dy \tag{5.2}$$

where $V(u, v)$ is the Fourier transform of $[W(s, t)]^{\frac{1}{2}}$. Observe now that the inner integrals in (5.2) are a convolution so that $T_n$ evidently is a version of Rosenblatt's density estimation-based statistic (1.2). However there is one important difference. Rosenblatt (1975) requires that a bandwidth parameter $b(n) \downarrow 0$ in order that the resulting density estimates be consistent. In the statistic (5.2) however, the effective bandwidth is constant and does not vary with $n$. This has implications for the asymptotic distributional character of the statistic. In particular, it is now apparent that *consistency* of the associated density estimators is neither a relevant nor helpful property for the dependence testing context. Furthermore, the Rosenblatt statistics (1.2) do not ordinarily correspond to a weight function $W$ in (5.1) which emphasizes the origin and axes in the manner of (4.1) thereby resulting in additional losses of power for many typical situations.

## 6 Asymptotic Distributions

Using the weak convergence result of Section 3 and standard arguments, we find that under $H_0$ (4.1) is of the form

$$n^{-1}\cdot \int\int |\zeta_n(s, t)|^2\, ds\, dt$$

where $\zeta_n(s, t)$ converges weakly to a Gaussian process $\zeta(s, t)$ having mean 0 and covariance function

$$\begin{aligned} v(s_1, t_1, s_2, t_2) &\equiv \operatorname{cov}(\zeta(s_1, t_1), \zeta(s_2, t_2)) \\ &= [e^{-(s_1-s_2)^2/2} - e^{-(s_1^2+s_2^2)/2}][e^{-(t_1-t_2)^2/2} - e^{-(t_1^2+t_2^2)/2}] \\ &\quad \times \left(\frac{W(s_1, t_1)}{(1-e^{-s_1^2})(1-e^{-t_1^2})}\right)^{\frac{1}{2}} \left(\frac{W(s_2, t_2)}{(1-e^{-s_2^2})(1-e^{-t_2^2})}\right)^{\frac{1}{2}}, \end{aligned} \tag{6.1}$$

so that $n$ times (4.1) has asymptotically a $\sum \lambda_j \cdot Z_j^2$ distribution where, for $j = 1, 2 \cdots$ the $Z_j$ are independent $N(0, 1)$ variables and the $\lambda_j$ are the solutions of the eigenvalue equation

$$\lambda_j \cdot f_j(s_1, t_1) = \int\int f_j(s_2, t_2)\cdot v(s_1, t_1, s_2, t_2)\, ds_2\, dt_2. \tag{6.2}$$

In the event that we have the factorization $W(s, t) = W_1(s) \cdot W_1(t)$, the covariance function (6.1) will factor as $v(s_1, s_2, t_1, t_2) = v_1(s_1, s_2) \cdot v_1(t_1, t_2)$ where

$$v_1(s_1, s_2) = [e^{-(s_1-s_2)^2/2} - e^{-(s_1^2+s_2^2)/2}]\left(\frac{W_1(s_1) \cdot W_1(s_2)}{(1-e)^{-s_1^2})(1-e^{-s_2^2})}\right)^{\frac{1}{2}},$$

so that equation (6.2) separates and the eigenvalues of (6.2) will range over all pairwise products of the eigenvalues of

$$\lambda_j \cdot f_j(s_1) = \int f_j(s_2) \cdot v_1(s_1, s_2)\, ds_2.$$

When $W$ is as in (4.2) we will have

$$v_1(s_1, s_2) = \frac{e^{-(s_1^2+s_2^2)/2}[e^{s_1 s_2} - 1]}{|s_1 \cdot s_2|}$$

however an exact solution for the eigenvalues here does not appear to be straightforward. Instead of solving such eigenvalue equations numerically, we shall prefer to obtain the distributions of the test statistics by Monte Carlo means.

Under the alternatives we have, following standard arguments, that (4.1) is asymptotically normal, i.e. $T_n \to N(\mu, \sigma^2)$ in distribution; the expressions for $\mu$ and $\sigma^2$ are very lengthy and here omitted.

## 7 Concerning Multivariate Extensions

The purpose of this section is to indicate briefly that extension to dimensions $k > 2$ is not straightforward. The $k$-variate forms for all quantities, including $\Gamma_n$ and $\Gamma'_n$, are immediate and their exact and limiting covariance structures may be readily determined; we omit these expressions here. However, factorization analogous to (3.2) and (3.3) does not occur for $k > 2$. Likewise, the factoring step which carries (4.3) into the form (4.4) also has no analogue for $k > 2$. Consequently there is no natural analogue for the results of Section 4.

Difficulties with dimension $k > 2$ have been noted elsewhere. Some useful ideas may be found in Section 3 of Blum, Kiefer & Rosenblatt (1961); see especially their form (3.3). See also Deheuvels (1981) and Lancaster (1960). However, the ensuing lines of inquiry are extensive and are not pursued here.

## 8 Some Numerical Studies

Some numerical trials and a brief Monte Carlo study were carried out and these confirm the practicality of the procedures proposed here. The definition of the approximate normal scores which we have used throughout was taken to be

$$X'_i = \Phi^{-1}\left(\frac{\text{rank}\,(X_i) - 3/8}{n + 1/4}\right) \tag{8.1}$$

where $\Phi$ is the $N(0, 1)$ distribution function (see e.g., Blom, 1958), and the version of the test statistic used here is $n \cdot T_n$ where

$$T_n = \frac{1}{n^2}\sum_j \sum_k |X'_j - X'_k| \cdot |Y'_j - Y'_k| - \frac{2}{n^3}\sum_j \sum_l \sum_m |X'_j - X'_l| \cdot |Y'_j - Y'_m| + \frac{4}{\pi}. \tag{8.2}$$

Table 8.1 gives the upper percentage points of the distribution of $n \cdot T_n$ under $H_0$

**Table 8.1**

*Critical values for* $n \cdot T_n$

| Sample size | Upper tail probability | | | | |
|---|---|---|---|---|---|
| | $p = 0{\cdot}10$ | $p = 0{\cdot}05$ | $p = 0{\cdot}02$ | $p = 0{\cdot}01$ | $p = 0{\cdot}005$ |
| $n = 100$ | 4·88 | 5·24 | 5·77 | 6·11 | 6·44 |
| $n = 200$ | 5·09 | 5·48 | 6·02 | 6·39 | 6·79 |
| $n = 300$ | 5·17 | 5·56 | 6·15 | 6·54 | 6·96 |
| $n = 400$ | 5·35 | 5·76 | 6·3 | 6·8 | 7·0 |
| $n = 500$ | 5·37 | 5·78 | 6·3 | 6·6 | 7·0 |

**Table 8.2**

*Power of the test statistic* (8.2) *for distributions, sample sizes, and levels shown*

| Dist'n | Level | Sample size ($n$) | | | | |
|---|---|---|---|---|---|---|
| | | 100 | 200 | 300 | 400 | 500 |
| Normal | 0·05 | 0·45 | 0·77 | 0·90 | 0·96 | 0·99 |
| rho = 0·2 | 0·01 | 0·21 | 0·56 | 0·77 | 0·90 | 0·95 |
| Normal | 0·05 | 1·000 | 1·000 | 1·000 | 1·000 | 1·000 |
| rho = 0·5 | 0·01 | 0·998 | 1·000 | 1·000 | 1·000 | 1·000 |
| annulus | 0·05 | 0·43 | 1·00 | 1·00 | 1·00 | 1·00 |
| | 0·01 | 0·05 | 0·85 | 1·00 | 1·00 | 1·00 |
| *UZ, UZ* | 0·05 | 0·19 | 0·45 | 0·79 | 0·87 | 0·97 |
| | 0·01 | 0·07 | 0·10 | 0·25 | 0·35 | 0·76 |

determined, for the sample sizes indicated, by Monte Carlo trials on a Sun 3/60 computer in conjunction with the *S* statistical package and double precision Fortran routines. Table 8.2 gives the powers of the test for various alternative distributions, sample sizes, and levels. The number of Monte Carlo trials used for Table 8.1 were 6,000, 14,000, 4,800, 500 and 2,400 for sample sizes 100, 200, 300, 400 and 500 respectively, and the number of trials used for Table 8.2 were at least 200 trials for each entry. The number of digits selected for reporting in each case is such that the standard error is ordinarily at most one or two units in the last digit provided, and usually much less. The four distributions indicated in Table 8.2 refer to normals with correlations 0.2 and 0.5, the annular distribution discussed in Section 3 in the context of Fig. 3a–d, and the $(U \cdot Z, U \cdot Z')$ distribution discussed in Section 3 in the context of Fig. 5a–d.

## Acknowledgements

The author would like to express his thanks to P.J. Bickel, K. Knight and P.G. Rooney for helpful conversations. This research was supported by an Operating Grant from the Natural Sciences and Engineering Research Council of Canada.

## References

Billingsley, P. (1968). *Convergence of Probability Measures.* New York: Wiley.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates.* New York: Wiley.

Blum, J.R., Kiefer, J. & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* **32,** 485–498.

Breiman, L. (1968). *Probability.* Reading, Massachusetts: Addison-Wesley.

Csörgő, S. (1985). Testing independence by the empirical characteristic function. *J. Multiv. Anal.* **16,** 290–299.

Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multiv. Anal.* **11,** 102–113.

Feuerverger, A. (1987). On some ECF procedures for testing independence. In: *Time Series and Econometric Modeling*, Eds. I.B. MacNeill and G.J. Umphrey, *Joshi Festschrift*, pp. 189–206. New York: Reidel.

Feuerverger, A. & Mureika, R.A. (1977). The empirical characteristic function and its applications. *Ann. Statist.* **5,** 88–97.

Feuerverger, A. & McDunnough, P. (1981). On some Fourier methods for inference. *J. Amer. Statist. Assoc.* **76,** 379–387.

Gradshteyn, I.S. & Ryzhik, I.M. (1980). *Table of Integrals, Series, and Products.* New York: Academic Press.

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Statist.* **19,** 546–557.

Kawata, T. (1972). *Fourier Analysis and Probability Theory.* New York: Academic Press.

Koziol, J.A. & Nemec, A.F. (1979). On a Cramér–von Mises type statistic for testing bivariate independence. *Canadian J. Statist.* **7,** 43–52.

Lancaster, H.O. (1960). On tests of independence in several dimensions. *J. Austral. Math. Soc.*, **1,** 241–254. Corrig. **1,** 496.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3,** 1–14.

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

Whitt, W. (1970). Weak convergence of probability measures on the function space $C[0, \infty)$. *Ann. Math. Statist.* **3,** 939–944.

## Résumé

Un nouveau test du rank convergent pour la dépendance bivariée est exposé dans cet article. Soit $X_i'$ et $Y_i'$ les scores normaux (approchés) associés aux vecteurs iid $(X_i, Y_i)$, $i = 1, \ldots, n$. La statistique du test proposée peut alors être obtenue en enlevant la première projection de Hájek de la quantité $\xi = n^{-2} \sum \sum |X_j' - X_k'| \cdot |Y_j' - Y_k'|$. Des considérations liées aux fonctions caractéristiques empiriques sont utilisées dans notre développement et des méthodes graphiques correspondantes sont proposées. Des difficultés apparaissant lors d'extensions aux dimensions $k > 2$ sont indiquées. Une étude de simulation atteste de l'efficacité de la nouvelle méthode.

[*Received March* 1991, *accepted February* 1992]