ON A NEW CORRELATION COEFFICIENT, ITS ORTHOGONAL DECOMPOSITION AND ASSOCIATED TESTS OF INDEPENDENCE

By Wicher Bergsma[1]

London School of Economics and Political Science

[2] A possible drawback of the ordinary correlation coefficient ρ for two real random variables X and Y is that zero correlation does not imply independence. In this paper we introduce a new correlation coefficient ρ^* which assumes values between zero and one, equalling zero iff the two variables are independent and equalling one iff the two variables are linearly related. The coefficients ρ^* and ρ^2 are shown to be closely related algebraically, and they coincide for distributions on a 2×2 contingency table. We derive an orthogonal decomposition of ρ^* as a positively weighted sum of squared ordinary correlations between certain marginal eigenfunctions. Estimation of ρ^* and its component correlations and their asymptotic distributions are discussed, and we develop visual tools for assessing the nature of a possible association in a bivariate data set. The paper includes consideration of grade (rank) versions of ρ^* as well as the use of ρ^* for contingency table analysis. As a special case a new generalization of the Cramér-von Mises test to K ordered samples is obtained.

Contents:

---

[1] Supported by The Netherlands Organization for Scientific Research (NWO), Project Number 400-20-001. Part of this research was carried out at the Methodology Department of the Faculty of Social Sciences at Tilburg University (Tilburg, the Netherlands) and EURANDOM (Eindhoven, the Netherlands).

[2] *AMS 2000 subject classifications.* Primary 62H20; secondary 62E99

Key words and phrases. correlation, covariance, test of independence, spectral decomposition, eigenvalues, eigenfunctions, Hilbert-Schmidt operator, Fréchet bounds, contingency tables, phi-square, canonical correlation, Cramér-von Mises tests, rank tests, Fredholm integral equation of the second kind.

1. Introduction We introduce a correlation coefficient ρ^* which has the potential advantage compared to the ordinary correlation ρ that it detects arbitrary forms of association between two real random variables X and Y. In fact ρ^*, to be defined below, can be viewed as a simple modification of ρ^2, as we now show. The ordinary covariance is defined as

$$\mathrm{cov}(X,Y) = E(X-EX)(Y-EY)$$

and the ordinary correlation as

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}}$$

Now suppose that Z, Z_1 and Z_2 are iid with distribution function F, that X and Y have marginal distributions F_1 and F_2, and that (X_1,Y_1), and (X_2,Y_2) are iid replications of (X,Y). Then with

$$u_F(z_1,z_2) = (z_1-EZ)(z_2-EZ) = E(z_1-Z_1)(z_2-Z_2) \tag{1.1}$$

it is easy to verify that

$$\mathrm{cov}(X,Y)^2 = Eu_{F_1}(X_1,X_2)u_{F_2}(Y_1,Y_2) \tag{1.2}$$

Now straightforward algebra based on the right hand side of (1.1) shows that we can rewrite u_F as

$$u_F(z_1,z_2) = -\tfrac{1}{2}E\Big(\,|z_1-z_2|^2 - |z_1-Z_2|^2 - |Z_1-z_2|^2 + |Z_1-Z_2|^2\,\Big)$$

Replacing the squares in u_F by absolute values then gives

$$h_F(z_1,z_2) = -\tfrac{1}{2}E\Big(\,|z_1-z_2| - |z_1-Z_2| - |Z_1-z_2| + |Z_1-Z_2|\,\Big) \tag{1.3}$$

and we can define a new 'covariance' κ by replacing u by h in (1.2):

$$\kappa(X,Y) = Eh_{F_1}(X_1,X_2)h_{F_2}(Y_1,Y_2)$$

Now we can also define

$$\rho^*(X,Y) = \frac{\kappa(X,Y)}{\sqrt{\kappa(X,X)\kappa(Y,Y)}}$$

Thus, whereas the squared covariance and ρ^2 are based on squared differences, κ and ρ^* are based on absolute differences. In this paper we demonstrate the perhaps surprising result that $0 \le \rho^*(X,Y) \le 1$, such that $\rho^*(X,Y) = 0$ iff X and Y are independent, and $\rho^*(X,Y) = 1$ iff X and Y are linearly related. A further main result we give is an orthogonal decomposition of ρ^* in terms of component correlations between eigenfunctions of h_{F_1} and h_{F_2}.

Based on their formulas, the following statistical interpretation of ρ^2 and ρ^* can be given: they measure how much two X observations which are 'far' apart tend to occur with Y observations which are 'far' apart, and similarly how much two X observations which are 'close' together tend to occur with Y observations which are 'close' together.

This paper is organized as follows. In Section 2, the properties of the kernel function h_F are investigated in detail. Some general properties are given, including conditions for its existence and a proof that it is positive, and a large part of the section is devoted to the spectral decomposition of h_F. We show that if h_F is square integrable, it has a mean square convergent spectral decomposition in terms of the eigenvalues and vectors of h_F. For discrete F, a set of difference equations is given which has this eigensystem as its solution, and for continuous F an analogous differential equation is given. The numerical solution of these equations is treated in some detail. Closed form solutions are only available in some special cases, for example, if F belongs to the uniform distribution on $[0,1]$, the eigenfunctions of h_F are the Fourier cosine functions.

The results of Section 2 are used in Section 3 to derive properties of ρ^*. We demonstrate the aforementioned result that $0 \le \rho^*(X,Y) \le 1$, such that $\rho^*(X,Y) = 0$ iff X and Y are independent, and $\rho^*(X,Y) = 1$ iff X and Y are linearly related. Furthermore, a decomposition of ρ^* is given in terms of a sum of squared correlations between marginal eigenfunctions of h_{F_1} and h_{F_2} weighted with the product of the corresponding marginal eigenvalues. We give a parameterization of the likelihood in terms of component correlations of ρ^* and the marginal eigenfunctions, somewhat analogous to the well-known canonical correlation decomposition. Fréchet bounds for the component correlations are discussed, which gives some insight into the possible structure of the dependence between two random variables. Finally in this section, component correlations for the normal distribution are discussed as an illustration.

In Section 4, we derive sample and unbiased estimators of $\kappa(X,Y)$ and related estimates of $\rho^*(X,Y)$, which can be calculated in time $O(n^2)$. The asymptotic distributions of the estimators under independence, which is a mixture of chi-squares, is derived. Finally, small sample permutation tests and Bonferroni corrections for testing the significance of component correlations are discussed.

Section 5 concerns grade versions of κ and ρ^*, copulas, and rank tests. Rank statistics, obtained from the grade versions of κ and ρ^*, are discussed. It is shown that the two sample Cramér-von Mises statistic is obtained as a special case, as well as a new generalization to the case of K ordered samples. Furthermore, it is shown that ρ^* is a weighted mean of phi-square coefficients obtained from collapsing the distribution onto a 2×2 table with respect to cut-points (x, y).

In Section 6 we propose a methodology for gaining an understanding of the association between two variables from a data set. The methodology is based on combining hypothesis tests with visual tools for displaying how much individual observations contribute to the association.

Although many of the results of the present paper are new, we have, naturally, also borrowed much from the literature, particularly concerning the eigensystems and orthogonal decompositions. Some important references here are Anderson and Darling (1952), Durbin and Knott (1972), De Wet and Venter (1973) and De Wet (1987), among others. However, the focus of much of the literature we refer to is on studying power of hypothesis tests. The aim of this paper, on the other hand, is on providing a meaningful coefficient for describing association, which we hope leads to a useful methodology for gaining an understanding of the association between two variables, and, along the way, to tests with high power against salient alternatives, the salience of the alternatives being determined by the size of ρ^*.

Throughout this paper, we use the following conventions and assumptions. We assume that (X, Y), (X_1, Y_1) and (X_2, Y_2) are iid with marginal distribution functions F_1 and F_2, respectively, and joint distribution function F_{12}. We impose no restrictions on the distributions, i.e., they may be continuous, discrete, or mixed continuous-discrete. For simplification of some of the derivations, we define distribution functions in the following slightly non-standard way:

$$F_1(x) = P(X < x) + \tfrac{1}{2}P(X = x)$$
$$F_2(y) = P(Y < y) + \tfrac{1}{2}P(Y = y)$$

and

$$F_{12}(x, y) = P(X < x, Y < y) + \tfrac{1}{2}P(X < x, Y = y) + \tfrac{1}{2}P(X = x, Y < y) + \tfrac{1}{4}P(X = x, Y = y)$$

2. Properties of the kernel function h_F

In this section a detailed description is given of the kernel function h_F defined by (1.3). Section 2.1 concerns existence, continuity, positivity, square integrability, existence of the trace and the shape of the graph of h_F. Methods for verifying whether several of these properties hold are given. In Section 2.2, under the assumption of square integrability of h_F, its spectral decomposition is given, and some properties of the associated eigenvalues and functions are derived. The eigensystem is the solution to an integral equation which may be difficult to solve. The problem is reformulated in terms of difference equations for the discrete case in Section 2.3 and in terms of a differential equation for the continuous case in Section 2.4. Both rewrites appear much easier to solve than the integral equation. In Section 2.5, efficient numerical approximation of the eigensystem of h_F for continuous F is discussed. For several well-known distributions, including the uniform and the normal, closed form solutions or numerical approximations of (parts of) the eigensystem are given. A new distribution F is introduced which has the seemingly rare property that h_F is square integrable but has infinite trace. In Section 2.6 the relation between h_F and a kernel introduced by Anderson and Darling (1952) is given. We are not aware of the kernel h_F, depending on F, having been described previously.

2.1. *Key properties of* h_F The kernel h_F *exists* if $h_F(z_1, z_2)$ is finite for some $(z_1, z_2) \in \mathbf{R}^2$. The kernel h_F is *positive* if

$$Eg(Z_1)g(Z_2)h_F(Z_1, Z_2) \geq 0 \tag{2.4}$$

for every function $g : \mathbf{R} \to \mathbf{R}$ for which the expectation exists. The kernel h_F is *square integrable* if

$$Eh_F(Z_1, Z_2)^2 = \int h_F(z_1, z_2)^2 dF(z_1)dF(z_2) \tag{2.5}$$

is finite. The kernel h_F is *trace class* if its trace

$$\mathrm{tr}(h_F) = Eh_F(Z, Z) = \int h_F(z, z)dF(z)$$

is finite. In several lemmas below we give some relatively easily verifiable conditions for checking whether these properties hold for h_F. The final Lemma 4 concerns the shape of the graph of h_F.

The next lemma may simplify verification of the existence of h_F, and asserts continuity and positivity as well as giving another integral representation of h_F. First we need the following notation:

$$\gamma(x, y) = \begin{cases} 0 & x > y \\ \frac{1}{2} & x = y \\ 1 & x < y \end{cases}$$

Note that

$$F(z) = E\gamma(Z, z) \tag{2.6}$$

$$1 - F(z) = E\gamma(z, Z) \tag{2.7}$$

Lemma 1. *If h_F exists it exists on $\mathbf{R}^2$. It is then continuous and positive, with equality in (2.4) only for the constant function, and has the representation*

$$h_F(z_1, z_2) = \int_{-\infty}^{\infty} [\gamma(z_1, w) - F(w)][\gamma(z_2, w) - F(w)]dw \qquad \forall z_1, z_2$$

Proof. We first show continuity of h_F on its domain. Let $\delta > 0$. Then if $|z_1 - z_1'| < \delta$ and $|z_2 - z_2'| < \delta$,

$$\begin{aligned} &|h_F(z_1', z_2') - h_F(z_1, z_2)| \\ &= \tfrac{1}{2}|E\left[(|z_1' - z_2'| - |z_1 - z_2|) - (|z_1' - Z_2| - |z_1 - Z_2|) - (|z_2' - Z_1| - |z_2 - Z_1|)\right]| \end{aligned}$$

$$\le \tfrac{1}{2}E\,[2\delta + \delta + \delta]$$
$$= 2\delta$$

Hence, h_F is continuous and bounded on any finite domain. Therefore, if h_F exists in one point it exists on $\mathbf{R}^2$.

We next derive the integral representation of h_F. We have

$$|z_1 - z_2| = \int_{-\infty}^{\infty} [\gamma(z_1, w) - \gamma(z_2, w)]^2 dw \tag{2.8}$$

and with $z_{i:4}$ the ith largest number in the set $\{z_1, z_2, z_3, z_4\}$, we have

$$\int_{-\infty}^{\infty} |[\gamma(z_1, w) - \gamma(z_3, w)][\gamma(z_2, w) - \gamma(z_4, w)]|\, dw =$$
$$\begin{cases} 0 & \text{if } z_1, z_3 < z_2, z_4 \text{ or } z_1, z_3 > z_2, z_4 \\ z_{3:4} - z_{2:4} & \text{otherwise} \end{cases} \tag{2.9}$$

Now we can derive the desired result first using (2.8), then applying Fubini's theorem which is justified because (2.9) is finite, and finally using (2.6):

$$\begin{aligned} h_F(z_1, z_2) &= -\tfrac{1}{2}E \int_{-\infty}^{\infty} \big([\gamma(z_1, w) - \gamma(z_2, w)]^2 - [\gamma(z_1, w) - \gamma(Z_2, w)]^2 - \\ &\qquad [\gamma(Z_1, w) - \gamma(z_2, w)]^2 + [\gamma(Z_1, w) - \gamma(Z_2, w)]^2\big)\, dw \\ &= E \int_{-\infty}^{\infty} [\gamma(z_1, w) - \gamma(Z_1, w)][\gamma(z_2, w) - \gamma(Z_2, w)]dw \\ &= \int_{-\infty}^{\infty} [\gamma(z_1, w) - F(w)][\gamma(z_2, w) - F(w)]dw \qquad (2.10) \end{aligned}$$

Finally, we show positivity of h_F. Let g be nonconstant. Then using (2.10) and Fubini's theorem,

$$\begin{aligned} &Eg(Z_1)g(Z_2)h_F(Z_1, Z_2) \\ &= E \int_{-\infty}^{\infty} g(Z_1)[\gamma(Z_1, w) - F(w)]g(Z_2)[\gamma(Z_2, w) - F(w)]dw \\ &= \int_{-\infty}^{\infty} \left(Eg(Z)[\gamma(Z, w) - F(w)]\right)^2 dw > 0 \end{aligned}$$

Hence, h_F is positive. If g is constant it is easily verified that the expression is zero. □

Note that from the lemma, it follows that for checking existence of h_F, it suffices to check existence of $h_F(0, 0)$. Now $h_F(0, 0)$ has the following convenient representations

$$\begin{aligned} h_F(0, 0) &= \tfrac{1}{2}E\,(|Z_1| + |Z_2| - |Z_1 - Z_2|) \\ &= \int_{-\infty}^{0} F(z)^2 dz + \int_{0}^{\infty} [1 - F(z)]^2 dz \qquad (2.11) \end{aligned}$$

These representations are immediately verified from (1.3) and from the representation of h_F given in Lemma 1.

An example of a random variable for which h_F does not exist is $Z = V^2$, where V has a Cauchy distribution. This can be verified by checking that (2.11) does not converge.

By giving some alternative representations of (2.5), the next lemma may be helpful in the verification of square integrability of h_F:

LEMMA 2. *We have:*

$$Eh_F(Z_1, Z_2)^2 = \frac{1}{6}E(Z_{2:4} - Z_{3:4})^2 = 2\int_{z_1<z_2} F(z_1)^2[1 - F(z_2)^2]dz_1dz_2 \tag{2.12}$$

PROOF. To prove the first equality, write $a_{ij} = |Z_i - Z_j|$. Then, since for example $Ea_{12} = Ea_{13}$ and $Ea_{34}a_{12} = Ea_{34}a_{15}$, we obtain

$$\begin{aligned}
Eh_F(Z_1, Z_2)^2 &= \frac{1}{4}E\left(a_{12} - a_{13} - a_{24} + a_{34}\right)\left(a_{12} - a_{15} - a_{26} + a_{56}\right)\\
&= \frac{1}{4}E\big(a_{12}^2 - a_{12}a_{15} - a_{12}a_{26} + a_{12}a_{56}\\
&\quad -a_{13}a_{12} + a_{13}a_{15} + a_{13}a_{26} - a_{13}a_{56}\\
&\quad -a_{24}a_{12} + a_{24}a_{15} + a_{24}a_{26} - a_{24}a_{56}\\
&\quad +a_{34}a_{12} - a_{34}a_{15} - a_{34}a_{26} + a_{34}a_{56}\big)\\
&= \frac{1}{4}E\left(a_{12}^2 - a_{12}a_{13} - a_{12}a_{24} + a_{12}a_{34}\right)\\
&= \frac{1}{16}E\left(a_{12} - a_{13} - a_{24} + a_{34}\right)^2
\end{aligned}$$

It may now be verified that $a_{12} - a_{13} - a_{24} + a_{34}$ equals 0 if $Z_1, Z_4 \le Z_2, Z_3$ or $Z_1, Z_4 \ge Z_2, Z_3$, and equals $\pm 2(Z_{2:4} - Z_{3:4})$ otherwise. Hence, and because $P(a_{12} - a_{13} - a_{24} + a_{34} \ne 0) = 2/3$, we have

$$\begin{aligned}
Eh_F(Z_1, Z_2)^2 &= \frac{1}{16}\left(a_{12} - a_{13} - a_{24} + a_{34}\right)^2\\
&= \frac{1}{16} \times \frac{2}{3} \times 4\, E(Z_{2:4} - Z_{3:4})^2 = \frac{1}{6}E(Z_{2:4} - Z_{3:4})^2
\end{aligned}$$

which is the first part of the lemma.

To prove the second equality, first note that

$$E\gamma(Z, z_1)\gamma(Z, z_2) = \min\{F(z_1), F(z_2)\} \tag{2.13}$$

We now have using Lemma 1, Fubini's theorem (for justification see proof of Lemma 1) and (2.13),

$$\begin{aligned}
Eh_F(Z_1, Z_2)^2 = E\int &[\gamma(Z_1, z_1) - F(z_1)][\gamma(Z_2, z_1) - F(z_1)]dz_1 \times\\
&\int [\gamma(Z_1, z_2) - F(z_2)][\gamma(Z_2, z_2) - F(z_2)]dz_2
\end{aligned}$$

$$
\begin{aligned}
&= \int E[\gamma(Z_1,z_1)-F(z_1)][\gamma(Z_1,z_2)-F(z_2)]\times \\
&\quad E[\gamma(Z_2,z_1)-F(z_1)][\gamma(Z_2,z_2)-F(z_2)]dz_1dz_2 \\
&= \int [\min\{F(z_1),F(z_2)\}-F(z_1)F(z_2)]^2dz_1dz_2 \\
&= 2\int_{z_1<z_2} F(z_1)^2[1-F(z_2)]^2dz_1dz_2
\end{aligned}
$$

□

An example of a distribution for which h_F exists but is not square integrable is the Cauchy distribution. In particular, $h_F(0,0)=2\pi^{-1}\log 2$, so h_F exists. In this case, nonexistence of $Eh_F(Z_1,Z_2)^2$ can most easily be verified using the right hand side of (2.12) and a computer algebra package such as provided in Mathematica.

The next lemma may be helpful in verifying whether or not h_F is trace class:

Lemma 3. *We have*

$$\mathrm{tr}(h_F)=\tfrac{1}{2}E|Z_1-Z_2|=\int F(z)[1-F(z)]dz$$

which is finite iff Z has finite mean.

Proof. The first equality follows directly from (1.3), the second is well-known and can be found using similar techniques as in the proof of the second equality in Lemma 2. As is well-known, integration by parts leads to the representation of the mean as

$$EZ=\int_0^\infty [1-F(z)]dz-\int_{-\infty}^0 F(z)dz$$

so the mean exists iff the terms on the right hand side exists. Now since

$$F(0)\int_0^\infty [1-F(z)]dz\le\int_0^\infty F(z)[1-F(z)]dz\le\int_0^\infty [1-F(z)]dz$$

and

$$[1-F(0)]\int_{-\infty}^0 [1-F(z)]dz\le\int_{-\infty}^0 F(z)[1-F(z)]dz\le\int_{-\infty}^0 F(z)dz$$

it follows that

$$\mathrm{tr}(h_F)=\int_0^\infty F(z)[1-F(z)]dz+\int_{-\infty}^0 F(z)[1-F(z)]dz$$

exists iff EZ exists. □

The quantity $E|Z_1-Z_2|$ is also called *Gini's mean difference*. Note that, by Lemmas 2 and 3, both $Eh_F(Z_1,Z_2)^2$ and $\mathrm{tr}(h_F)$ can be used as measures of dispersion.

An example of a distribution function F for which h_F is square integrable but not trace class is given in Example 2 in the next subsection.

We conclude this section with a lemma concerning the shape of the graph of h_F. In Figure 1 a representation of the graph of h_F with F the CDF of the normal distribution is given. The statements of Lemma 4 can be verified in the plot. Then:

LEMMA 4. *Suppose F is such that h_F exists. Then:*

1. *For given z_1, $h_F(z_1, z_2)$ is strictly decreasing in z_2 on the domain $\{z : z \geq z_1, F(z) < 1\}$ and strictly increasing in z_2 on the domain $\{z : z \leq z_1, F(z) > 0\}$.*
2. *$h_F(z, z)$ is strictly increasing in z on the domain $\{z : F(z) > \frac{1}{2}\}$ and strictly decreasing in z on the domain $\{z : F(z) < \frac{1}{2}\}$.*

PROOF. Part 1: With z_1, z_2, z_3 such that $z_1 < z_2 \leq z_3$ and $F(z_3) < 1$, we obtain using Lemma 1 that

$$\begin{aligned} h_F(z_1, z_3) - h_F(z_1, z_2) &= \int [\gamma(z_1, w) - F(w)][\gamma(z_3, w) - \gamma(z_2, w)]dw \\ &= -\int_{z_2}^{z_3} [1 - F(w)]dw \\ &< 0 \end{aligned}$$

where the strict inequality holds because $F(z_3) < 1$. This proves the strict decreasingness part, the strict increasingness is proven by appropriately reversing signs in the above.

Part 2: For $z < z'$ we have

$$\begin{aligned} h_F(z, z) - h_F(z', z') &= \int_{-\infty}^{\infty} [\gamma(z, w) - F(w)]^2 - [\gamma(z', w) - F(w)]^2 dw \\ &= \int_{z}^{z'} [1 - 2F(w)]dw \end{aligned}$$

which is positive if $F(z') \geq F(z) > \frac{1}{2}$ and negative if $F(z) \leq F(z') < \frac{1}{2}$, proving the monotonicity relations. □

2.2. *Spectral decomposition of h_F* A sequence of random variables $Z^{(N)}$ is said to converge to Z in mean square if

$$E\left(Z - Z^{(N)}\right)^2 \to 0 \text{ as } N \to \infty$$

For a distribution function F, we define

$$L_2(F) = \left\{ g : \mathbf{R} \to \mathbf{R} \,\middle|\, \int g(z)^2 dF(x) < \infty \right\}$$

as the set of square integrable functions with respect to F. A set of functions $\{g_k\}$ is said to be *orthonormal* if

$$Eg_k(Z)g_l(Z) = \delta_{kl} \tag{2.14}$$

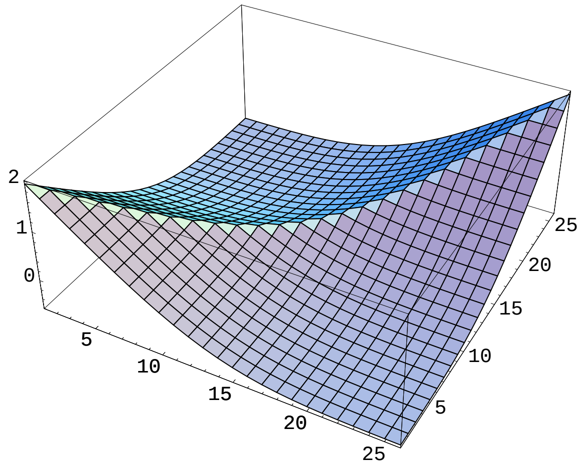


FIG. 1. *Graph of h_F with F the CDF for the standard normal distribution*

(with δ the Kronecker delta) and

$$Eg_k(Z)^2 = 1$$

An orthonormal system $\{g_k\}$ is *complete* if for any function $g \in L_2(F)$ there exist numbers $\{\alpha_k\}$ such that

$$g(Z) = \sum_{k=1}^{\infty} \alpha_k g_k(Z)$$

with convergence in mean square. Then the system $\{g_k\}$ is also called a *basis* of $L_2(F)$. A number λ is called an eigenvalue of h_F with corresponding eigenvector g if $Eg(Z)^2 = 1$ and

$$\lambda g(z) = Eh_F(z, Z)g(Z) \tag{2.15}$$

Then we have:

THEOREM 1. *Suppose h_F is square integrable. Then there exists a complete system of orthonormal functions $\{g_k\}$ of $L_2(F)$ consisting of eigenfunctions of h_F, the corresponding eigenvalues $\{\lambda_k\}$ being nonnegative. Each g_k is continuous and satisfies*

$$Eg_k(Z) = 0 \tag{2.16}$$

if $\lambda_k \neq 0$. Furthermore, h_F has the spectral decomposition

$$h_F(Z_1, Z_2) = \sum_{k=1}^{\infty} \lambda_k g_k(Z_1) g_k(Z_2) \tag{2.17}$$

where convergence is in mean square.

Proof. A continuous square integrable kernel on a sigma-finite measure space is a Hilbert-Schmidt kernel which by a generalization of Mercer's theorem has the desired spectral decomposition (Zaanen, 1960). It is easy to verify that $(\mathbf{R}, B, F)$ is a σ-finite measure space, so the operator mapping the function g to $\int h_F(x,y)g(y)dF(y)$ is a Hilbert-Schmidt operator. Nonnegativity of the eigenvalues follows from positivity of h_F (see Lemma 1). Furthermore, from (2.15), $\lambda E g_k(Z) = 0$ so $E g_k(Z) = 0$ for nonzero λ_k. □

Note that if g is a solution to (2.15), then so is $-g$. To identify the solutions, we proceed as follows. A function g is *initially positive* if there exists a z such that $g(z) > 0$ and $g(z') < 0$ for all $z' < z$. Without loss of generality, we may assume that the g_k are initially positive. We also assume the eigenvalues are ordered: $\lambda_1 > \lambda_2 > \ldots$. An interesting property of eigenfunctions of homogeneous positive Fredholm integral equations of the second kind is that they are oscillating, in the sense that for every k, g_k has k distinct zeroes and no more than $k-1$ local extremes.

Some further results are as follows:

Lemma 5. *For square integrable* h_F*:*

1. *if finite,* $\mathrm{tr}(h_F) = \sum_{k=0}^{\infty} \lambda_k$
2. $E h_F(Z_1, Z_2)^2 = \sum_{k=0}^{\infty} \lambda_k^2$.

Proof. From the spectral decomposition (2.17) and Fubini's theorem,

$$E h_F(Z, Z) = E \sum \lambda_k g_k(Z)^2 = \sum \lambda_k$$

For the second part, the desired result is obtained by Parseval's theorem. □

As an example, we consider the dichotomous case which is the simplest possible and has closed form solutions:

Example 1. *Consider the dichotomous case that* $P(Z = 0) = 1 - P(Z = 1) = p$. *Then*

$$\begin{aligned} h_F(0,0) &= (1-p)^2 \\ h_F(0,1) &= -p(1-p) \\ h_F(1,0) &= -p(1-p) \\ h_F(1,1) &= p^2 \end{aligned}$$

The eigenvalues are $(\lambda_0, \lambda_1) = (0, p(1-p))$ *and the eigenfunctions are* $g_0(z) = 1$ *and*

$$g_1(z) = \frac{1}{\sqrt{p(1-p)}}[z - (1-p)]$$

which has mean zero and variance one. Now $h_F(z_1, z_2) = \lambda_1 g_1(z_1) g_1(z_2)$.

Solutions to (2.15) for various other F are given in Table 1, see Section 2.4 for an explanation.

Equation (2.15) is a homogeneous Fredholm integral equation of the second kind based on the degenerate kernel h_F (see, for example, Tricomi, 1985). In general these equations are difficult to solve. In Sections 2.3 and 2.4, (2.15) is reduced to a simpler problem for the discrete and continuous case, respectively. We conclude this section with an alternative formula to (2.15) for finding the eigensystem which we use in the next two subsections. With

$$G(z) = \int_{-\infty}^{z} g(y)dF(y),$$

we have:

LEMMA 6. *The nonzero eigenvalues and eigenvectors of h_F are solutions to the equation*

$$\lambda\,[g(z) - g(z')] + \int_{z}^{z'} G(w)dw = 0$$

for all $z < z'$.

PROOF. Substitution of the expression for h_F of Lemma 1 into (2.15) yields

$$\begin{aligned}\lambda g(z) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[\gamma(z,w) - F(w)][\gamma(z_2,w) - F(w)]g(z_2)dF(z_2)dw \\ &= \int_{-\infty}^{\infty}[\gamma(z,w) - F(w)]G(w)dw \\ &= -\int_{-\infty}^{z} F(w)G(w)dw + \int_{z}^{\infty}[1 - F(w)]G(w)dw\end{aligned}$$

The Lemma follows from this and since by (2.16) $G(z) \to 0$ as $|z| \to \infty$. □

We have the following interesting implication:

COROLLARY 1. *If $\langle a, b\rangle$ is an interval with zero probability mass, i.e., $F(a) = F(b)$, then a solution $g(z)$ to (2.15) is linear on $\langle a, b\rangle$.*

PROOF. If $F(a) = F(b)$ then from its definition it follows that G is constant on $\langle a, b\rangle$. Therefore, for any $z, z' \in \langle a, b\rangle$, we obtain from Lemma 6 that

$$\lambda\,[g(z) - g(z')] = c(z - z')$$

for some constant c. Hence, g is linear on $\langle a, b\rangle$. □

Note that for discrete distributions, it follows that the eigenfunctions g are piecewise linear.

2.3. *Obtaining the eigensystem in the discrete case* We consider the case that Z is a.s. in a finite set, say $P(Z \in \{z_1, z_2, \ldots, z_K\}) = 1$. We use the shorthand $P(Z = z_i) = p_i$ and assume without loss of generality that $p_i > 0$ and $z_i < z_{i+1}$ for all i. Then we obtain:

LEMMA 7. *With* $c_i = (z_i - z_{i-1})^{-1}$ *the nonzero eigenvalues and eigenvectors of* h_F *are solutions to the equations*

$$p_1 g(z_1) = \lambda c_2[g(z_1) - g(z_2)]$$

$$p_K g(z_K) = \lambda c_K[g(z_{K-1}) - g(z_K)]$$

$$p_i g(z_i) = \lambda \left[c_i g(z_{i-1}) - (c_i + c_{i+1})g(z_i) + c_{i+1} g(z_{i+1})\right] \qquad 2 \le i \le K-1$$

PROOF. First note that from the definition for $z \in \langle z_i, z_{i+1}\rangle$,

$$G(z) = \sum_{j=1}^{i} p_j g(z_j)$$

It follows that

$$\int_{z_i}^{z_{i+1}} G(w)dw = (z_{i+1} - z_i)\sum_{j=1}^{i} p_j g(z_j) \tag{2.18}$$

The first two displayed equations of the lemma now follow from (2.18), Lemma 6 and from $\sum p_i g(z_i) = 0$. From (2.18) we further obtain

$$c_{i+1}\int_{z_i}^{z_{i+1}} G(w)dw - c_i \int_{z_{i-1}}^{z_i} G(w)dw = p_i g(z_i) \tag{2.19}$$

Again from Lemma 6, we have

$$\lambda\left[g(z_i) - g(z_{i+1})\right] + \int_{z_i}^{z_{i+1}} G(w)dw = 0$$

$$\lambda\left[g(z_{i-1}) - g(z_i)\right] + \int_{z_{i-1}}^{z_i} G(w)dw = 0$$

Multiplying these equations by c_{i+1} and c_i respectively, taking differences and the use of 2.19 yields

$$p_i g_k(z_i) = \lambda\left[c_i g_k(z_{i-1}) - (c_i + c_{i+1})g_k(z_i) + c_{i+1} g_k(z_{i+1})\right]$$

which completes the proof. □

More details on difference equations of the form given in Lemma 7 can be found in Agarwal (1992). Lemma 7 allows fast and memory efficient computation of the eigenvalues and vectors. In matrix notation, we must solve the generalized eigenvalue problem

$$D_p g = \lambda C g$$

where g is the eigenvector with corresponding eigenvalue λ, D_p is a diagonal matrix with the coordinates of p on the main diagonal, and

$$C = \begin{pmatrix} c_2 & c_2 & 0 & 0 & & \\ c_2 & -(c_2+c_3) & c_3 & 0 & & \cdots \\ 0 & c_3 & -(c_3+c_4) & c_4 & & \\ 0 & 0 & c_4 & -(c_4+c_5) & & \\ & & & & \ddots & \\ & \vdots & & & & -(c_{K-1}+c_K) & c_K \\ & & & & & c_K & c_K \end{pmatrix}$$

This method can also be used for fast approximation of continuous systems (see Section 2.5). The exact solution for continuous systems is given in the next subsection.

2.4. *Obtaining the eigensystem in the continuous case* If F is differentiable, the problem of finding the eigenvalues and vectors can be reduced to a differential equation which is sometimes easier to solve:

LEMMA 8. *Suppose f is the derivative of F. Then the nonzero eigenvalues and eigenvectors of h_F are solutions to the equation*

$$\lambda g''(z) + f(z)g(z) = 0 \tag{2.20}$$

subject to the side condition

$$g'(z) \to 0 \qquad \text{as } |z| \to \infty$$

PROOF. Letting $z' \to z$ in Lemma 6 we obtain

$$\lambda g'(z) + G(z) = 0$$

Differentiating both sides with respect to z yields (2.20). Finally, since by (2.16) $G(z) \to 0$ as $|z| \to \infty$, the side condition follows. □

Equation (2.20) leads to an interesting observation on the behavior of the eigenfunctions g_k: the second derivative $g_k''(x)$ is proportional to the local density $f(x)$ times $g_k(x)$. As mentioned in Section 2.2 and as follows from the theory of Sturm-Liouville differential equations, the eigenfunctions oscillate. Now if the local density is low, the oscillatory behavior will be slower. In particular, on intervals where the local density is zero the second derivative of an eigenfunction is zero so the eigenfunction is linear. (NB: this also holds for non-continuous distributions, see Corollary 1).

If F is both differentiable and invertible, we can reformulate differential equation (2.20) in the standard Sturm-Liouville form:

LEMMA 9. *Suppose F is invertible. Let $q(u) = g[F^{-1}(u)]$ and suppose q is twice differentiable. Then the eigenvalues and eigenvectors are solutions to the equation*

$$\frac{d}{du} f[F^{-1}(u)]q'(u) + \lambda^{-1} q(u) = 0 \tag{2.21}$$

subject to the side condition

$$f[F^{-1}(u)]q'(u) \to 0 \text{ as } u \downarrow 0 \text{ or } u \uparrow 1$$

PROOF. Note that

$$\frac{d}{du} F^{-1}(u) = \frac{1}{f[F^{-1}(u)]}$$

so that

$$f[F^{-1}(u)]q'(u) = f[F^{-1}(u)] \frac{d}{du} g[F^{-1}(u)] = g'[F^{-1}(u)]$$

From this the side condition follows. Substituting $z = F^{-1}(u)$ into the left hand side of (2.20) yields

$$\begin{aligned} &\lambda g''[F^{-1}(u)] - f(F^{-1}(u))g[F^{-1}(u)] \\ &= \lambda \frac{du}{dF^{-1}(u)} \frac{d}{du} g'[F^{-1}(u)] - f(F^{-1}(u))g[F^{-1}(u)] \\ &= \lambda f[F^{-1}(u)] \frac{d}{du} f[F^{-1}(u)]q'(u) - f[F^{-1}(u)]q(u) \end{aligned}$$

Hence dividing both sides of (2.20) by $\lambda f[F^{-1}(u)]$ yields the desired result. □

Note that for the q_k the orthonormality condition (2.14) reduces to

$$\int q_k(u) q_l(u) du = \delta_{kl}$$

In general, the differential equations (2.20) and (2.21) do not have closed form solutions (i.e., solutions in terms of well-known functions). We obtained the solutions for the uniform, the logistic and the exponential distributions which are given in Table 1, where P_k is the kth Legendre polynomial, J_i is the ith Bessel function of the first kind, α_k the kth zero of J_1 and

$$\beta_k = \left(J_0(\alpha_k)^2 + J_1(\alpha_k)^2\right)^{-1/2}$$

Numerical solutions were obtained for the normal, Laplace and chi-square distribution with one degree of freedom, see the next subsection for details on obtaining these solutions. For the normal and Laplace distributions we obtained the exact value of $\sum \lambda_k$ and for the normal distribution of $\sum \lambda_k^2$ using Lemmas 2 and 3 combined with results for order statistics by Bose and Gupta (1959) and Govindarajulu (1963) summarized in Johnson, Kotz, and Balakrishnan (1994) and using Mathematica.

We also obtained a closed form expression for the eigensystem for the distribution introduced in the next example. It is an example of a distribution F for which h_F is square integrable but not trace class, i.e., by Lemma 5, $\sum \lambda_k^2$ is finite but $\sum \lambda_k = \infty$. We are not aware of any previous studies of this distribution.

EXAMPLE 2. *With V standard normally distributed, let Z be defined as the following function of V:*

$$Z = \sqrt{\frac{\pi}{2}} \int_0^V \exp(t^2/2) dt$$

where the convention is used that for $a > b$,

$$\int_a^b f(t)dt = -\int_b^a f(t)dt$$

Close to zero, Z has approximately a normal density, but far from zero the density is much lower, that is, Z has much heavier tails than a standard normal.

With F the distribution function of Z, we now show that h_F is square integrable but not trace class. To show this, we shall derive an expression for $f[F^{-1}(u)]$ to be plugged into Equation (2.21) so that it can be solved. The derivation involves the so-called imaginary error function. The ordinary error function *is defined as*

$$\mathrm{erf}(z) = \sqrt{2\pi} \int_0^z \exp(-t^2) dt$$

Note that the CDF of the standard normal distribution is $\Phi(v) = (1 + \mathrm{erf}(v/\sqrt{2}))/2$. The imaginary error function *is defined as*

$$\mathrm{erfi}(z) = -i\, \mathrm{erf}(i\, z)$$

(Here $i = \sqrt{-1}$. See Weisstein, 1999, for some of the properties of erf *and* erfi*.) We define the inverse $\mathrm{erfi}^{-1}(z)$ as the unique real y satisfying $z = \mathrm{erfi}(y)$.*

We see that

$$Z = \pi\, \mathrm{erfi}\left(\frac{V}{\sqrt{2}}\right)$$

Now since V has a standard normal distribution we obtain for the CDF of Z:

$$F(z) = P(Z < z) = P\left(V < \sqrt{2}\, \mathrm{erfi}^{-1}(z/\pi)\right) = \Phi\left(\sqrt{2}\, \mathrm{erfi}^{-1}(z/\pi)\right)$$

From this,

$$F^{-1}(u) = \pi\, \mathrm{erfi}\left(\Phi^{-1}(u)/\sqrt{2}\right)$$

Some tedious but straightforward algebra then gives

$$f(F^{-1}(u)) = 1/\phi[\Phi^{-1}(u)]^2$$

Plugging this expression into (2.21) leads to the solution

$$\lambda_k = 1/k$$
$$q_k(u) \equiv H_k\left[\Phi^{-1}(u)\right] \sqrt{\phi\left[\Phi^{-1}(u)\right]}$$

where H_k is the kth Hermite polynomial. The '$\equiv$' symbol is used to indicate that the expression for q_k needs to be suitably normalized. This solution of (2.21) was

Distribution	Logistic	Uniform	Normal
$1/f[F^{-1}(u)]$	$u(1-u)$	1	$\phi[\Phi^{-1}(u)]$
λ_k	$\frac{1}{k(k+1)}$	$\frac{1}{k^2\pi^2}$	[a]
$\sum \lambda_k$	1	$\frac{1}{6}$	$\frac{1}{\sqrt{\pi}}$
$\sum \lambda_k^2$	$\frac{1}{3}(\pi^2-9)$	$\frac{1}{90}$	$\frac{1}{3}-\frac{\sqrt{3}-1}{\pi}$
$q_k(u)$	$\sqrt{2k+1}\,P_k(2u-1)$	$2\cos(k\pi u)$	[a]
$\lambda_1/\sum\lambda_k$	0.5000	0.6079	0.5269
$\lambda_2/\sum\lambda_k$	0.1667	0.1520	0.1635
$\lambda_3/\sum\lambda_k$	0.0833	0.0675	0.0795
$\lambda_4/\sum\lambda_k$	0.0500	0.0380	0.0470

Distribution	Exponential	Laplace	Chi-square	Example 2
$1/f[F^{-1}(u)]$	$1-u$	$\min(u,1-u)$	$\frac{\phi[\Phi^{-1}(\frac{u+1}{2})]}{\Phi^{-1}(\frac{u+1}{2})}$	$\phi[\Phi^{-1}(u)]^2$
λ_k	$\frac{4}{\alpha_k^2}$	[a]	[a]	$\frac{1}{k}$
$\sum \lambda_k$	$\frac{1}{2}$	$\frac{3}{4}$	0.6360	∞
$\sum \lambda_k^2$	$\frac{1}{12}$	0.1458	0.1399	$\frac{6}{\pi^2}$
$q_k(u)$	$\beta_k J_0(\alpha_k\sqrt{1-u})$	[a]	[a]	$H_k[\Phi^{-1}(u)]\sqrt{\phi[\Phi^{-1}(u)]}$ [b]
$\lambda_1/\sum\lambda_k$	0.5453	0.4611	0.5567	0
$\lambda_2/\sum\lambda_k$	0.1627	0.1816	0.1615	0
$\lambda_3/\sum\lambda_k$	0.0774	0.0875	0.0758	0
$\lambda_4/\sum\lambda_k$	0.0451	0.0542	0.0438	0

TABLE 1

Eigenvalues and eigenvectors of kernel function h_F for various F.

[a] *No closed form expression is available.*

[b] *Expression not normalized.*

derived by De Wet and Venter (1973), who provided a method for solving differential equations of the form $\frac{d}{du}w(u)q_k'(u)+\lambda_k^{-1}q_k(u)=0$ for certain types of weights $w(u)$, including $w(u)=1/\phi[\Phi^{-1}(u)]^2$. Note that for g_k we obtain

$$g_k(x) \equiv H_k\left[\sqrt{2}\,\mathrm{erfi}^{-1}(\pi\, z)\right]\sqrt{\phi\left[\sqrt{2}\,\mathrm{erfi}^{-1}(\pi\, z)\right]}$$

It is well-known that here $\sum\lambda_k$ is divergent and $\sum\lambda_k^2=\pi^2/6$. Note that, by Lemmas 3 and 5, divergence of $\sum\lambda_k$ implies that Z does not have a finite mean.

The differential equation (2.21) with $f[F^{-1}(z)]$ replaced by a weight function $w(z)$ is given in Anderson and Darling (1952), see also De Wet and Venter (1973) and De Wet (1987). We are not aware of equation (2.20) having been studied previously.

2.5. *Discrete approximation of the continuous case* For many continuous distribution functions F the differential equations (2.20) and (2.21) do not have a closed form solution and the first t eigenvalues and eigenfunctions can be approximated by using a discrete approximation of F and solving the difference equations

	True value	Estimate ($t = 101$)	Estimate ($t = 1001$)
$\sum \lambda_k$	1	0.99303	0.99931
$\sum \lambda_k^2$	0.28987	0.29027	0.28988
λ_1^*	0.50000	0.50370	0.50035
λ_{10}^*	9.0909×10^{-3}	9.3093×10^{-3}	9.1056×10^{-3}
λ_{100}^*	9.9010×10^{-5}	9.9708×10^{-5}	9.9145×10^{-5}
λ_{1000}^*	9.9900×10^{-7}	-	9.9970×10^{-7}

TABLE 2

Eigenvalues and their estimates based on discrete approximations for the kernel h_F with F the logistic distribution function.

of Lemma 7. For $i = 1, \ldots, t$, let $c_i = F^{-1}\left(\frac{i-1/2}{t}\right)$ and let $Z^{(t)}$ be a discrete random variable with $P(Z^{(t)} = c_i) = p_i = F(i/t) - F((i-1)/t)$. The eigenvalues and eigenvectors of $h_{F^{(t)}}$, with $F^{(t)}$ the distribution function of $Z^{(t)}$, can then be calculated using the method of Section 2.3. For large t, this method seems to give good approximations of the eigenvalues and eigenvectors of h_F. An idea of the quality of the approximations can be gained from Table 2. The numerical results in Table 1 were obtained using this method. For further details on discrete approximations of eigenvalues and vectors of kernels see Tricomi (1985).

To obtain a good approximation, t should of course be chosen as large as possible. Using Mathematica 5.2 on a Pentium IV computer at 3.0MHz, using no special routines for calculating the eigensytem of tridiagonal matrices, calculation of a complete solution for $t = 1000$ took 29 seconds. We expect that using software with such special routines, it is possible to obtain solutions of the first few eigenvalues and eigenvectors much more quickly and for much larger t.

For calculation of the eigensystem from a sample, see Section 4.1.

2.6. *Relation to Anderson-Darling kernel* A related kernel was studied by Anderson and Darling (1952) and De Wet and Venter (1973), among others, in the context of Cramér-von Mises tests. With w a nonnegative weight function, they considered the kernel

$$r_w(u,v) = \int_0^1 [\gamma(u,t) - t][\gamma(v,t) - t]w(t)dt$$

The kernel r_w is closely related to the kernel h_F: with $w(t) = \frac{1}{f[F^{-1}(t)]}$, we obtain

$$r_w(u,v) = h_F[F^{-1}(u), F^{-1}(v)]$$

Note that this conversion does not work for discrete F. The eigensystem for r_w is given by the set of solutions to (2.21) with $f[F^{-1}(z)]$ replaced by the weight function $w(z)$ (Anderson & Darling, 1952).

Other closely related kernels have been given in the context of two-sample tests by Zech and Aslan (2003) and Baringhaus and Franz (2004). (See Section 5.2 for the relation between two-sample tests and tests of independence.)

3. Properties of κ and ρ^* We now apply the results of Section 2 in order to derive properties of κ and ρ^*. In Section 3.1, some key properties are derived, including that $0 \le \rho^*(X,Y) \le 1$, with $\rho^*(X,Y) = 0$ iff $X \perp\!\!\!\perp Y$ and $\rho^*(X,Y) = 1$ iff X and Y are linearly related. In Section 3.2 we give a decomposition of $\kappa(X,Y)$ and $\rho^*(X,Y)$ as weighted sums of squared correlations between the marginal eigenfunctions of h_{F_1} and h_{F_2}, weighted by functions of the eigenvalues. In Section 3.3, we describe a decomposition of the likelihood in terms of marginal eigenfunctions and component correlations of ρ^*. In Section 3.4, Fréchet bound for the component correlations are given, which gives some insight into their meaning.

3.1. *Key properties of κ and ρ^** Some key properties of κ and of ρ^*, are given in the following theorem:

THEOREM 2. *Suppose X and Y and Z are real random variables for which the marginal kernels h_{F_1} and h_{F_2} exist. Then:*

1. *If a, b, c and d are constants, then $\kappa(aX+b, cY+d) = a\,c\,\kappa(X,Y)$ and $\rho^*(aX+b, cY+d) = \rho^*(X,Y)$.*
2. *$\kappa(X,Y) \ge 0$ with equality iff $X \perp\!\!\!\perp Y$.*
3. *If $\kappa(X,X) < \infty$ and $\kappa(Y,Y) < \infty$ then $\kappa(X,Y) \le \sqrt{\kappa(X,X)\kappa(Y,Y)}$ with equality iff X and Y are a.s. linearly related.*
4. *If both X and Y are dichotomous then $\kappa(X,Y) = \operatorname{cov}(X,Y)^2$ and $\rho^*(X,Y) = \rho(X,Y)^2$*
5. *With $Z_{i:n}$ the ith order statistic in a sample of size n, $\kappa(Z,Z) = \frac{1}{6}E(Z_{2:4} - Z_{3:4})^2$.*

The proof of the theorem is given at the end of this section. Note that $\kappa(X,X) = Eh_{F_1}(X_1,X_2)^2$ and $\kappa(Y,Y) = Eh_{F_2}(Y_1,Y_2)^2$ so the condition in Part 3 is equivalent to square integrability of the marginal kernels. From Theorem 2, Parts 2 and 3, we immediately have the following:

COROLLARY 2. *Suppose $\kappa(X,X) < \infty$ and $\kappa(Y,Y) < \infty$. Then $0 \le \rho^*(X,Y) \le 1$, with $\rho^*(X,Y) = 0$ iff $X \perp\!\!\!\perp Y$ and $\rho^*(X,Y) = 1$ iff X and Y are a.s. linearly related.*

We may compare Corollary 2 to the related well-known result for the ordinary correlation ρ: If $\operatorname{var}(X) < \infty$ and $\operatorname{var}(Y) < \infty$ then $0 \le \rho^2 \le 1$ with $\rho^2 = 0$ if $X \perp\!\!\!\perp Y$ and $\rho^2 = 1$ iff X and Y are a.s. linearly related. The important difference is that $X \perp\!\!\!\perp Y$ is equivalent to $\rho^* = 0$ but $X \perp\!\!\!\perp Y$ only implies $\rho = 0$, not vice versa. By Part 5 of Theorem 2, $\kappa(Z,Z)$ can be used as a measure of dispersion for a real random variable Z. Note the relation with the variance, which can be defined as $E(Z_{1:2} - Z_{2:2})^2$.

Note that, even though $\rho^*(X,Y) = 1$ iff X and Y are linearly related, ρ^* is not a measure of linear association in the following sense: if the slope of the linear regression line of Y given X is zero, $\rho^*(X,Y)$ need not be equal to zero.

From Lemma 2 and Lemma 3, a sufficient condition for $\rho^*(X,Y)$ to exist is that EX and EY exist. An example showing that this condition is not necessary is Example 2. Note that the existence of the ordinary correlation ρ has the much stronger requirement of finite marginal variances. Summarizing, we have

$$\begin{aligned}\rho(X,Y) \text{ exists } &\Leftrightarrow \left\{\sigma^2(X) \text{ and } \sigma^2(Y) \text{ exist }\right\} \Rightarrow \{EX \text{ and } EY \text{ exist }\} \Rightarrow \\ &\Rightarrow \left\{E(X_{2:4}-X_{3:4})^2 \text{ and } E(Y_{2:4}-Y_{3:4})^2 \text{ exist }\right\} \Leftrightarrow \rho^*(X,Y) \text{ exists}\end{aligned}$$

where the one-way implications are strict.

We now proceed to the proof of Theorem 2:

Proof of Theorem 2. Part 1 follows directly from the definition.

Part 2: From Lemma 1 we obtain

$$h_{F_i}(a,b) = \int [\gamma(a,w) - F_i(w)][\gamma(b,w) - F_i(w)]dw$$

Furthermore, note that

$$F_{12}(x,y) - F_1(x)F_2(y) = E[\gamma(X,x) - F_1(x)][\gamma(Y,y) - F_2(y)]$$

which is easy to verify. Using these results and the finiteness of each side of (2.9) which allows us to apply Fubini's theorem, we obtain

$$\begin{aligned}\kappa(X,Y) &= Eh_{F_1}(X_1,X_2)h_{F_2}(Y_1,Y_2)\\ &= E\int [\gamma(X_1,x) - F_1(x)][\gamma(X_2,x) - F_1(x)]dx \times \\ &\qquad \int [\gamma(Y_1,y) - F_2(y)][\gamma(Y_2,y) - F_2(y)]dy\\ &= \int E[\gamma(X_1,x) - F_1(x)][\gamma(Y_1,y) - F_2(y)] \times \\ &\qquad E[\gamma(X_2,x) - F_1(x)][\gamma(Y_2,y) - F_2(y)]dxdy\\ &= \int [F_{12}(x,y) - F_1(x)F_2(y)]^2 dxdy \qquad (3.22)\end{aligned}$$

If $X \perp\!\!\!\perp Y$, the integrand is zero so $\kappa(X,Y) = 0$. It remains to be shown that $X \not\!\perp\!\!\!\perp Y$ implies $\kappa(X,Y) \neq 0$. We next sketch the proof.

If $X \not\!\perp\!\!\!\perp Y$ then there is an (a,b) such that $D(a,b) = F_{12}(a,b) - F_1(a)F_2(b) \neq 0$. We now show that if $D(a,b) \neq 0$, then there is an open interval, which has (a,b) as a limit point, such that $D \neq 0$ on that interval. It then follows that (3.22) is nonzero. Let $\varepsilon_{12} \geq 0$ be the probability mass in (a,b), $\varepsilon_1 \geq 0$ be the marginal probability mass in a and $\varepsilon_2 \geq 0$ be the marginal probability mass in b. Denote by $D(a^\pm, b^\pm)$

the limit approaching from anywhere in one of four open ‘quadrants’ defined by (a, b). Then from the definition of F_{12}, F_1 and F_2 given in the introduction,

$$\begin{aligned}
D(a^-, b^-) &= D(a,b) - \tfrac{1}{4}\varepsilon_{12} + \tfrac{1}{2}\varepsilon_1 F_2(b) + \tfrac{1}{2}\varepsilon_2 F_1(a) - \tfrac{1}{4}\varepsilon_1\varepsilon_2 \\
D(a^-, b^+) &= D(a,b) + \tfrac{1}{4}\varepsilon_{12} + \tfrac{1}{2}\varepsilon_1 F_2(b) - \tfrac{1}{2}\varepsilon_2 F_1(a) + \tfrac{1}{4}\varepsilon_1\varepsilon_2 \\
D(a^+, b^-) &= D(a,b) + \tfrac{1}{4}\varepsilon_{12} - \tfrac{1}{2}\varepsilon_1 F_2(b) + \tfrac{1}{2}\varepsilon_2 F_1(a) + \tfrac{1}{4}\varepsilon_1\varepsilon_2 \\
D(a^+, b^+) &= D(a,b) + \tfrac{3}{4}\varepsilon_{12} - \tfrac{1}{2}\varepsilon_1 F_2(b) - \tfrac{1}{2}\varepsilon_2 F_1(a) - \tfrac{1}{4}\varepsilon_1\varepsilon_2
\end{aligned}$$

Now if $D(a, b) \neq 0$ these four expressions cannot all be zero, so there must be an open set, in at least one of the four ‘quadrants’ and with (a, b) as a limit point, where D is nonzero. Hence, (3.22) cannot be zero.

Part 3: By definition $\kappa(X,Y)^2 \leq \kappa(X,X)\kappa(Y,Y)$ is equivalent to

$$[Eh_{F_1}(X_1,X_2)h_{F_2}(Y_1,Y_2)]^2 \leq Eh_{F_1}(X_1,X_2)^2 Eh_{F_2}(Y_1,Y_2)^2$$

This is a Cauchy-Schwartz inequality so it holds, and equality holds iff

$$h_{F_1}(X_1,X_2) \overset{\text{a.s.}}{=} c\, h_{F_2}(Y_1,Y_2) \tag{3.23}$$

for some constant c. If $Y \overset{\text{a.s.}}{=} aX + b$ for certain constants a and b then it is immediately verified that (3.23) holds with $c = |a|$.

The reverse implication that (3.23) implies linearity remains to be shown. Suppose that (3.23) holds. Then

$$\begin{aligned}
&h_{F_1}(X_1,X_2) - h_{F_1}(X_1,X_3) - h_{F_1}(X_2,X_4) - h_{F_1}(X_3,X_4) \overset{\text{a.s.}}{=} \\
&c\,[h_{F_2}(Y_1,Y_2) - h_{F_2}(Y_1,Y_3) - h_{F_2}(Y_2,Y_4) - h_{F_2}(X_3,X_4)]
\end{aligned}$$

which reduces to

$$\begin{aligned}
&|X_1 - X_2| - |X_1 - X_3| - |X_2 - X_4| + |X_3 - X_4| \overset{\text{a.s.}}{=} \\
&c\,(|Y_1 - Y_2| - |Y_1 - Y_3| - |Y_2 - Y_4| + |Y_3 - Y_4|)
\end{aligned}$$

But this is equivalent to

$$X_{3:4} - X_{2:4} \overset{\text{a.s.}}{=} c\,(Y_{3:4} - Y_{2:4}) \tag{3.24}$$

Now without loss of generality suppose $Y_{3:4} = cX_{3:4} + b$ and $Y_{2:4} = cX_{2:4} + b'$ for some b and b'. Substitution into (3.24) then yields $b = b'$, so the second and third order statistics for X and Y are linearly related. Now the distribution function of the second order statistic for X is

$$F_{1;2:4}(x) = \int_{-\infty}^{x} F_1(t)[1 - F_1(t)]^2 dF_1(t)$$

and for Y

$$F_{2;2:4}(y) = \int_{-\infty}^{y} F_2(t)[1 - F_2(t)]^2 dF_2(t)$$

It is now straightforward to show that the equation $F_{2;2:4}(y) = F_{1;2:4}(cx + b)$ leads to $F_2(y) = F_1(cx + b)$, so X and Y are linearly related.

Part 4: Without loss of generality assume $X \in \{0,1\}$ and $Y \in \{0,1\}$ with probability one. Then $h_{F_1} = u_{F_1}$ and $h_{F_2} = u_{F_2 1}$ (both u and h defined in Section 1), so $\kappa(X,Y) = Eh_{F_1}(X_1,X_2)h_{F_2}(Y_1,Y_2) = Eu_{F_1}(X_1,X_2)u_{F_2}(Y_1,Y_2) = \mathrm{cov}(X,Y)$.

Part 5: Since $\kappa(Z,Z) = Eh_F(Z_1,Z_2)^2$, this follows directly from Lemma 2 □

We conclude this section by giving some representations of κ in terms of h_F and the (conditional) distribution functions of X and Y. Let

$$F_{2|1}(y|x) = P(Y < y|X = x) + \tfrac{1}{2}P(Y = y|X = x)$$

be the conditional distribution function of Y given $X = x$.

LEMMA 10. *The following equalities hold for κ:*

1. $\kappa(X,Y) = \int [F_{12}(x,y) - F_1(x)F_2(y)]^2 dxdy$
2. $\kappa(X,Y) = Eh_{F_1}(X_1,X_2) \int [F_{2|1}(y|X_1) - F_2(y)][F_{2|1}(y|X_2) - F_2(y)]dy$

PROOF. Part 1: This follows from the proof of Theorem 2, Part 2

Part 2: First note that $d_{xy}F_{12}(x,y) = d_xF_1(x)d_yF_{2|1}(y|x)$ which we write in shorthand $dF_{12}(x,y) = dF_1(x)dF_{2|1}(y|x)$. Hence,

$$\begin{aligned}
\kappa(X,Y) &= Eh_{F_1}(X_1,X_2)h_{F_2}(Y_1,Y_2) \\
&= \tfrac{1}{2}\int h_{F_1}(x_1,x_2)\int [\gamma(y_1,y) - F_2(y)][\gamma(y_2,y) - F_2(y)]dydF_{12}(x_1,y_1)dF_{12}(x_2,y_2) \\
&= \int h_{F_1}(x_1,x_2)\int [\gamma(y_1,y) - F_2(y)]dF_{2|1}(y_1|x_1)[\gamma(y_2,y) - F_2(y)]dF_{2|1}(y_2|x_2)dydF_1(x_1)dF_1(x_2) \\
&= \int h_{F_1}(x_1,x_2)\int [F_{2|1}(y|x_1) - F_2(y)][F_{2|1}(y|x_2) - F_2(y)]dydF_1(x_1)dF_1(x_2) \\
&= Eh_{F_1}(X_1,X_2)\int [F_{2|1}(y|X_1) - F_2(y)][F_{2|1}(y|X_2) - F_2(y)]dy
\end{aligned}$$

□

Note the similarity of Part 1 of Lemma 10 with the formula for the covariance given by Hoeffding (1940):

$$\mathrm{cov}(X,Y) = \int [F_{12}(x,y) - F_1(x)F_2(y)]dxdy$$

3.2. *Orthogonal decomposition* Let us assume h_{F_1} and h_{F_2} are square integrable and have the spectral decompositions

$$h_{F_1}(x_1,x_2) = \sum_{k=0}^{\infty} \lambda_k g_{1k}(x_1)g_{1k}(x_2) \tag{3.25}$$

$$h_{F_2}(y_1,y_2) = \sum_{k=0}^{\infty} \mu_k g_{2k}(y_1)g_{2k}(y_2) \tag{3.26}$$

See Lemma 2 on how to check for square integrability. For ease of notation, we write the correlations between marginal eigenfunctions as

$$\rho_{kl}(X,Y) = \rho[g_{1k}(X), g_{2l}(Y)]$$

We now have the orthogonal decomposition given as follows:

THEOREM 3. *Suppose h_{F_1} and h_{F_2} are square integrable with spectral decompositions as above. Then with convergence in mean square,*

$$\kappa(X,Y) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \lambda_k \, \mu_l \, \rho_{kl}(X,Y)^2$$

and

$$\rho^*(X,Y) = \frac{1}{\sqrt{\sum \lambda_k^2}\sqrt{\sum \mu_l^2}} \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \lambda_k \, \mu_l \, \rho_{kl}(X,Y)^2$$

PROOF. Write

$$\kappa^{(N,N)}(X,Y) = E\left[\sum_{k=1}^{N} \lambda_k g_{1k}(X_1) g_{1k}(X_2) \sum_{l=1}^{N} \mu_l g_{2l}(Y_1) g_{2l}(Y_2)\right]$$

$$\kappa^{(.,N)}(X,Y) = E\left[h_{F_1}(X_1,X_2) \sum_{l=1}^{N} \mu_l g_{2l}(Y_1) g_{2l}(Y_2)\right]$$

$$\kappa^{(N,.)}(X,Y) = E\left[\sum_{k=1}^{N} \lambda_k g_{1k}(X_1) g_{1k}(X_2) \, h_{F_2}(Y_1,Y_2)\right]$$

Then straightforward algebra gives

$$\kappa^{(N,N)}(X,Y) = \sum_{k=1}^{N}\sum_{l=1}^{N} \lambda_k \, \mu_l \, \rho[g_{1k}(X), g_{2l}(Y)]^2$$

By the Cauchy-Schwartz inequality we obtain

$$\Big(\kappa(X,Y) - \kappa^{(N,.)}(X,Y) - \kappa^{(.,N)}(X,Y) + \kappa^{(N,N)}(X,Y)\Big)^2$$

$$= E\left(\left[h_{F_1}(X_1,X_2) - \sum_{k=1}^{N} \lambda_k g_{1k}(X_1) g_{1k}(X_2)\right]\left[h_{F_2}(Y_1,Y_2) - \sum_{l=1}^{N} \mu_k g_{2l}(Y_1) g_{2l}(Y_2)\right]\right)^2$$

$$\leq E\left[h_{F_1}(X_1,X_2) - \sum_{k=1}^{N} \lambda_k g_{1k}(X_1) g_{1k}(X_2)\right]^2 E\left[h_{F_2}(Y_1,Y_2) - \sum_{l=1}^{N} \mu_k g_{2l}(Y_1) g_{2l}(Y_2)\right]^2$$

By mean square convergence of the spectral decomposition the latter goes to zero as $N \to \infty$ so

$$\kappa(X,Y) - \kappa^{(N,.)}(X,Y) - \kappa^{(.,N)}(X,Y) + \kappa^{(N,N)}(X,Y) \to 0$$

as $N \to \infty$. Similarly we find

$$\kappa(X,Y) - \kappa^{(N,.)}(X,Y) \to 0$$
$$\kappa(X,Y) - \kappa^{(.,N)}(X,Y) \to 0$$

as $N \to \infty$. It follows that

$$\kappa^{(N,N)} \to \kappa(X,Y)$$

as $N \to \infty$, which is the desired result. □

The simplest example of a decomposition is if both variables are dichotomous, say $P(X = 0) = 1 - P(X = 1) = p$ and $P(Y = 0) = 1 - P(Y = 1) = q$, we obtain $\lambda_1 = 2p(1 - p)$, $\mu_1 = 2q(1 - q)$, and $\lambda_k = \mu_k = 0$ for $k > 1$ so that $\rho^*(X,Y) = \rho(X,Y)^2$, see Theorem 2, Part 4. In this special case the decomposition consists of just one component.

3.3. *Parameterization of the likelihood* Let f_{12} be the joint density of (X,Y) with corresponding marginal densities f_1 and f_2. Since

$$\rho_{kl} = \int \frac{f_{12}(x,y)}{f_1(x)f_2(y)} g_{1k}(x)g_{2l}(y)dF_1(x)dF_2(y)$$

we can decompose the joint density as:

$$f_{12}(x,y) = f_1(x)f_2(y)\left(1 + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} g_{1k}(x)g_{2l}(y)\rho_{kl}\right) \tag{3.27}$$

A similar equation can be given for discrete distributions, and a general treatment can be given using the Radon-Nikodym derivative.

Decomposition (3.27) may be compared to the well-known canonical correlation decomposition

$$f_{12}(x,y) = f_1(x)f_2(y)\left(1 + \sum_{k=1}^{\infty} a_{1k}(x)a_{2k}(y)\rho_k\right)$$

Here, a_{1k} and a_{2k} are those functions maximizing the correlation between X and Y, subject to the restraint (for $k > 1$) that they are orthogonal to $a_{11}, \ldots, a_{1,k-1}$ and $a_{21}, \ldots, a_{2,k-1}$, respectively, and ρ_k is the correlation between $a_{1k}(X)$ and $b_{2k}(Y)$.

3.4. *Fréchet bounds for component correlations* Below, we discuss the interpretation and properties of the component correlations. In particular, we look at bounds for the component correlations.

For two random variables X and Y with joint distribution function F_{12} and marginal distribution functions F_1 and F_2, the well-known Fréchet upper bound F_{12}^+ is defined by

$$F_{12}^+(x,y) = \min\{F_1(x), F_2(y)\}$$

and the Fréchet lower bound F_{12}^- is defined by

$$F_{12}^-(x,y) = \max\{0, 1 - F_1(x) - F_2(y)\}$$

Then $\rho(X,Y) = 1$ if and only if $F_{12} = F_{12}^+$ and $\rho(X,Y) = -1$ if and only if $F_{12} = F_{12}^-$. A more general question is, for functions g and h, for which F_{12} the correlation between $g(X)$ and $h(Y)$ is maximal or minimal. Let

$$S_{g,h}^+ = \left\{(x,y) \in \mathbf{R}^2 | g(x) = h(y)\right\}$$

and

$$S_{g,h}^- = \left\{(x,y) \in \mathbf{R}^2 | g(x) = -h(y)\right\}$$

Then we have:

Lemma 11. *For functions g and h, $\rho[g(X), h(Y)] = 1$ iff the support of the distribution of (X,Y) is a subset of $S_{g,h}^+$ and and $\rho[g(X), h(Y)] = -1$ iff the support is a subset of $S_{g,h}^-$.*

Proof. Suppose for simplicity that g and h are standardized. Then $\rho[g(X), h(Y)] = 1$ iff $P[g(X) = h(Y)] = 1$ and $\rho[g(X), h(Y)] = -1$ iff $P[g(X) = -h(Y)] = 1$, and the lemma immediately follows. □

For the dichotomous case we obtain the following:

Example 3. *If both variables are dichotomous, say $P(X = 0) = 1 - P(X = 1) = p$ and $P(Y = 0) = 1 - P(Y = 1) = q$, we obtain $S_{11}^+ = \{(0,0), (1,1)\}$ and $S_{11}^- = \{(0,1), (1,0)\}$.*

Note that the bounds need not be attainable since it may be the case that, for example, for certain x, there is no y such that $(x,y) \in S_{g,h}^+$.

Here we are interested in the bounds for the component correlations ρ_{ij} of ρ^*. For simplicity, we write $S_{ij}^\pm = S_{g_i,g_j}^\pm$. The next example shows that if both X and Y have uniform distributions on $[0,1]$, then the component correlations of ρ^* can attain the bounds 1 and -1:

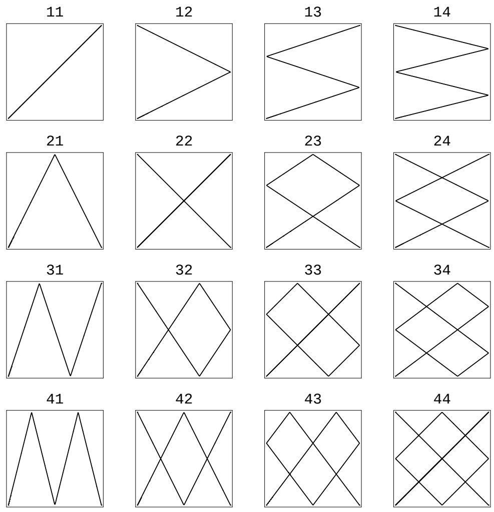
Example 4. *Suppose X and Y are uniformly distributed on $[0,1]$. Then the eigenfunctions are the Fourier cosine functions (see Table 1). The set S_{kl}^+ is formed by the solutions (x,y) to the equation*

$$\cos(k\pi x) = \cos(l\pi y)$$

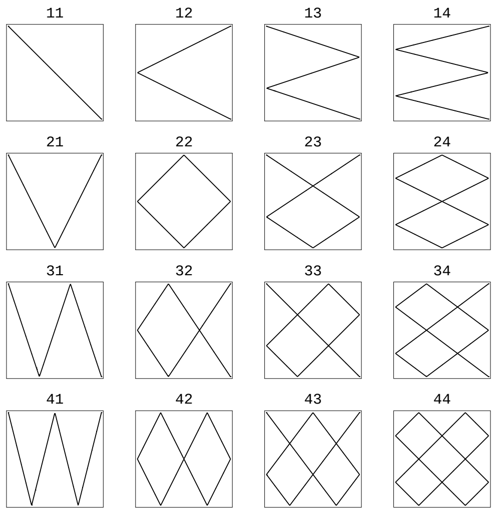
and S_{kl}^- by the solutions of

$$\cos(k\pi x) = -\cos(l\pi y)$$

The solutions are plotted in Figure 2, for $k = 1, \ldots, 4$ and $l = 1, \ldots, 4$. The bounds for the ρ_{kl} are attainable since $\rho_{kl}(X,Y) = 1$ for the uniform distribution on S_{kl}^+ and $\rho_{kl}(X,Y) = -1$ for the uniform distribution on S_{kl}^-.



(a) Positive bounds S_{ij}^{+}: $\rho_{ij} = 1$ if support of (X, Y) is subset of S_{ij}^{+}



(b) Negative bounds S_{ij}^{-}: $\rho_{ij} = -1$ if support of (X, Y) is subset of S_{ij}^{-}

FIG. 2. *Supports of the Fréchet bounds for the component correlations ρ_{ij} of $\rho^*(X, Y)$ when X and Y are uniformly distributed on an interval.*

Note that, if $\rho_{11} = 1$, then $\rho_{22} = 1$, since $S_{11}^+ \subset S_{22}^+$. More generally, by the same reasoning, we have for all $i > 1$,

$$\rho_{kl}(X, Y) = 1 \Rightarrow \rho_{i\times k, i\times l}(X, Y) = 1$$

and

$$\rho_{kl}(X, Y) = -1 \Rightarrow \rho_{i\times k, i\times l}(X, Y) = (-1)^i$$

By a symmetry argument, we also have $\rho_{11} = 1 \Rightarrow \rho_{12} = 0$, and there are various other similar implications.

An overview of a large amount of literature on Fréchet bounds is given in Rüschendorf (1991)

4. Estimation and tests of independence In this section we discuss estimation of κ and ρ^* by U- and V-statistics. Roughly speaking, the U-statistic estimator of a parameter is an unbiased estimator based on taking averages (Hoeffding, 1948a), and the V-statistic estimator is the estimator based on the distribution obtained by assigning a probability weight $1/n$ to each sample point. For κ, both the U- and V-statistic estimators are available, but for ρ^* only the latter is. However, we can estimate ρ^* by a function of U-statistic estimators.

In Section 4.1, it is shown how estimators of κ and ρ^* by U- and V-statistics are obtained. In Section 4.2 permutation tests, useful for small samples, are described. In Section 4.3, the asymptotic distribution of these estimators is derived under the null hypothesis of independence. In Section 4.4, Bonferroni corrections for tests of significance of the component correlations are described.

4.1. *U and V statistic estimators of κ* We first give a method for calculating the U- and V-statistic estimators of κ based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, then we give the related estimates for ρ^*.

The V-statistic estimator $\hat{\kappa}$ is the value of κ based on the sample distribution functions $\hat{F}_1$ and $\hat{F}_2$, and is obtained as follows. Let

$$A_{1k} = \frac{1}{n}\sum_{i=1}^{n} |X_k - X_i|$$

$$A_{2k} = \frac{1}{n}\sum_{i=1}^{n} |Y_k - Y_i|$$

and

$$B_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |X_i - X_j|$$

$$B_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |Y_i - Y_j|$$

Then we have for $k, l = 1, \ldots, n$,

$$h_{\hat{F}_1}(x_1, x_2) = -\tfrac{1}{2}\left(|x_1 - x_2| - A_{1k} - A_{1l} + B_1\right)$$
$$h_{\hat{F}_2}(y_1, y_2) = -\tfrac{1}{2}\left(|y_1 - y_2| - A_{2k} - A_{2l} + B_2\right)$$

and the sample or V-statistic estimator of κ is given as

$$\hat{\kappa} = \frac{1}{n^2}\sum_{i,j=1}^{n} h_{\hat{F}_1}(X_i, X_j) h_{\hat{F}_2}(Y_i, Y_j)$$

Now with

$$\tilde{h}_{\hat{F}_1}(x_1, x_2) = -\tfrac{1}{2}\left(|x_1 - x_2| - \frac{n}{n-1}A_{1k} - \frac{n}{n-1}A_{1l} + \frac{n}{n-1}B_1\right)$$
$$\tilde{h}_{\hat{F}_2}(y_1, y_2) = -\tfrac{1}{2}\left(|y_1 - y_2| - \frac{n}{n-1}A_{2k} - \frac{n}{n-1}A_{2l} + \frac{n}{n-1}B_2\right)$$

for $k, l = 1, \ldots, n$, the unbiased or U-statistic estimator of κ is given as

$$\tilde{\kappa} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \tilde{h}_{\hat{F}_1}(X_i, X_j)\tilde{h}_{\hat{F}_2}(Y_i, Y_j)$$

By Hoeffding's theory of U-statistics we have that $\tilde{\kappa}$ is an unbiased estimator of κ (Hoeffding, 1948a; Randles & Wolfe, 1979). Note that $\hat{\kappa} \geq 0$ but $\tilde{\kappa}$ may be negative.

The related estimators of ρ^* are the following:

$$\hat{\rho}^*(X, Y) = \frac{\hat{\kappa}(X, Y)}{\sqrt{\hat{\kappa}(X, X)\hat{\kappa}(Y, Y)}}$$
$$\tilde{\rho}^*(X, Y) = \frac{\tilde{\kappa}(X, Y)}{\sqrt{\tilde{\kappa}(X, X)\tilde{\kappa}(Y, Y)}}$$

For both types of estimators, the computational complexity of the above method is $O(n^2)$.

The marginal eigenvalues and functions can be computed numerically from $h_{\hat{F}_1}$ and $h_{\hat{F}_2}$ or from $\tilde{h}_{\hat{F}_1}$ and $\tilde{h}_{\hat{F}_2}$. See also Section 2.5 for computational aspects.

4.2. *Permutation tests* Under independence, the sample marginal distributions of X and Y are ancillary statistics for $\hat{\rho}^*$ and $\tilde{\rho}^*$, so by Fisher's theory of fiducial inference we should condition on the sample marginals when testing for independence using $\hat{\rho}^*$ and $\tilde{\rho}^*$. If $X \perp\!\!\!\perp Y$, conditioning on the marginals ensures that $\hat{\rho}^*$ and $\tilde{\rho}^*$ are distribution free, and exact conditional p-values can be calculated using the permutation method. Evaluating all permutations quickly becomes computationally prohibitive even for moderately large sample sizes, and we recommend using a set of random permutations. Note that permutation tests may also be applied to the component correlations $\hat{\rho}_{ij}$ and $\tilde{\rho}_{ij}$.

Permutation tests may be computationally intensive. Using non-optimized software, our experience shows that (bootstrap) permutation tests for up to a several hundred observations are feasible: for $n = 100$, the permutation test based on 1000 random permutations took less than four minutes, and for $n = 500$ it took a bit more than one hour. Techniques for the fast *exact* evaluation of permutation tests using generating functions are described by, among others, Baglivo, Pagano, and Spino (1996) and Van de Wiel, Di Bucchianico, and Van der Laan (1999), but it is not clear whether these techniques extend to statistics such as $\hat{\rho}^*$ which are not based on ranks.

For categorical data the permutation test is better known as an *exact conditional test* (where the conditioning is, again, on the marginal distributions), the Fisher exact test being the best known example. There is a large body of literature on fast evaluation of exact conditional p-values for contingency tables, for an overview see Agresti (1992) and for more recent developments see Forster, McDonald, and Smith (1996), Diaconis and Sturmfels (1998), Booth and Butler (1999).

If the permutation test is too computationally intensive, an asymptotic test may be done using the results of the next section.

4.3. *Asymptotic distribution of estimators under independence* For the asymptotic distribution of the estimators we obtain the following:

THEOREM 4. *Suppose h_{F_1} and h_{F_2} are square integrable with spectral decompositions (3.25) and (3.26). Then if $X \perp\!\!\!\perp Y$ and with Z_{ij} iid standard normal variables, we obtain*

$$n\tilde{\kappa}(X,Y) \to_D \sum_{i,j=0}^{\infty} \lambda_i \mu_j (Z_{ij}^2 - 1)$$

If additionally h_{F_1} and h_{F_2} are trace class, we obtain

$$n\hat{\kappa}(X,Y) \to_D \sum_{i,j=0}^{\infty} \lambda_i \mu_j Z_{ij}^2$$

PROOF. By the Hoeffding (1961) decomposition we can write with $R_n = O(n^{-3})$

$$\begin{aligned}
\tilde{\kappa} &= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h_{F_1}(x_i, x_j) h_{F_2}(y_i, y_j) + R_n \\
&= \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \left(\sum_{k=1}^{\infty} \lambda_k g_{1k}(x_i) g_{1k}(x_j) \right) \left(\sum_{l=1}^{\infty} \mu_l g_{2l}(y_i) g_{2l}(y_j) \right) + R_n \\
&= \binom{n}{2}^{-1} \sum_{k,l=1}^{\infty} \lambda_k \mu_l \left[\left(\sum_{i=1}^{n} g_{1k}(x_i) g_{2l}(y_i) \right)^2 - \sum_{i=1}^{n} g_{1k}(x_i)^2 g_{2l}(y_i)^2 \right] + R_n
\end{aligned}$$

Since $n^{-1}\sum_{i=1}^n g_{1k}(x_i)^2 g_{2l}(y_i)^2 \to 1$ a.s., and $n^{-1}\left(\sum_{i=1}^n g_{1k}(x_i)g_{2l}(y_i)\right)^2 \to_D Z_{kl}^2$ we obtain using the Cramér-Wold device that

$$n\tilde{\kappa} \to_D \sum_{i,j=0}^{\infty} \lambda_i \mu_j (Z_{ij}^2 - 1)$$

The proof for $\hat{\kappa}$ is similar; we have

$$\begin{aligned}
\hat{\kappa} &= \frac{1}{n^2}\sum_{i,j=1}^{n} h_{F_1}(x_i,x_j)h_{F_2}(y_i,y_j) + R_n \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\left(\sum_{k=1}^{\infty}\lambda_k g_{1k}(x_i)g_{1k}(x_j)\right)\left(\sum_{l=1}^{\infty}\mu_l g_{2l}(y_i)g_{2l}(y_j)\right) + R_n \\
&= \frac{1}{n^2}\sum_{k,l=1}^{\infty}\lambda_k\mu_l\left(\sum_{i=1}^{n} g_{1k}(x_i)g_{2l}(y_i)\right)^2 + R_n
\end{aligned}$$

Since $n^{-1}\left(\sum_{i=1}^n g_{1k}(x_i)g_{2l}(y_i)\right)^2 \to_D Z_{kl}^2$ we obtain using the Cramér-Wold device that

$$n\hat{\kappa} \to_D \sum_{i,j=0}^{\infty} \lambda_i\, \mu_j\, Z_{ij}^2$$

under the condition that

$$\lim_{n\to\infty} E(n\hat{\kappa}) = \sum_{i,j=0}^{\infty}\lambda_i\,\mu_j = \sum_{i=1}^{\infty}\lambda_i\sum_{j=0}^{\infty}\mu_j$$

is finite. Now by Lemma 5, the two factors on the right hand side are finite iff h_{F_1} and h_{F_2} are trace class, completing the proof. □

The proof is similar to an adaptation by De Wet (1987) of a proof by Eagleson (1979). See also Gregory (1977) and Hall (1979).

Note that by Lemma 3, h_{F_1} and h_{F_2} are trace class iff EX and EY exist. As follows from the theorem and noted earlier by De Wet (1987) for related statistics, the U-statistic estimator has an asymptotic distribution in more cases than the V-statistic estimator. For example, if the marginal distribution of at least one of X and Y is the distribution of Example 2, $n\hat{\kappa}$ does not have an asymptotic distribution but $n\tilde{\kappa}$ does have one.

4.4. *Bonferroni corrections for testing significance of component correlations* As well as testing the significance of $\hat{\rho}^*$ directly, we can test for the significance of the empirical component correlations $\hat{\rho}_{ij}$. We recommend using $\hat{\rho}^*$ rather than $\tilde{\rho}^*$ for calculating component correlations, since $\tilde{\rho}^*$ may be negative in which case no component correlations with nonnegative eigenvalues exist.

The proof of Theorem 4 suggests that the component correlations $\hat{\rho}_{kl}$ are asymptotically normal and independent. Since there are many component correlations, a simultaneous test of their significance needs a Bonferroni correction. The ordinary

Bonferroni correction, i.e., multiplying the exceedance probabilities by the number of tests done, which in this case is n^2, would be unreasonable since the multiplication factor increases rapidly with n. Instead we propose dividing the exceedance probability for $\hat{\rho}_{ij}$ by

$$\frac{\hat{\lambda}_i \hat{\mu}_j}{\sum_{i=1}^n \hat{\lambda}_i \sum_{j=1}^n \hat{\mu}_j}$$

Note that these numbers will converge to zero in probability if at least one of h_{F_1} and h_{F_2} is not of trace class, i.e., by Lemma 3, if at least one of EX or EY does not exist, in which case the correction may not be the most appropriate one.

The idea of looking at components of a test seems to have first appeared in Durbin and Knott (1972), where components of the Cramér-von Mises test were investigated. This test is a special case of the tests based on ρ^* described above (see Section 5.2).

Other related work is by Kallenberg and Ledwina (1999), who looked at correlations between orthogonal functions of the marginal cumulative distribution functions, in particular, the Legendre polynomials. This work is an extension of the so-called smooth tests of fit of Neyman (1937). Rather than looking at all correlations between the orthogonal functions, they considered just the first few, and developed a selection method based on Schwartz's rule for determining how many correlations to base the overall test on.

5. Grade versions of κ and ρ^*, copulas, and rank tests For ordinal random variables X and Y, any given scale is arbitrary and it may be desirable to use scales based on the grades $F_1(X)$ and $F_2(Y)$ of X and Y. A general way to base κ and ρ^* on grades is as follows. For given invertible distribution functions K_1 and K_2, we can define

$$\kappa_{K_1,K_2}(X,Y) = \kappa[K_1^{-1} \circ F_1(X), K_2^{-1} \circ F_2(Y)]$$

and

$$\rho^*_{K_1,K_2}(X,Y) = \rho^*[K_1^{-1} \circ F_1(X), K_2^{-1} \circ F_2(Y)]$$

Note that

$$\kappa_{F_1,F_2}(X,Y) = \kappa(X,Y)$$

and

$$\rho^*_{F_1,F_2}(X,Y) = \rho^*(X,Y)$$

With K_1 and K_2 uniform distribution functions, $\rho^*_{K_1,K_2}(X,Y)$ is to ρ^* what Spearman's rho is to the ordinary correlation ρ.

We can use the results of Section 4 to obtain an orthogonal decomposition of κ_{K_1,K_2} and $\rho^*_{K_1,K_2}$ in terms of component correlations. These component correlations then parameterize the *copula*, which is defined as the joint distribution of $(F_1(X), F_2(Y))$. From (3.27), and since the marginal eigenfunctions of h_F with F

the uniform distribution are the cosine functions given in Table 1, we obtain the following decomposition of c_{12}, the density function of the copula:

$$c_{12}(u,v) = 1 + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \cos(k\pi u)\cos(l\pi v)\rho_{kl}$$

where $\rho_{kl} = \int \cos(k\pi u)\cos(l\pi v)c_{12}(u,v)dudv$. This decomposition was earlier given in De Wet (1980) and Deheuvels (1981). An overview of copula theory is given in Nelsen (2006). Possible drawbacks of using $\rho^*_{K_1,K_2}$ for some given K_1 and K_2 is the arbitrariness of any choice of K_1 and K_2 and the loss of scale information, but these issues are hotly debated (Mikosch, 2006).

In Section 5.1, a brief description of rank tests based on κ is given. In Section 5.2 a generalization of the Cramér-von Mises test to the case of K ordered samples is shown to be a special case, and a convenient representation is given. In Section 5.3 we write κ_{K_1,K_2} as a weighted average of ϕ-coefficients.

5.1. *Rank tests* Rank statistics which are distribution free under independence in the continuous case are obtained as follows. For invertible distribution functions K_1 and K_2 let

$$\hat\rho^*_{K_1,K_2}(X,Y) = \hat\rho^*[K_1^{-1}\circ \hat F_1(X), K_2^{-1}\circ \hat F_2(Y)]$$
$$\tilde\rho^*_{K_1,K_2}(X,Y) = \tilde\rho^*[K_1^{-1}\circ \hat F_1(X), K_2^{-1}\circ \hat F_2(Y)]$$

The derivation of the asymptotic distribution of these statistics is slightly more involved than that of the asymptotic distribution of $\hat\rho^*(X,Y)$. De Wet (1980) has done this derivation for statistics related to $\tilde\rho^*_{K_1,K_2}(X,Y)$. He gave the weights for optimal tests in the Bahadur sense for certain classes of alternatives, such as the bivariate normal.

With K_1 and K_2 the uniform distribution functions, $n\hat\rho^*_{K_1,K_2}(X,Y)$ is a statistic discussed by Blum, Kiefer, and Rosenblatt (1961), see also Deheuvels (1981). It can be viewed as a generalization of the ordinary Cramér-von Mises test (see next subsection). Similarly, with K_1 and K_2 the logistic distribution functions, $n\hat\rho^*_{K_1,K_2}(X,Y)$ can be viewed as a generalization of the Anderson-Darling test.

Hoeffding (1948b) described a related test, namely based on the U-statistic estimator of

$$\int [F_{12}(x,y) - F_1(x)F_2(y)]^2\, dF_{12}(x,y)$$

which can be obtained from the representation of κ in Lemma 10, Part 1, by replacing $dxdy$ by $dF_{12}(x,y)$. Hoeffding's coefficient does not fall in the framework of the present paper.

5.2. *A new class of K-sample Cramér-von Mises tests as a special case* Suppose we have K samples, the kth sample having n_k iid observations, say $\{U_{k1},\ldots,U_{kn_k}\}$. Then a test whether the distributions of the observations in the different samples are equal is called a *K-sample test*. A K-sample test can, in fact, be viewed as a test of independence, namely, whether 'response' depends on 'group membership,' the groups referring to the different samples. Let us consider the case that the score $c_k \in \mathbf{R}$ is assigned to sample k $(k = 1, \ldots, K)$. With $N_0 = 0$ and $N_k = \sum_{i=1}^{k} n_i$ let $(X_{N_{k-1}+i_k}, Y_{N_{k-1}+i_k}) = (c_i, U_{k,i_k})$ for $k = 1, \ldots, K$ and $i_k = 1, \ldots, n_k$. Then it can be seen that the K sample test is a test of independence of the X observations and the Y observations. (Note that here the X observations are not random). A K-sample test can then be based on $\hat{\rho}^*$ or $\tilde{\rho}^*$.

If samples are ordered but have no numerical scores assigned to them, rank scores can be assigned, for example $c_k = N_k$.

In order to arrive at the Cramér-von Mises test, we now use Lemma 10, Part 2 to give a representation of κ in terms of the conditional distribution functions. Let G_k be the distribution function of U_k, the response for sample k, and let $p_k = n_k/N_K$ be the proportion of observations in sample k. Then we obtain

$$\kappa(X, Y) = \sum_{i,j} p_i p_j h_F(c_i, c_j) \int [G_i(y) - F_2(y)]\,[G_j(y) - F_2(y)]\, dy$$

Some straightforward algebra shows that for the two-sample case this reduces to

$$\kappa(X, Y) = p_1^2 p_2^2 \int [G_1(y) - G_2(y)]^2\, dy$$

A grade version of κ is

$$\kappa_{F_1,K}(X, Y) = \sum_{i,j} p_i p_j h_F(c_i, c_j) \int [G_i(y) - F_2(y)]\,[G_j(y) - F_2(y)]\, w[F_2(y)] dF_2(y)$$

where

$$w(u) = \frac{1}{k[K^{-1}(u)]}$$

In the two-sample case, the sample version of $\kappa_{F_1,K}(X, Y)$ with K the uniform distribution function reduces (essentially) to the ordinary Cramér-von Mises statistic, so we have a generalization to the case of K ordered samples. With K the logistic distribution, $\hat{\kappa}_{F_1,K}(X, Y)$ reduces to the Anderson-Darling statistic.

A different generalization of the two-sample Cramér-von Mises test was given by Kiefer (1959), namely to the case of K *unordered* samples.

5.3. *κ as a weighted ϕ-coefficient* From Lemma 10, Part 1, we directly obtain

$$\kappa_{K_1,K_2}(X, Y) = \int [F_{12}(x, y) - F_1(x)F_2(y)]^2\, dK_1^{-1} \circ F_1(x)\, dK_2^{-1} \circ F_2(y) \tag{5.28}$$

This result leads to an interesting interpretation of κ_{K_1,K_2}. The ϕ coefficient for measuring the dependence in the 2×2 table obtained by collapsing the distribution with respect to the cut point (x, y) is given as

$$\phi(x, y) = \frac{|F_{12}(x, y) - F_1(x)F_2(y)|}{\sqrt{F_1(x)[1 - F_1(x)]F_2(y)[1 - F_2(y)]}} \tag{5.29}$$

Now suppose ψ is such that

$$\phi(x, y) = \psi[F_1(x), F_2(y)]$$

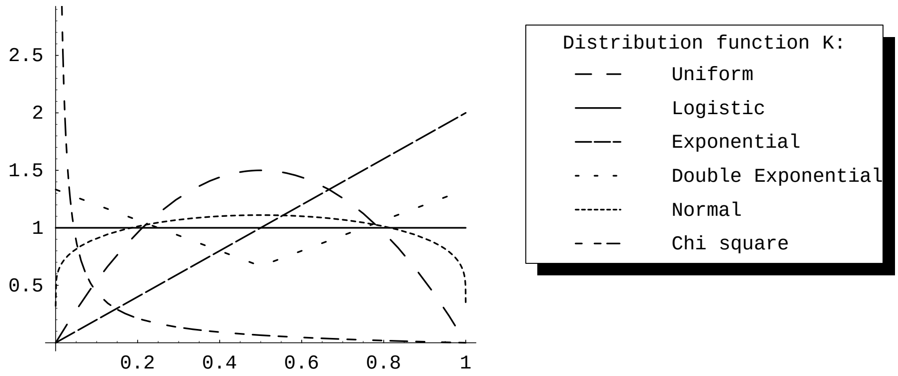


FIG. 3. *Plots of $\bar{w}_K(u)$ for several distributions K*

Then from (5.28) we obtain that κ_{K_1,K_2} can be written as a weighted average ψ-square:

$$\kappa_{K_1,K_2}(X,Y) = \int \psi(u,v)^2\, w_{K_1}(u)\, w_{K_2}(v) du dv$$

where the weight function w is defined by

$$w_K(u) = \frac{u(1-u)}{k[K^{-1}(u)]}$$

The normalized weight function (integrating to one) is

$$\bar{w}_K(u) = \frac{w_K(u)}{\int_0^1 w_K(u)du}$$

where

$$\int_0^1 w_K(u)du = \int_0^1 u(1-u)dK^{-1}(u) = \int_{-\infty}^{\infty} K(x)[1-K(x)]dx$$

In Figure 3, $\bar{w}_K$ is plotted for the distribution functions K given in Table 1.

From Figure 3 we can deduce which ϕ-coefficients are given most weight for different marginal distributions. As the reference marginal, we take the logistic, which is the horizontal line in the figure, i.e., assigning uniform weights. We see that for a uniform marginal, the weight goes to zero in the tails. The weights for a normal marginal are for most u intermediate between the weights for the uniform and logistic marginals. In contrast to uniform and normal marginals, the Laplace distribution gives more weight to the tails than a logistic marginal. An exponential marginal gives little weight to the lower tail, but much weight to the upper tail. Finally, the chi-square distribution gives very large weight to the lower tail and very small weight to the upper tail. Among the distributions considered, the biggest difference is between the chi-square and the exponential distribution.

6. Data analysis: investigating the nature of the association Gaining an understanding of the nature of an association between two random variables is probably best viewed as an art rather than a science, and in this section we present some visual tools based on ρ^* and its components which may be helpful in reaching this objective. For an iid bivariate sample $\{(X_i, Y_i)\}$ we propose the following two procedures.

Firstly, we calculate $\hat{\rho}^*$ from the sample and test its significance. If found to be significant, then for each data point (X_i, Y_i) we calculate the weight

$$W_i = \frac{\frac{1}{n}\sum_{j=1}^n h_{\hat{F}_1}(X_i, X_j)\, h_{\hat{F}_2}(Y_i, Y_j)}{\sqrt{\hat{\kappa}(X,X)\,\hat{\kappa}(Y,Y)}}$$

Since

$$\hat{\rho}^*(X,Y) = \frac{1}{n}\sum_{i=1}^n W_i$$

the weights W_i give an indication of how much the sample element (X_i, Y_i) contributes to $\hat{\rho}^*$, and so can be used to discover the nature of a possible association between X and Y.

Secondly, we calculate the component correlations $\hat{\rho}_{kl}$ of $\hat{\rho}^*$ and test their significance using the Bonferroni corrections described in Section 4.4. Then for each significant component correlation $\hat{\rho}_{kl}$, we compute the weights

$$W_i^{(k,l)} = \hat{g}_{1k}(X_i)\hat{g}_{2l}(Y_i)$$

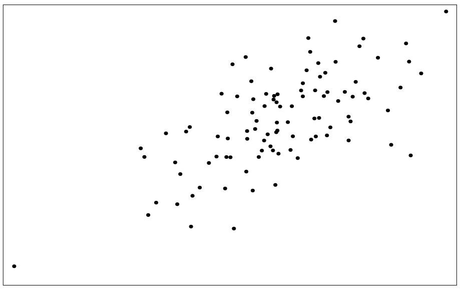
where g_{1k} and g_{2l} are the eigenfunctions belonging to h_{F_1} and h_{F_2}. Since

$$\hat{\rho}_{kl} = \frac{1}{n}\sum_{i=1}^n W_i^{(k,l)}$$

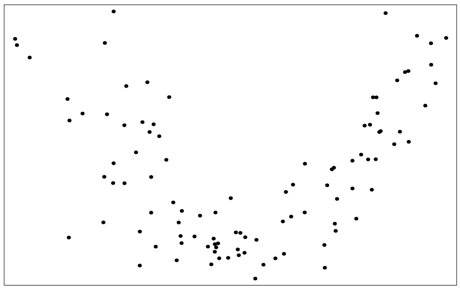
the weight $W_i^{(k,l)}$ is the amount the sample element (X_i, Y_i) contributes to $\hat{\rho}_{kl}$ (conditionally on the marginals), and so, like W_i, can be used to investigate the association between X and Y.

In this section we show how to visualize the weights W_i and $W_i^{(k,l)}$, both for continuous and categorical data, and show how this can be used to gain an understanding of the association. Some artificial continuous data sets are considered in Section 6.1, a real categorical data set is considered in Section 6.2 and a real time series data set is considered in Section 6.3

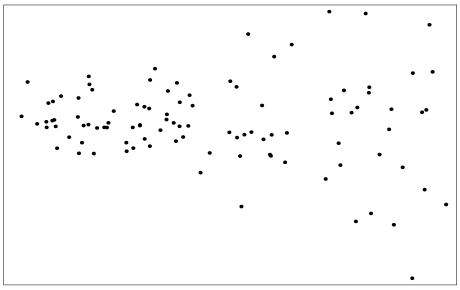
6.1. *Some artificial data sets* In Figure 4, four artificial data sets are plotted, each consisting of 100 iid points. For completeness, we explain how the data were generated. In the following, U is uniformly distributed on $[0,1]$ and Z_1 and Z_2 are iid standard normal random variables, and $Z(u)$ has a normal distribution with mean zero and standard deviation u. The data in Figure 4(a) are from a bivariate normal distribution with $\rho = 2/3$. The data in Figure 4(b) are of the form $(X,Y) = (U, (U-1/2)^2) + (Z_1/10, Z_2/10)$. The data in Figure 4(c) are of the form $(X,Y) = (U, Z(1/5+U))$. Finally, the data in Figure 4(d) are of the form $(X,Y) = (U, Z(1/5+\min\{U, 1-U\}))$.



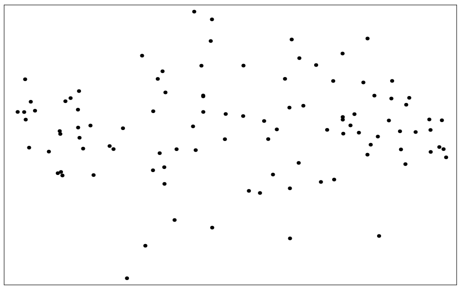
(a) Increasing location ($\hat{\rho}^* = .36$)



(b) Decreasing then increasing location ($\hat{\rho}^* = .17$)

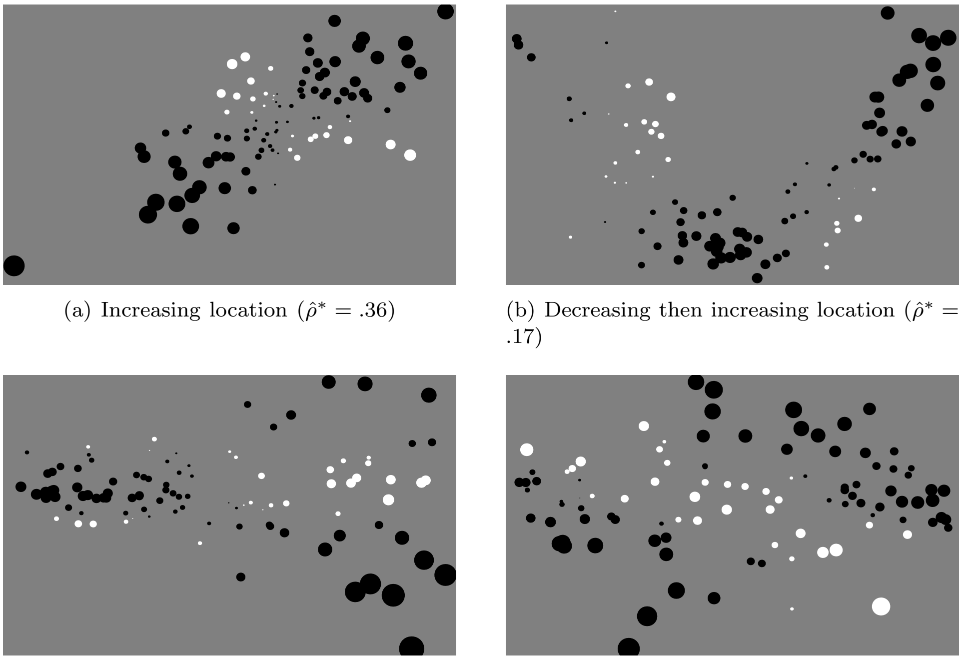


(c) Increasing dispersion ($\hat{\rho}^* = .11$)



(d) Increasing then decreasing dispersion ($\hat{\rho}^* = .02$)

FIG. 4. *Scatterplots of artificial data sets. The captions denote what happens to the conditional Y distribution as X increases.*



(a) Increasing location ($\hat{\rho}^* = .36$)

(b) Decreasing then increasing location ($\hat{\rho}^* = .17$)

(c) Increasing dispersion ($\hat{\rho}^* = .11$)

(d) Increasing then decreasing dispersion ($\hat{\rho}^* = .02$)

FIG. 5. *Representation of weights W_i for data in Figure 4. The size of each dot is proportional to $|W_i|$; black dots represent positive W_i and white represent negative W_i. Total area of dots is scaled to a fixed constant for each plot.*

For all four data sets, we performed permutation tests for the significance of $\hat{\rho}^*$ and its component correlations $\hat{\rho}_{ij}$ based on 10,000 random permutations. This took us about 48 minutes for each data set. For the significant component correlations we took 1 million random permutations to get a more accurate p-value, and this took about 110 seconds per component correlation. We also computed the ordinary correlation, and, not surprisingly, only for the data in Figure 4(a) it is significantly different from zero. There we found that $\hat{\rho} = .44$ ($p = .000$).

For the data in Figure 4(a), we found that $\hat{\rho}^* = .36$ ($p = .000$), i.e., there is significant association. In Figure 5(a), the data are plotted again, this time each sample element (X_i, Y_i) is represented with size proportional to $|W_i|$; black dots represent positive W_i and white represent negative W_i. In all these plots, the total area of the dots is scaled to a constant to make it easier to study them. From Figure 5(a),we see that the association seems to consist of a linearity in the data. It may be worthwhile however to check if there are other forms of association present by looking at the component correlations of $\hat{\rho}^*$. We found two significant components: $\hat{\rho}_{11} = .61$ ($p = .000$) and $\hat{\rho}_{22} = .38$ ($p = .004$). In Figures 6(a) and 6(b) the weights $W_i^{(11)}$ and $W_i^{(22)}$ are visualized. The interpetation of the black and white

	Mental Health Status			
Parents' Socioeconomic Status	Well	Mild Symptom Formation	Moderate Symptom Formation	Impaired
A (high)	64	94	58	46
B	57	94	64	40
C	57	105	65	60
D	72	141	77	94
F	36	97	54	78
G (low)	21	71	54	71

Table 3
Cross-classification of Mental Health Status and Socioeconomic Status

dots is the same as above. The gridlines correspond to the zeroes of the marginal eigenfunctions. Therefore, within any rectangle the dots have the same color. Also in this case, the plots point to a linearity in the data.

For the data in Figure 4(b), we found $\hat{\rho}^* = .17$ ($p = .000$) and we found two significant component correlations: $\hat{\rho}_{21} = -.78$ ($p = .000$) and $\hat{\rho}_{42} = .54$ ($p = .000$). The plots in Figures 5(b), 6(c) and 6(d) all point to a curved relationship.

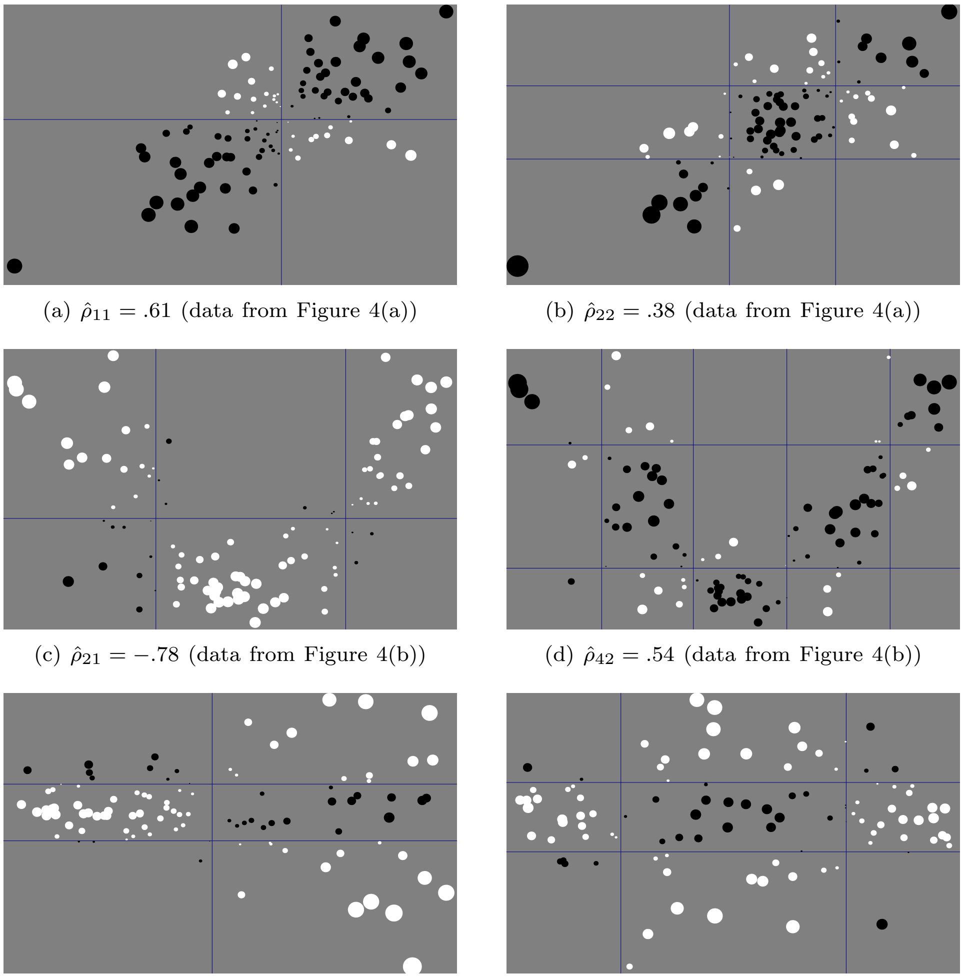
For the data in Figure 4(c), we found that $\hat{\rho}^* = .11$ ($p = .001$). There is significant association at the 5% level, but the evidence is not as overwhelming as in the previous two cases. We only found one significant component correlation: $\hat{\rho}_{12} = -.51$ ($p = .000$). Figures 5(c) and 6(e) both point towards and increase in dispersion of the Y variable as X increases.

For the data in Figure 4(d), we found that $\hat{\rho}^* = .02$ ($p = .522$). This time, the test based on ρ^* does not yield a significant association. However, there is one significant component correlation: $\hat{\rho}_{22} = -.38$ ($p = .004$). Figure 6(f) indicates that the association is due to an increase and then decrease in dispersion. Since $\hat{\rho}^*$ is not significant, we should refrain from giving an interpretation based on Figure 5(d).

For comparative purposes, we plotted the weights W_i for data drawn from a distribution satisfying independence in Figure 7. Both sets consist of 100 sample elements. The most common pattern is that of Figure 7(a), with two diagonally opposing clusters of black dots and two diagonally opposing clusters of white dots. In a very limited investigation, this type of pattern occured about half of the time. Otherwise more complex patterns were obtained, such as the one in Figure 7(b). These figures indicate that it doesn't seem to make sense to interpret this kind of plot if $\hat{\rho}^*$ is not significant, such as Figure 5(d).

Concluding, we see that significance tests combined with an inspection of the two types of plots in Figures 5 and 6 can give us insight into the kind of association present between two random variables.

6.2. *Mental health data* Table 3 describes the relationship between child's mental impairment and parents' socioeconomic status for a sample of residents of Man-



(a) $\hat{\rho}_{11} = .61$ (data from Figure 4(a))

(b) $\hat{\rho}_{22} = .38$ (data from Figure 4(a))

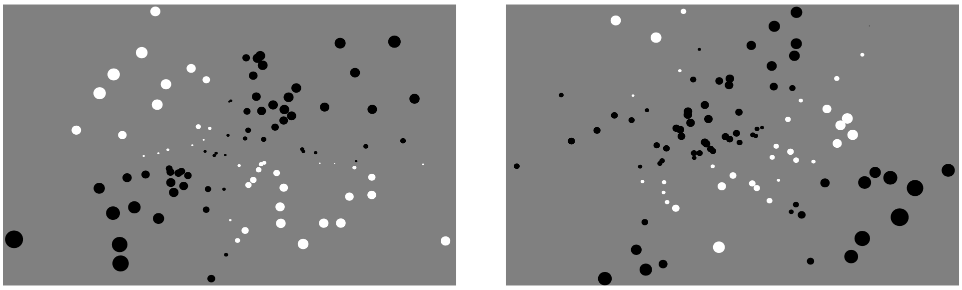
(c) $\hat{\rho}_{21} = -.78$ (data from Figure 4(b))

(d) $\hat{\rho}_{42} = .54$ (data from Figure 4(b))

(e) $\hat{\rho}_{12} = -.51$ (data from Figure 4(c))

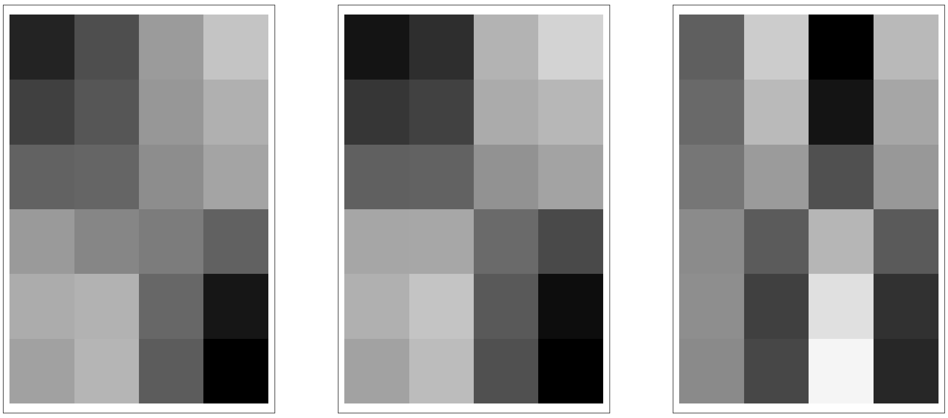
(f) $\hat{\rho}_{22} = -.38$ (data from Figure 4(d))

FIG. 6. *Representation of weights $W_i^{(k,l)}$ contributing to $\hat{\rho}_{kl}$ for data in Figure 4. The meaning of the dots is otherwise the same as in Figure 5.*



(a) Most common pattern (b) Less common pattern

FIG. 7. *Representation of weights W_i for data drawn from distributions satisfying independence.*



(a) $\hat{\rho}^* = .02$ (b) $\hat{\rho}_{11} = .13$ (c) $\hat{\rho}_{13} = .08$

FIG. 8. *Representation of weights W_{ab}, $W_{ab}^{(11)}$ and $W_{ab}^{(13)}$, respectively, for mental health data of Table 3. The darker the shade of gray in cell (a, b) the larger W_{ab}; $W_{ab} = 0$ is represented by an intermediate shade of gray.*

hattan (Goodman, 1985; Agresti, 2002, and references therein). Goodman used this table to illustrate various association models for categorical data, including the so-called linear by linear association model, the row and columns effects model, and correspondence analysis based on canonical correlations. Here we illustrate the use of $\hat{\rho}^*$ and its components as yet an alternative method for analyzing these data. We relied on asymptotic p-values because with 1670 observations approximate evaluation of the permutation tests would have been too time consuming using our implementation.

In the categorical case, for an $I \times J$ contingency table, it suffices to calculate weights for the cells, i.e., it is not necessary to calculate separately a weight for each individual observation. For an observation (X_i, Y_i) in cell (a, b), the weight W_i reduces to

$$W_{ab} = p_{ab} \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} h_{\hat{F}_1}(i,a) h_{\hat{F}_2}(j,b)}{\sqrt{\hat{\kappa}(X,X)\hat{\kappa}(Y,Y)}}$$

where p_{ab} is the proportion of observations in cell (a, b). Similarly, the weights belonging to component correlation ρ_{kl} are

$$W_{ab}^{(k,l)} = p_{ab} g_{1k}(a) g_{2l}(b)$$

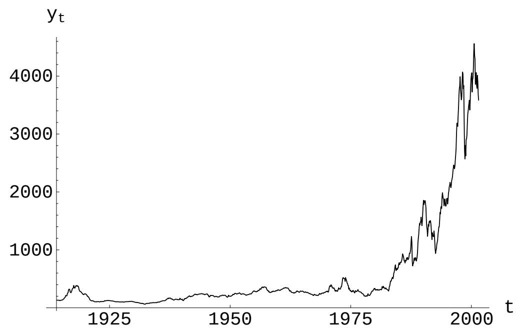
for $a = 1, \ldots, I$, $b = 1, \ldots, J$, $k = 1, \ldots, I$ and $l = 1, \ldots, J$. Note that

$$\rho^* = \sum_{a=1}^{I} \sum_{b=1}^{J} W_{ab}$$

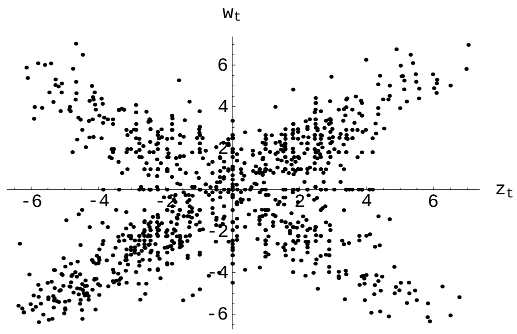
and

$$\rho_{kl} = \sum_{a=1}^{I} \sum_{b=1}^{J} W_{ab}^{(k,l)}$$

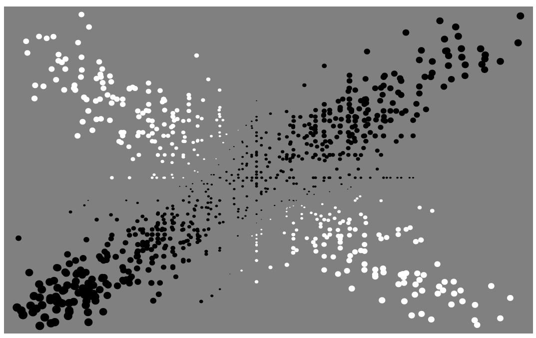
We found that $\hat{\rho}^* = .02$ ($p = .000$), i.e., there is significant association in the data. The weights W_{ab} for the cells are represented in Figure 8(a). Here, the grayscale represents the size of W_{ab}: the darker the cell, the larger W_{ab}; $W_{ab} = 0$ is represented by a fixed intermediate shade of gray. From Figure 8(a), it can be seen that most of the association is of a monotone nature: the higher the parents' socioeconomic status, the better the mental health status of their children. We also investigated the component correlations and found two components to be significant at the 5% level: $\hat{\rho}_{11} = .13$ ($p = .000$) and $\hat{\rho}_{13} = .08$ ($p = .026$). In Figures 8(b) and 8(c) we represented the $W_{ab}^{(11)}$ and $W_{ab}^{(13)}$ using grayscales as above. From Figure 8(b), we see that $\hat{\rho}_{11}$ indicates linearity again. However, in Figure 8(c) we see evidence of some nonlinearity in the data, namely an apparent reversal of the association if only the middle categories 'Mild Symptom Formation' and 'Moderate Symptom Formation' are considered. Hence, it appears that the association which is present in the data cannot be fully explained by linearity.



(a) Time series of stock price index



(b) Scatterplot of (Z_t, W_t) showing association between successive jumps in Figure 9(a). $\hat{\rho}^* = .09$



(c) Weights W_i for data in Figure 9(b)

Fig. 9. *Monthly Norwegian stock price indices, 1914-2001*

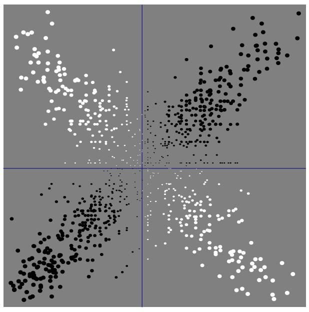
6.3. *Norwegian stock exchange* The Norwegian stock exchange data represented in Figure 9(a) yield an example with an especially interesting form of association. In Figure 9(a) the original time series data Y_t are plotted. In Figure 9(b) we plotted

$$(Z_t, W_t) = (\text{arcsinh}(Y_t - Y_{t-1}), \text{arcsinh}(Y_{t+1} - Y_t))$$

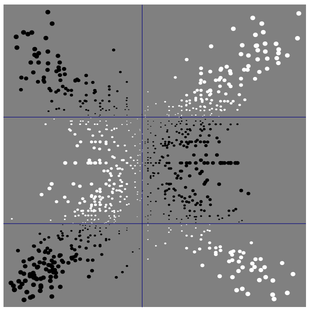
The arcsinh transformation was done to make the marginal distributions less heavy tailed. We found a highly significant association with $\hat{\rho}^*(Z, W) = .09$ ($p = .000$). From Figure 9(b) we can already interpret the association: large jumps (up or down) in stock prices tend to be followed by large jumps, and small jumps by small jumps, indicating periods of volatility. In Figure 9(c), the weights W_i are represented by the size and color of the dots (see Section 6.1 for further explanation). This plot points to the same conclusion that large jumps are followed by large jumps and small jumps by small jumps. Note that in this case, not only the positive weights (the black dots) but also the negative weights (the white dots) are highly indicative of association. The plot indicates that the up-arm (with the black dots) is 'heavier' than the down-arm (with the white dots), that is, there is evidence that in the data generating process a jump tends to be of the same sign as the previous jump.

Seven component correlations were found to be significant at the 5% level after applying the Bonferroni correction: $\hat{\rho}_{11} = .25$, $\hat{\rho}_{12} = .13$, $\hat{\rho}_{22} = .64$, $\hat{\rho}_{24} = .19$, $\hat{\rho}_{33} = .19$, $\hat{\rho}_{44} = .38$ and $\hat{\rho}_{66} = .25$, all with $p = .000$. In Figure 10, the weights corresponding to the component correlations are represented. Figures 10(c), (d), (f) and (g) point to the cross-like nature of the data. Figures 10(a) and (e) indicate that the up-arm is heavier than the down-arm. We did not find a meaningful explanation for Figure 10(b).

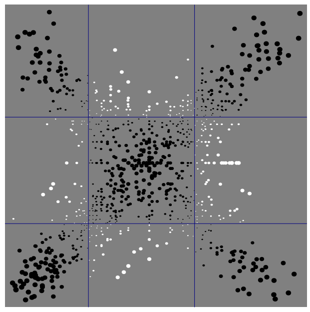
6.4. *Discussion* If a researcher investigating the association between two variables decides on the use of ρ^*, we recommend the following approach. First a test of the significance of $\hat{\rho}^*$ should be done, and, if found to be significant, the weights W_i should be visualized as described above in order to determine the nature of the association. If this does not yield the desired insight, it can be worthwhile to investigate the component correlations and visualize the corresponding weights $W_i^{(k,l)}$. These components form an orthogonal decomposition of the 'infinite-dimensional' object ρ^* into 'one-dimensional' objects ρ_{kl}, and the orthogonality ensures, in a limited sense, that the different components measure different things; by the latter we mean that for large samples and close to independence, the sample component correlations are approximately independent. The question may arise: why not investigate correlations between other sets of orthogonal functions? A sketch of an answer is as follows. Because of the various optimality properties of the eigenfunctions in describing the marginal kernels, these component correlations are likely to be a better choice for investigating the deviation of ρ^* from zero than correlations between arbitrarily chosen functions for the marginal distributions. One way to make this intuitive is as follows: in those regions where the marginal distributions are sparse, the eigenfunctions vary relatively slowly (in second derivative sense, see remark after Lemma 7), and therefore power of a test based on a component correlation will be concentrated in those regions where there are many observations.



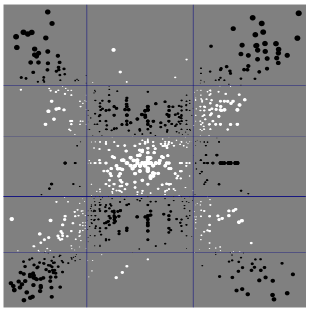
(a) $\hat{\rho}_{11} = .25$



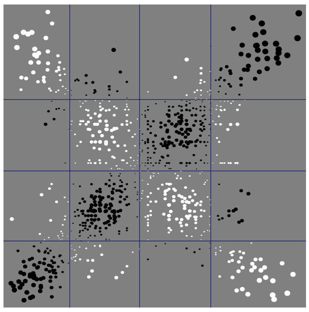
(b) $\hat{\rho}_{12} = .13$



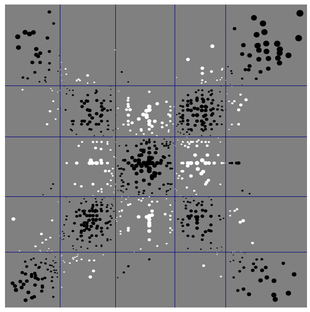
(c) $\hat{\rho}_{22} = .64$



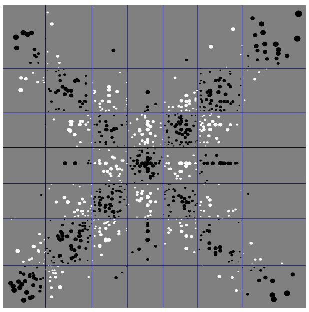
(d) $\hat{\rho}_{24} = .19$



(e) $\hat{\rho}_{33} = .19$



(f) $\hat{\rho}_{44} = .38$



(g) $\hat{\rho}_{66} = .25$

FIG. 10. *Representation of weights contributing to $\hat{\rho}_{ij}$ for data in Figure 9(b)*

Thus, if we believe ρ^* to be a good measure of deviation from independence, then good 'one-dimensional' objects to look at are the correlations between the marginal eigenfunctions of h_{F_1} and h_{F_2}.

REFERENCES

Agarwal, R. P. (1992). *Difference equations and inequalities.* New York: Marcel Dekker.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, *Vol. 7, No. 1*, 131-177.

Agresti, A. (2002). *Categorical Data Analysis, 2nd edition.* New York: Wiley.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory for certain 'goodness of fit' criteria based on stochastic processes. *Ann. Math. Stat.*, *47*, 193-212.

Baglivo, J., Pagano, M., & Spino, C. (1996). Permutation distributions via generating functions, with applications to sensitivity analysis of discrete data. *J. Am. Stat. Ass.*, *91*, 1037-1046.

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis*, *88*, 190-206.

Blum, J. R., Kiefer, J., & Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The annals of mathematical statistics*, *32*, 485-498.

Booth, J. G., & Butler, R. (1999). An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrica*, *86*, 321-332.

Bose, R. C., & Gupta, S. S. (1959). Moments of order statistics from a normal population. *Biometrika*, *46*, 433 - 440.

De Wet, T. (1980). Cramér-von Mises tests for independence. *J. Multivariate Anal.*, *10*, 38-50.

De Wet, T. (1987). Degenerate U- and V-statistics. *South African Statistical Journal*, *21*, 99-129.

De Wet, T., & Venter, J. H. (1973). Asymptotic distributions for quadratic forms with application to tests of fit. *Annals of Statistics*, *1*, 380-387.

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

Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.*, *13*, 363-397.

Durbin, J., & Knott, M. (1972). Components of the Cramér-von Mises statistics, I. *J. R. Statist. Soc. B*, *34*, 260-307.

Eagleson, G. K. (1979). Orthogonal expansions and U-statistics. *Australian Journal of Statistics*, *21*, 221-237.

Forster, J. J., McDonald, J. W., & Smith, P. W. F. (1996). Monte Carlo exact conditional tests for log-linear and logistic models. *J. Roy. Stat. Soc. Ser B*, *58*, 445-453.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.*, *13*, 10-69.

Govindarajulu, Z. (1963). On moments of order statistics and quasi-ranges from normal populations. *Ann. Math. Statist.*, *34*, 633-651.

Gregory, G. G. (1977). Large sample theory for U-statistics. *Annals of Statistics*, *5*, 110-123.

Hall, P. (1979). On the invariance principle for U-statistics. *Stoch. Proc. Appl.*, *9*, 163-174.

Hoeffding, W. (1940). Masstabinvariante Korrelationtheorie. *Schriften Math. Inst Univ. Berlin*, *5*, 181-233.

Hoeffding, W. (1948a). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293-325.

Hoeffding, W. (1948b). A non-parametric test of independence. *Annals of Mathematical Statistics*, *19*, 546-557.

Hoeffding, W. (1961). The strong law of large numbers for U-statistics. *Institute of Statistics, University of North Carolina, Mimeograph Series No. 302.*

Johnson, Kotz, & Balakrishnan. (1994). *Continuous univariate distributions: volume 1.* New York: Wiley.

Kallenberg, W. C. M., & Ledwina, T. (1999). Data driven rank tests for independence. *J. Amer. Statist. Assoc.*, *94*, 285-301.

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Ann. Math. Stat.*, *30*, 420-447.

Mikosch, T. (2006). Copulas: tales and facts (with discussion). *Extremes*, *To appear.*

Nelsen, R. B. (2006). *An introduction to copulas.* New York: Springer.

Neyman, J. (1937). "Smooth" tests for goodness of fit. *Skand. Aktuarietidskr.*, *20*, 150-199.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric statistics.* New York: Wiley.

Rüschendorf, L. (1991). Fréchet bounds and their applications. In G. Dall'Aglio, S. Kotz, & G. Salinetti (Eds.), *Advances in probabability distributions with given marginals* (p. 151-188). Kluwer.

Tricomi, F. G. (1985). *Integral equations.* Dover Publications.

Van de Wiel, M. A., Di Bucchianico, A. D., & Van der Laan, P. (1999). Symbolic computation and exact distributions of nonparametric test statistics. *The statistician*, *48*, 507-516.

Weisstein, E. W. (1999). "Erf" and "Erfi". *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Erfi.html*.

Zaanen, A. (1960). *Linear analysis.* Amsterdam: North Holland Publishing Co.

Zech, G., & Aslan, B. (2003). A multivariate two-sample test based on the concept of minimum energy. *PHYSTAT2003*, 97-100.