Statistical Science
2022, Vol. 37, No. 1, 90–109
https://doi.org/10.1214/21-STS823

Statistical Dependence: Beyond Pearson's ρ

Dag Tjøstheim, Håkon Otneim and Bård Støve

Abstract. Pearson's ρ is the most used measure of statistical dependence. It gives a complete characterization of dependence in the Gaussian case, and it also works well in some non-Gaussian situations. It is well known; however, that it has a number of shortcomings; in particular, for heavy tailed distributions and in nonlinear situations, where it may produce misleading, and even disastrous results. In recent years, a number of alternatives have been proposed. In this paper, we will survey these developments, especially results obtained in the last couple of decades. Among measures discussed are the copula, distribution-based measures, the distance covariance, the HSIC measure popular in machine learning and finally the local Gaussian correlation, which is a local version of Pearson's ρ. Throughout, we put the emphasis on conceptual developments and a comparison of these. We point out relevant references to technical details as well as comparative empirical and simulated experiments. There is a broad selection of references under each topic treated.

Key words and phrases: Statistical dependence, Pearson's ρ, nonlinear dependence, distance covariance, HSIC, mutual information, local Gaussian correlation.

1. INTRODUCTION

Pearson's ρ, the product moment correlation, was not invented by Pearson, but rather by Francis Galton. Galton, a cousin of Charles Darwin, needed a measure of association in his hereditary studies (Galton, 1888, 1890). This was formulated in a scatter diagram and regression context, and he chose r (for regression) as the symbol for his measure of association. Pearson (1896) gave a more precise mathematical development and used ρ as a symbol for the population value and r for its estimated value. The product moment correlation is now universally referred to as Pearson's ρ. Galton died in 1911, and Karl Pearson became his biographer, resulting in a massive four-volume biography (Pearson, 1922, 1930). All of this and much more is detailed in Stigler (1989) and Stanton (2001). Some other relevant historical references are Fisher (1915, 1921), von Neumann (1941, 1942) and the survey paper by King (1987).

Dag Tjøstheim is Professor Emeritus, Department of Mathematics, University of Bergen, P.b. 7803, 5020 Bergen, Norway (e-mail: dag.tjostheim@uib.no). Håkon Otneim is Associate Professor, Department of Business and Management Science, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway (e-mail: hakon.otneim@nhh.no). Bård Støve is Professor, Department of Mathematics, University of Bergen, P.b. 7803, 5020 Bergen, Norway (e-mail: bard.stove@uib.no).

Write the covariance between two random variables X and Y having finite second moments as $\mathrm{Cov}(X, Y) = \sigma(X, Y) = \mathrm{E}(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))$. The Pearson's ρ, or the product moment correlation, is defined by

$$\rho = \rho(X, Y) = \frac{\sigma(X, Y)}{\sigma_X \sigma_Y}$$

with $\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\mathrm{E}(X - \mathrm{E}(X))^2}$ being the standard deviation of X and similarly for σ_Y. The correlation takes values between and including -1 and $+1$. For a given set of pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ of X and Y, an estimate of ρ is given by

$$r = \widehat{\rho} = \frac{\sum_{j=1}^n (X_j - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum_{j=1}^n (X_j - \overline{X})^2} \sqrt{\sum_{j=1}^n (Y_j - \overline{Y})^2}} \tag{1.1}$$

with $\overline{X} = n^{-1} \sum_{j=1}^n X_j$, and similarly for $\overline{Y}$. Consistency and asymptotic normality can be proved using an appropriate law of large numbers and a central limit theorem, respectively.

The correlation coefficient ρ has been, and probably still is, the most used measure for statistical association, and it is generally accepted as *the* measure of dependence, not only in statistics, but in most applications of statistics to the natural and social sciences. There are several reasons for this:

(i) It is easy to compute (estimate).

(ii) Linear models are much used, and in a linear regression model of Y on X, say, ρ is proportional to the slope of the regression line.

(iii) In a bivariate Gaussian density

$$f(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}\sigma_X\sigma_Y}\times\exp\Big\{-\frac{1}{2(1-\rho^2)}\Big(\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\Big)\Big\},$$

the dependence between X and Y is completely characterized by ρ. In particular, two jointly Gaussian variables (X,Y) are independent if and only if they are uncorrelated. For a considerable number of data sets, the Gaussian distribution works at least as a fairly good approximation. Moreover, joint asymptotic normality often appears as a consequence of the central limit theorem for many statistics, and the joint asymptotic behavior of such statistics are therefore generally well defined by the correlation coefficient.

(iv) The product moment correlation is easily generalized to the multivariate case. For p stochastic variables $X_1,\dots,X_p$, their joint dependencies can simply (but not always accurately) be characterized by their covariance matrix $\Sigma=\{\sigma_{ij}\}$, with σ_{ij} being the covariance between X_i and X_j. Similarly the correlation matrix is defined by $\Lambda=\{\rho_{ij}\}$, with ρ_{ij} being the correlation between X_i and X_j. Again, for a column vector $x=(x_1,\dots,x_p)^T$, the joint normality density is defined by

$$f(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\Big\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big\},$$

where $|\Sigma|$ is the determinant of the covariance matrix Σ (whose inverse Σ^{-1} is assumed to exist), and $\mu=\mathrm{E}(X)$. Then the complete dependence structure of the Gaussian vector is given by the *pairwise* covariances σ_{ij}, or equivalently the *pairwise* correlations ρ_{ij}. This is remarkable: the entire dependence structure is determined by pairwise dependencies.

(v) It is easy to extend the correlation concept to time series. For a time series $\{X_t\}$, the autocovariance and autocorrelation function, respectively, are defined, assuming stationarity and existence of second moments, by $c(t)=\sigma(X_{t+s},X_s)$ and $\rho(t)=\rho(X_{t+s},X_s)$ for arbitrary integers s and t. For a Gaussian time series, the dependence structure is completely determined by $\rho(t)$. Even for nonlinear time series and nonlinear regression models, the autocovariance function has often been made to play a major role. In the frequency domain all of the traditional spectral analysis is based again on the autocovariance function.

In spite of these assets, there are several serious weaknesses of Pearson's ρ. These will be briefly reviewed in Section 2. In the remaining sections of this paper, a number of alternative dependence measures going beyond the Pearson ρ will be described. The emphasis will be on concepts, conceptual developments and comparisons of these. We do provide some illustrative plots of key properties, but when it comes to technical details, empirical and simulated experiments with numerical comparisons, we point out relevant references instead.

In Section 3, we briefly review the copula and its use in dependence modeling. Global dependence functionals based on distribution functions and density functions, if they exist, are treated in Section 4, where we cover the distance based functionals such as the distance covariance function and the mutual information criterion as well as related criteria. We also treat the HSIC criterion, which is popular in machine learning, and its relationship to the distance covariance. We discuss all of this in light of seven properties that a dependence criterion ideally should possess according to Rényi (1959). In Section 5, we shift our emphasis to local dependence measures, allowing the statistical dependence to vary across different regions of the support of the distribution functions. In particular, we treat the recently introduced local Gaussian approximation in Section 6. To improve the focus of the paper, some details, especially those related to Section 6, have been moved to an online supplementary note (Tjøstheim, Otneim and Støve, 2022).

2. WEAKNESSES OF PEARSON'S ρ

We have subsumed, somewhat arbitrarily, the problems of Pearson's ρ under three issues.

2.1 The Non-Gaussianity Issue

A natural question to ask is whether the close connection between Gaussianity and correlation/covariance properties can be extended to larger classes of distributions. The answer to this question is a conditional yes. The multivariate Gaussian distribution is a member of the vastly larger class of elliptical distributions. That class of distributions is defined both for discrete and continuous variables, but we limit ourselves to the continuous case. An elliptical distribution can be defined in terms of a parametric representation of the characteristic function or the density function. For an overview of elliptical distributions see, for example, Gómez, Gómez-Villegas, and Marín (2003).

Unfortunately, the equivalence between uncorrelatedness and independence is generally not true for elliptical distributions. Consider, for instance, the multivariate

t-distribution with ν degrees of freedom

$$f(x) = \frac{\Gamma(\frac{p+\nu}{2})}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\Sigma|^{1/2}} \times \left(1 + \frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{\nu}\right)^{-\frac{p+\nu}{2}}. \tag{2.1}$$

Unlike the multinormal distribution where the exponential form of the distribution forces the distribution to factor if Σ is a diagonal matrix (uncorrelatedness), this is not true for the t distribution defined in equation (2.1) if Σ is diagonal. In other words, if two components of a bivariate t distribution are uncorrelated, they are not necessarily independent. This pinpoints a serious deficiency of the Pearson's ρ in measuring dependence in t distributions, and indeed in general elliptical (and, of course, nonelliptical) distributions.

2.2 The Robustness Issue

As is the case for regression, it is well known that the product moment estimator is sensitive to outliers. Even just one single outlier may be very damaging. There are therefore several robustified versions of ρ, primarily based on ranks. The idea of rank correlation goes back at least to Spearman (1904), and it is most easily explained through its sample version. Given scalar observations $\{X_1, \ldots, X_n\}$, we denote by $R_{i,X}^{(n)}$ the rank of X_i among $X_1, \ldots, X_n$. (There are various rules for treating ties.) The estimated Spearman rank correlation function given n pairwise observations of two random variables X and Y is given by

$$\widehat{\rho}_S = \frac{n^{-1}\sum_{i=1}^n R_{i,X}^{(n)} R_{i,Y}^{(n)} - (n+1)^2/4}{(n^2-1)/12}.$$

The rank correlation is thought to be especially effective in picking up linear trends in the data, but it suffers in a very similar way as the Pearson ρ to certain nonlinearities of the data, which are treated in the next subsection.

Another way of using the ranks is Kendall's τ rank correlation coefficient given by Kendall (1938). Again, consider the situation of n pairs (X_i, Y_i) of the random variables X and Y. Two pairs of observations (X_i, Y_i) and (X_j, Y_j), $i \neq j$ are said to be concordant if the ranks for both elements agree; that is, if both $X_i > X_j$ and $Y_i > Y_j$ or if both $X_i < X_j$ and $Y_i < Y_j$. Similarly, they are said to be discordant if $X_i > X_j$ and $Y_i < Y_j$ or if $X_i < X_j$ and $Y_i > Y_j$. If one has equality, they are neither concordant nor discordant, even though there are various rules for treating ties in this case as well. The estimated Kendall τ is then given by

$$\widehat{\tau} = \big((\text{number of concordant pairs}) - (\text{number of discordant pairs})\big)/\big(n(n-1)/2\big).$$

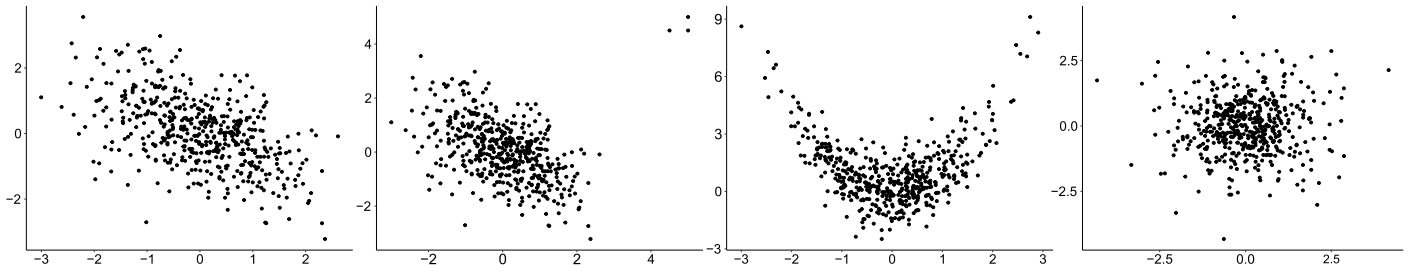
We will illustrate the robustness issue using a simple example. In Figure 1(a), we see 500 observations that have been simulated from the bivariate Gaussian distribution having correlation $\rho = -0.5$. The sample value for Pearson's ρ is $\widehat{\rho} = -0.53$. If we add just three outliers to the data, however, as shown in Figure 1(b), the sample correlation changes to $\widehat{\rho} = -0.36$. The sample versions of Spearman's ρ for the simulated data in Figures 1(a) and 1b are on the other hand very similar: $\widehat{\rho}_S = -0.52$ and $\widehat{\rho}_S = -0.49$, and the corresponding values for the estimated Kendall's τ are $\widehat{\tau} = -0.37$ and $\widehat{\tau} = -0.35$.

2.3 The Nonlinearity Issue

This is probably the most serious issue with Pearson's ρ, and it is an issue also for the rank based correlations of Spearman and Kendall. All of these (and similar measures), are designed to detect rather specific types of statistical dependencies, namely those for which large values of X tend to be associated with large values of Y, and small values of X with small values of Y (positive dependence), or the opposite case of negative dependence in which large values of one variable tend to be associated with small values of the other variable. It is easy to find examples where this is not the case, but where nevertheless there is strong dependence. A standard introductory text book example is the case where

$$Y = X^2. \tag{2.2}$$

Here, Y is uniquely determined once X is given; that is, basically the strongest form of dependence one can have. Nevertheless, if X has a symmetric distribution on the real



(a) Gaussian (b) Gaussian with outliers (c) Nonlinear (d) Garch

FIG. 1. *Illustration of some problems related to the Pearson correlation coefficient.*

line, and if sufficient moments exist in the case of ρ, then the population quantities corresponding to these three test statistics are all zero. A version of this situation is illustrated in Figure 1(c), where we have generated 500 observations of the standard normally distributed independent variables X and ϵ, and calculated Y as $Y = X^2 + \epsilon$. Still, $\rho(X, Y) = 0$. The sample values for Pearson's ρ, Spearman's ρ_S and Kendall's τ are $\widehat{\rho} = -0.001$, $\widehat{\rho}_S = -0.03$ and $\widehat{\tau} = -0.02$, respectively, and none of them are significantly different from zero. It may be noted that there is a consistent nonnegative modification of Kendall's τ; see Bergsma and Dassios (2014).

Essentially the same problem will occur if $X = UW$ and $Y = VW$, where U and V are independent of each other and independent of W. It is trivial to show that $\rho(X, Y) = 0$ if $\mathrm{E}(U) = \mathrm{E}(V) = 0$, whereas X and Y are clearly dependent. This example typifies the kind of dependence one has in ARCH/GARCH time series models: If $\{\varepsilon_t\}$ is a time series of zero-mean i.i.d. variables and if the non-negative time series $\{h_t\}$ is independent of $\{\varepsilon_t\}$, and $\{X_t\}$ and $\{h_t\}$ are given by the recursive relationship

$$X_t = \varepsilon_t h_t^{1/2}, \quad h_t = \alpha + \beta h_{t-1} + \gamma X_{t-1}^2, \tag{2.3}$$

where the stochastic process $\{h_t\}$ is the so-called volatility process, then the resulting model is a GARCH(1, 1) model. Further, $\alpha > 0$, and β and γ are nonnegative constants satisfying $\beta + \gamma < 1$. This model can be extended in many ways and the ARCH/GARCH models are extremely important in finance. A comprehensive book is Francq and Zakoïan (2011). The work on these kinds of models was initiated by Engle (1982), and he was awarded the Nobel Memorial Prize in Economic Sciences for his work. The point as far as Pearson's ρ is concerned is that X_t and X_s are uncorrelated for $t \neq s$, but they are in fact strongly dependent through the volatility process $\{h_t\}$, which can be taken to measure financial risk.

In Figure 1(d), we see some simulated data from a GARCH(1, 1)-model with $\epsilon_t \sim$ i.i.d. $N(0, 1)$, $\alpha = 0.1$, $\beta = 0.7$ and $\gamma = 0.2$, with X_t on the horizontal axis, and X_{t-1} on the vertical axis. In this particular case, $\widehat{\rho}(X_t, X_{t-1}) = 0.018$, despite the strong serial dependence that is seen to exist directly from equation (2.3).

In the following sections, we will look at ways of detecting nonlinear and non-Gaussian structures by going beyond Pearson's ρ.

3. BEYOND PEARSON'S ρ: THE COPULA

For two variables, one may ask, why not just take the joint density function $f(x, y)$ or the cumulative distribution function $F(x, y)$ as a descriptor of the joint dependence? The answer is quite obvious. If a parametric density model is considered, it is usually quite difficult to give an interpretation of the parameters in terms of the strength of the dependence. If one looks at nonparametric estimates for multivariate density functions, to a certain degree one may get an informal indication of strength of dependence in certain regions from a display of the density, but the problems increase quickly with dimension due both to difficulties of producing a graphical display and to the lack of precision of the estimates due to the curse of dimensionality. Another problem in analyzing a joint density function is that it may be difficult to disentangle effects due to the shape of marginal distributions and effects due to dependence among the variables involved.

This last problem is resolved by the copula construction. Sklar's (1959) theorem states that a multivariate cumulative distribution function $F(x) = F(x_1, \ldots, x_p)$ with marginals $F_i(x_i)$, $i = 1, \ldots, p$ can be decomposed as

$$F(x_1, \ldots, x_p) = C\big(F_1(x_1), \ldots, F_p(x_p)\big), \tag{3.1}$$

where $C(u_1, \ldots, u_p)$ is a distribution function over the unit cube $[0, 1]^p$. Klaassen and Wellner (1997) point out that Hoeffding (1940) had the basic idea of summarizing the dependence properties of a multivariate distribution by its associated copula, but he chose to define the corresponding function on the interval $[-1/2, 1/2]$ instead of on the interval $[0, 1]$. In the continuous case, C is a function of uniform variables $U_1, \ldots, U_p$, using the well-known fact that for a continuous random variable X_i, $F_i(X_i)$ is uniform on $[0, 1]$. Further, in the continuous case C is uniquely determined by Sklar's (1959) theorem.

The theorem continues to hold for discrete variables under certain regularity conditions securing uniqueness. We refer to Nelsen (1999) and Joe (2014) for extensive treatments of the copula. Joe (2014), in particular, contains a large section on copulas in the discrete case. See also Genest and Nešlehová (2007).

The decomposition (3.1) very effectively disentangle the distributional properties of a multivariate distribution into a dependence part measured by the copula C and a marginal part described by the univariate marginals. Note that C is invariant with respect to one-to-one transformations of the marginal variables X_i. In this respect, it is analogous to the invariance of the Kendall and Spearman rank based correlation coefficients.

The decomposition in (3.1) is very useful in that it leads to large classes of models that can be specified by defining the marginals and the copula function separately. It has great flexibility in that very different models can be chosen for the marginal distribution, and there is a large catalog of possible parametric models available for the copula function C; it can also be estimated nonparametrically. In particular, the Clayton copula has been important in economics and finance. It is defined by

$$C_C(u_1, u_2) = \max\{u_1^{-\theta} + u_2^{-\theta} - 1; 0\}^{-1/\theta} \quad \text{with } \theta \in [-1, \infty)/0 \tag{3.2}$$

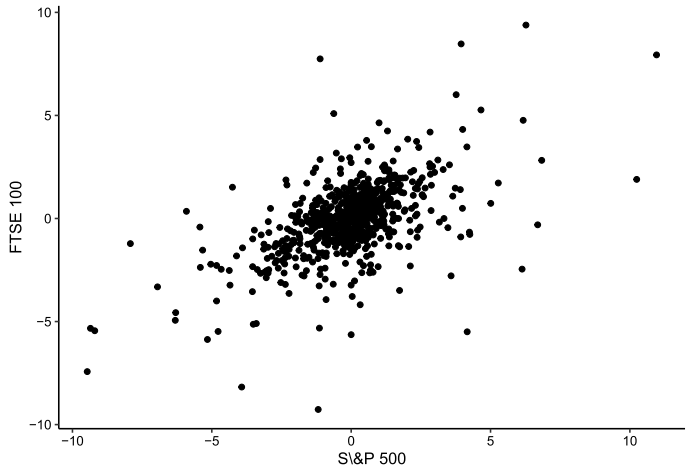
in the two-dimensional case. It can be extended to higher dimensions. For a connection between Kendall's τ and the copula parameter θ, see Genest et al. (2011).

We will throughout this paper and its supplement illustrate several points using a bivariate data set on some financial returns. We use daily international equity price index data for the United States (i.e., the S&P 500) and the United Kingdom (i.e., the FTSE 100). The data are obtained from Datastream (2018), and the returns are defined as

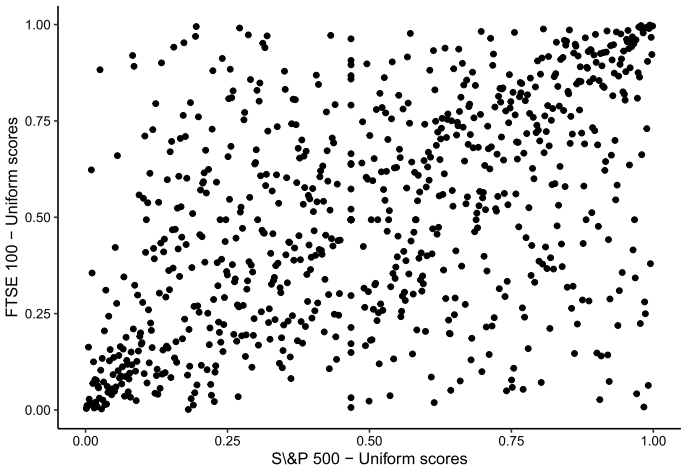
$$r_t = 100 \times \big(\log(p_t) - \log(p_{t-1})\big),$$

where p_t is the price index at time t. The observation span covers the period from January 1, 2007, through December 31, 2009, in total 784 observations. In Figure 2, four scatterplots are presented.

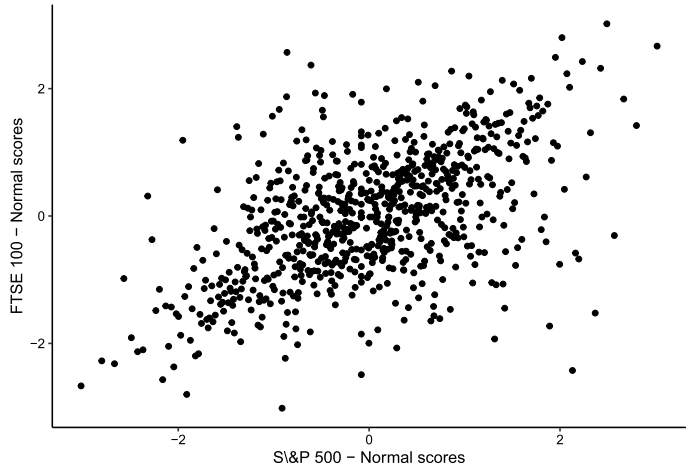
Figure 2(a) displays a scatterplot of the observed log-returns with S&P 500 on the horizontal axis, and the FTSE 100 on the vertical axis. Figure 2(b) displays the uniform scores of the same data, that is, $(\widehat{U}_{1i}, \widehat{U}_{2i}) = (\widehat{F}_1(X_{1i}), \widehat{F}_2(X_{2i}))$ where $\widehat{F}_1$ and $\widehat{F}_2$ are the empirical distribution functions of X_1: S&P500 and X_2: FTSE 100, and we see indications of a somewhat cluttered behaviour of the scores in the lower left and upper right corners of the unit square corresponding to the tails of the joint distribution. The plot in Figure 2(b) then is a scatter diagram of the copula dependence function C in the formula (3.1) when $p = 2$. In Figure 2(c), the observations have been transformed to normal scores given by $\widehat{Z}_i = \Phi^{-1}(\widehat{F}(X_i))$, $i = 1, 2$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal. The transformation to a standard normal scale plays an important role in the theory and application of the local Gaussian correlation measure; see, in particular, the discussion following equation (6.4). For now, it is sufficient to note that it more clearly reveals the tail properties of the underlying distribution



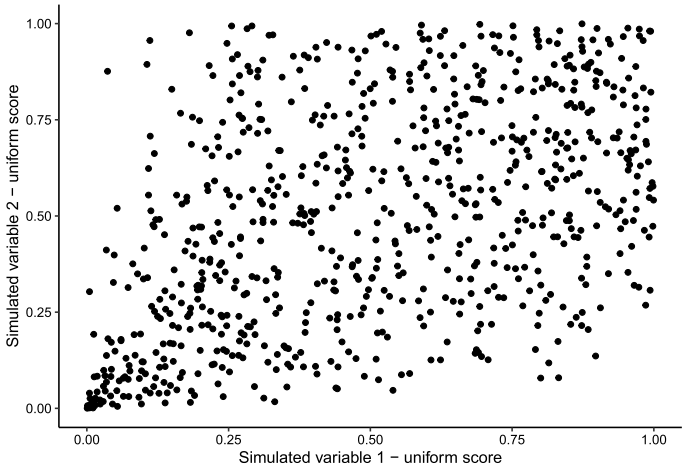
(a) The observed log-returns of the daily data



(b) Uniform scores of the financial returns data set



(c) Normal scores of the financial returns data set



(d) Simulated data from a Clayton copula fitted to the financial returns data set

FIG. 2. *Illustrations using the financial returns data set.*

than is the case of Figure 2(b). Finally, Figure 2(d) shows the scatterplot of 784 simulated pairs of variables, on uniform scale, from a Clayton copula fitted to the return data. The plot resembles Figure 2(b), in particular in the lower left corner. However, there are some differences in the upper right corner. We will look into this discrepancy in Section 6, and in Section 2 of the supplement.

In Figure 2(a), and perhaps more clearly in Figure 2(c), we see that there seems to be stronger dependence between the variables when the market is going either up or down, which is very sensible from an economic point of view, but it is not easy to give an interpretation of the parameter θ of the Clayton copula in terms of such type of dependence. In fact, in this particular case, $\widehat{\theta} = 0.96$. The difficulty of giving a clear and concrete interpretation of copula parameters in terms of measuring strength of dependence can be stated as a potential issue of the copula representation. In this respect, it is very different from Pearson's ρ. We will return to this point in Section 6, where we define a local correlation.

Another issue of the original copula approach has been the lack of good practical models as the dimension increases, as it would, for example, in a portfolio problem in finance. This has recently been sought solved by the so-called pair copula construction. To simplify, in a trivariate density $f(x_1, x_2, x_3)$, by conditioning this can be written $f(x_1, x_2, x_3) = f_1(x_1) f_{23|1}(x_2, x_3|x_1)$, and a bivariate copula construction, for example, a Clayton copula, can be applied to the conditional density $f_{23|1}(x_2, x_3|x_1)$ with x_1 fixed and with a parameter $\theta = \theta(x_1)$ depending on x_1. This conditioning can be extended to higher dimensions under a few simplifying assumptions, resulting in a so-called vine copula, of which there are several types. The procedure is well described by Aas et al. (2009), and has found a number of applications.

4. BEYOND PEARSON'S ρ: GLOBAL DEPENDENCE FUNCTIONALS AND TESTS OF INDEPENDENCE

Studies of statistical dependence may be said to center mainly around two problems: (i) definition and estimation of measures of dependence and (ii) tests of independence. Of course, these two themes are closely related. Measures of association such as Pearson's ρ can also be used in tests of independence, or more precisely: tests of uncorrelatedness. On the other hand, test functionals for tests of independence can in many, but not all, cases be used as a measures of dependence. A disadvantage with measures derived from tests is that they are virtually always based on a distance function and, therefore, nonnegative. This means that they cannot distinguish between negative and positive dependence.

Most of the test functionals are based on the definition of independence in terms of cumulative distribution functions or in terms of density functions. Consider p stochastic variables $X_1, \ldots, X_p$. These variables are independent if and only if their joint cumulative distribution function is the product of the marginal distribution functions: $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = F_1(x_1) \cdots F_p(x_p)$, and the same is true for all subsets of variables of $(X_1, \ldots, X_p)$. If the variables are continuous, this identity can be phrased in terms of the corresponding density functions instead. A typical test functional is then designed to measure the distance between the estimated joint distributions/densities and the product of the estimated marginals. One would usually estimate the involved distributions non- or semiparametrically, which, for joint distributions, may be problematic for moderate and large p's due to the curse of dimensionality. We will treat these problems in some detail in Sections 4.2–4.5.

Before starting on the description of the various dependence measures, let us remark that Rényi (1959) proposed that a measure of dependence $\delta(X, Y)$ between two stochastic variables X and Y should ideally have the following seven properties:

(i) $\delta(X, Y)$ is defined for any X, Y neither of which is constant with probability 1.
(ii) $\delta(X, Y) = \delta(Y, X)$.
(iii) $0 \le \delta(X, Y) \le 1$.
(iv) $\delta(X, Y) = 0$ if and only if X and Y are independent.
(v) $\delta(X, Y) = 1$ if either $X = g(Y)$ or $Y = f(X)$, where f and g are measurable functions.
(vi) If the Borel-measurable functions f and g map the real axis in a one-to-one way to itself, then $\delta(f(X), g(Y)) = \delta(X, Y)$.
(vii) If the joint distribution of X and Y is normal, then $\delta(X, Y) = |\rho(X, Y)|$, where $\rho(X, Y)$ is Pearson's ρ.

The product moment correlation ρ satisfies only (ii) and (vii).

One can argue that the rules (i)–(vii) do not take into account the difference between positive and negative dependence; it only looks at the strength of the measured dependence. If this wider point of view were to be taken into account, (iii) could be changed into (iii′): $-1 \le \delta(X, Y) \le 1$, (v) into (v′): $\delta(X, Y) = 1$ or $\delta(X, Y) = -1$ if there is a deterministic relationship between X and Y. Finally, (vii) should be changed into (vii′) requiring $\delta(X, Y) = \rho(X, Y)$. Moreover, some will argue that property (vi) may be too strong to require. It means that the strength of dependence is essentially independent of the marginals as for the copula case.

We will discuss these properties as we proceed in the paper. Before we begin surveying the test functionals as announced above, we start with the maximal correlation which, it will be seen, is intertwined with at least one of the test functionals to be presented in the sequel.

4.1 Maximal Correlation

The maximal correlation is based on the Pearson ρ. It is constructed to avoid the problem demonstrated in Section 2.3 that ρ can easily be zero even if there is strong dependence.

It seems that the maximal correlation was first introduced by Gebelein (1941). He introduced it as

$$S(X,Y)=\sup_{f,g}\rho\big(f(X),g(Y)\big),$$

where ρ is Pearson's ρ. Here, the supremum is taken over all Borel-measurable functions f, g with finite and positive variance for $f(X)$ and $g(Y)$. The measure S gets rid of the nonlinearity issue of ρ. It is not difficult to check that $S=0$ if and only if X and Y are independent, and in fact all of the Renyi's seven criteria hold for the maximum correlation; see Lancaster (1957) for property (vii). On the other hand, S cannot distinguish between negative and positive dependence, and it is in general difficult to compute.

Two more recent publications are Huang (2010), where the maximal correlation is used to test for conditional independence, and Yenigün, Székely and Rizzo (2011), where it is used to test for independence in contingency tables. The latter paper introduces a new example where $S(X,Y)$ can be explicitly computed.

4.2 Measures and Tests Based on the Distribution Function

We start with, and in fact put the main emphasis on, the bivariate case. Let X and Y be stochastic variables with cumulative distribution functions F_X and F_Y. The problem of measuring the dependence between X and Y can then be formulated as a problem of measuring the distance between the joint cumulative distribution function $F_{X,Y}$ of (X,Y) and the distribution function F_XF_Y formed by taking the product of the marginals. Let $\Delta(\cdot,\cdot)$ be a candidate for such a distance functional. It will be assumed that Δ is a metric, and it is natural to require (see, e.g., Skaug and Tjøstheim, 1996), that

$$\begin{aligned}&\Delta(F_{X,Y},F_XF_Y)\ge 0\\ \text{(4.1)}\quad\text{and}\quad&\Delta(F_{X,Y},F_XF_Y)=0\quad\text{if and only if}\\ &F_{X,Y}=F_XF_Y.\end{aligned}$$

Clearly, such a measure is capable only of measuring the strength of dependence, not its direction.

A natural estimate $\widehat{\Delta}$ of a distance functional Δ is obtained by setting

$$\widehat{\Delta}(F_{X,Y},F_XF_Y)=\Delta(\widehat{F}_{X,Y},\widehat{F}_X\widehat{F}_Y),$$

where $\widehat{F}$ may be taken to be the empirical distribution functions given by

$$\widehat{F}_X(x)=\frac{1}{n}\sum_{j=1}^{n}1(X_j\le x),\qquad \widehat{F}_Y(y)=\frac{1}{n}\sum_{j=1}^{n}1(Y_j\le y)$$

and

$$\widehat{F}_{X,Y}(x,y)=\frac{1}{n}\sum_{j=1}^{n}1(X_j\le x)1(Y_j\le y),$$

for given observations $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$.

Conventional distance measures between two distribution functions F and G are the Kolmogorov–Smirnov distance

$$\Delta_1(F,G)=\sup_{(x,y)}\big|F(x,y)-G(x,y)\big|$$

and the Cramér–von Mises type distance of a distribution G from a distribution F

$$\Delta_2(F,G)=\int\big\{F(x,y)-G(x,y)\big\}^2\,\mathrm{d}F(x,y).$$

Here, both Δ_1 and Δ_2 satisfy (4.1).

Most of the work pertaining to measuring dependence and testing of independence has been done in terms of the Cramér–von Mises distance. This work started already by Hoeffding (1948) who looked at i.i.d. pairs (X_i,Y_i), and studied finite sample distributions in some special cases. With considerable justification, it has been named the Hoeffding-functional by some. This work was continued by Blum, Kiefer and Rosenblatt (1961) who provided an asymptotic theory, still for the i.i.d. case. It was extended to the time series case with a resulting test of serial independence in Skaug and Tjøstheim (1993a). A paper using a copula framework is Kojadinovic and Holmes (2009). We will briefly review the time series case in an online supplement that accompanies this paper because it illustrates some of the problems, and because some of the same ideas as for the Hoeffding-functional have been used in more recent work on the distance covariance, which we treat in Section 4.3.

As mentioned in the beginning of this section, an independence test for $p>2$ should test the cummulative distribution funcion for all subsets of $X_1,\dots,X_p$. Deheuvels (1981a, 1981b) does exactly that using the Möbius transformation. A recent follow-up is Ghoudi and Rémillard (2018).

Instead of stating independence in terms of cumulative distribution functions this can alternatively be expressed in terms of the characteristic function. Székely, Rizzo and Bakirov (2007) and Székely and Rizzo (2009), as will be seen in Section 4.3, make systematic use of this in their introduction of the distance covariance test. Two random variables X and Y are independent if and only if the characteristic functions satisfy

$$\phi_{X,Y}(u,v)=\phi_X(u)\phi_Y(v)\quad\forall(u,v),$$

where

$$\phi_{X,Y}(u,v)=\mathrm{E}\big(e^{iuX+ivY}\big),\quad \phi_X(u)=\mathrm{E}\big(e^{iuX}\big),$$
$$\phi_Y(v)=\mathrm{E}\big(e^{ivY}\big).$$

This was exploited by Csörgö (1985) and Pinkse (1998) to construct tests for independence based on the characteristic function in the i.i.d. and time series case, respectively. Further work on testing of conditional independence was done by Su and White (2007). Hong (1999) put this into a much more general context by focussing on

$$\sigma_k(u,v) = \phi_{X_t,X_{t-|k|}}(u,v) - \phi_{X_t}(u)\phi_{X_{t-|k|}}(v).$$

By taking Fourier transform of this quantity, one obtains

$$f(\omega,u,v) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\sigma_k(u,v)e^{-ik\omega}. \tag{4.2}$$

Hong (1999) called (4.2) the generalized spectral density function and based a test of serial independence on this. More work related to this has been done by Escanciano and Velasco (2006). Some related ideas can be found in Hong (2000), and more recently in Escanciano and Hualde (2019).

4.3 Distance Covariance

We have seen that there are at least two ways of constructing functionals that are consistent against all forms of dependence, namely those based on the empirical distribution function initiated by Hoeffding (1948) and briefly reviewed above, and those based on the characteristic function represented by Csörgö (1985) in the i.i.d. case and Pinkse (1998) in the serial dependence case, and continued in Hong (1999, 2000) in a time series generalized spectrum approach. Both Pinkse and Hong use a kernel type weight function in their functionals.

The authors of two remarkable papers, Székely, Rizzo and Bakirov (2007) and Székely and Rizzo (2009), take up the characteristic function test statistic again in the nontime series case. But what distinguishes these from earlier papers is an especially judicious choice of weight function reducing the empirical characteristic function functional to empirical moments of differences between the variables, or distances in the vector case, this leading to covariance of distances. Some of these ideas go back to what the authors term an "energy statistic"; see Székely (2002), Székely and Rizzo (2013). It has been extended to time series and multiple dependencies by Davis et al. (2018), Fokianos and Pitsillou (2017), Zhou (2012) and Dueck et al. (2014), and Yao, Zhang and Shao (2018). In the locally stationary time series case, there is also a theory; see Jentsch et al. (2020). The distance covariance, dcov, seems to work well in a number of situations, and it has been used as a yardstick by several authors writing on dependence and tests of independence. In particular, it has been used as a measure of comparison in the work on local Gaussian correlation to be detailed in Section 6 and the supplementary material. There are also points of contacts, as will be seen in Section 4.4, with the HSIC measure of dependence popular in the machine learning community.

The central ideas and derivations are more or less all present in Székely, Rizzo and Bakirov (2007). The framework is that of pairs of i.i.d. vector variables (X,Y) in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, and the task is to construct a test functional for independence between X and Y. Let $\phi_{X,Y}(u,v) = \mathrm{E}(e^{i(\langle X,u\rangle+\langle Y,v\rangle)})$, $\phi_X(u) = \mathrm{E}(e^{i\langle X,u\rangle})$ and $\phi_Y(v) = \mathrm{E}(e^{i\langle Y,v\rangle})$ be the characteristic functions involved, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. The starting point is again the weighted characteristic functional

$$\mathcal{V}^2(X,Y;w) = \int_{\mathbb{R}^{p+q}}\left|\phi_{X,Y}(u,v) - \phi_X(u)\phi_Y(v)\right|^2 \times w(u,v)\,\mathrm{d}u\,\mathrm{d}v, \tag{4.3}$$

where w is a weight function to be chosen. Similarly, one defines

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}}\left|\phi_{X,X}(u,v) - \phi_X(u)\phi_X(v)\right|^2 \times w(u,v)\,\mathrm{d}u\,\mathrm{d}v \tag{4.4}$$

and $\mathcal{V}^2(Y;w)$. The distance correlation, dcor, is next defined by, assuming $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$,

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}.$$

These quantities can be estimated by the empirical counterparts given n observations of the vector pair (X,Y) with

$$\mathcal{V}_n^2(X,Y;w) = \int_{\mathbb{R}^{p+q}}\left|\phi_{X,Y}^n(u,v) - \phi_X^n(u)\phi_Y^n(v)\right|^2 \times w(u,v)\,\mathrm{d}u\,\mathrm{d}v, \tag{4.5}$$

where, for a set of observations $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$ the empirical characteristic functions are given by

$$\phi_{X,Y}^n(u,v) = \frac{1}{n}\sum_{k=1}^n \exp\{i(\langle X_k,u\rangle + \langle Y_k,v\rangle)\}$$

and

$$\phi_X^n(u) = \frac{1}{n}\sum_{k=1}^n \exp\{i\langle X_k,u\rangle\},$$

$$\phi_Y^n(v) = \frac{1}{n}\sum_{k=1}^n \exp\{i\langle Y_k,v\rangle\}.$$

It turns out that it is easier to handle the weight function in the framework of the empirical characteristic functions. It will be seen below that

$$w(u,v) = \left(c_p c_q |u|_p^{1+p}|v|_q^{1+q}\right)^{-1} \tag{4.6}$$

is a good choice. Here, $|\cdot|_p$ is the Euclidean norm in $\mathbb{R}^p$ and similarly for $|\cdot|_q$. Moreover, the normalizing constants are given by $c_j = \pi^{(1+j)/2}/\Gamma((1+j)/2)$, $j = p,q$.

For it to make sense to introduce the weight function on the empirical characteristic function, one must show that the empirical functionals $\mathcal{V}_n$ converges to the theoretical functionals $\mathcal{V}$ for this weight function. This is not trivial because of the singularity at 0 for w given by (4.6). A detailed argument is given in the proof of Theorem 2 in Székely, Rizzo and Bakirov (2007).

The advantage of introducing the weight function for the empirical characteristic functions is that one can compute the squares in (4.5) and then interchange summation and integration. The resulting integrals can be computed using trigonometric identities. The details are given in the proof of Theorem 1 in Székely, Rizzo and Bakirov (2007) and in Lemma 1 of the Appendix of Szekely and Rizzo (2005) who in turn refer to Prudnikov, Brychkov and Marichev (1986) for the fundamental lemma

$$\int_{\mathbb{R}^d} \frac{1-\cos\langle x,u\rangle}{|u|_d^{d+\alpha}}\,\mathrm{d}u = C(d,\alpha)|x|_d^{\alpha}$$

for $0<\alpha<2$ with

$$C(d,\alpha)=\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)}, \tag{4.7}$$

and where the weight function considered above corresponds to $\alpha=1$ and $d=p$ or $d=q$ in (4.6). The general α-case corresponds to a weight function

$$w(u,v;\alpha)=\big(C(p,\alpha)C(q,\alpha)|u|_p^{p+\alpha}|v|_q^{q+\alpha}\big)^{-1}.$$

With the simplification $\alpha=1$ all of this implies that $\mathcal{V}_n^2$ as defined in (4.5), can be computed as

$$\mathcal{V}_n^2(u,v)=S_1+S_2-2S_3,$$

where

$$S_1=\frac{1}{n^2}\sum_{k,l=1}^{n}|X_k-X_l|_p|Y_k-Y_l|_q,$$

$$S_2=\frac{1}{n^2}\sum_{k,l=1}^{n}|X_k-X_l|_p\frac{1}{n^2}\sum_{k,l=1}^{n}|Y_k-Y_l|_q$$

and

$$S_3=\frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n}|X_k-X_l|_p|Y_k-Y_m|_q, \tag{4.8}$$

which explains the appellation distance covariance. In fact, it is possible to further simplify this by introducing

$$a_{kl}=|X_k-X_l|_p,\quad \overline{a}_{k.}=\frac{1}{n}\sum_{l=1}^{n}a_{kl},\quad \overline{a}_{.l}=\frac{1}{n}\sum_{k=1}^{n}a_{kl},$$

$$\overline{a}_{..}=\frac{1}{n^2}\sum_{k,l=1}^{n}a_{kl},\quad A_{kl}=a_l-\overline{a}_{k.}-\overline{a}_{.l}+a_{..},$$

for $k,l=1,\dots,n$. Similarly, one can define $b_{kl}=|Y_k-Y_l|_q$ and $B_{kl}=b_{kl}-\overline{b}_{k.}-\overline{b}_{.l}+\overline{b}_{..}$ and

$$\mathcal{V}_n^2(X,Y)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}B_{kl}$$

and

$$\mathcal{V}_n^2(X)=\mathcal{V}_n^2(X,X)=\frac{1}{n^2}\sum_{k,l=1}^{n}A_{kl}^2$$

and similarly for $\mathcal{V}_n^2(Y)$. From this, one can easily compute $\mathcal{R}_n^2(X,Y)$. The computations are available in the R-package `energy` by Rizzo and Szekely (2018).

As is the case of the empirical joint distribution functional, it can be expected that the curse of dimensionality will influence the result for large and moderate values of p and q. Obviously, in the time series case, it is possible to base oneself on pairwise distances, which has been done in Yao, Zhang, and Shao (2018).

Letting $n\to\infty$, it is not difficult to prove that an alternative expression for $\mathcal{V}(X,Y)$ is given by (assuming $\mathrm{E}|X|_p<\infty$ and $\mathrm{E}|Y|_q<\infty$)

$$\begin{aligned}\mathcal{V}^2(X,Y)&=\mathrm{E}_{X,X',Y,Y'}\big\{|X-X'|_p|Y-Y'|_q\big\}\\&\quad+\mathrm{E}_{X,X'}\big\{|X-X'|_p\big\}\mathrm{E}_{Y,Y'}\big\{|Y-Y'|_q\big\}\\&\quad-2\mathrm{E}_{X,Y}\big\{\mathrm{E}_{X'}|X-X'|_p\mathrm{E}_{Y'}|Y-Y'|_q\big\},\end{aligned} \tag{4.9}$$

where (X,Y), (X',Y') are i.i.d. This expression will be useful later in Section 4.4 in a comparison with the HSIC statistic. Properly scaled $\mathcal{V}_n^2$ has a limiting behavior under independence somewhat similar to that described in Theorem 2 of Skaug and Tjøstheim (1993a); see also equation (1.2) in Section 1 in the online supplement. One can also obtain an empirical process limit theorem, Theorem 5 of Székely, Rizzo and Bakirov (2007). In the R-package `energy`, as for the case of the empirical distribution function, it has been found advantageous to rely on re-sampling via permutations. This is quite fast since the algebraic formulas (4.8) are especially amenable to permutations. Both Székely, Rizzo and Bakirov (2007) and Székely and Rizzo (2009) in their experiments only treat the case of $\alpha=1$ in (4.7).

Turning to the properties (i)–(vii) of Rényi (1959) listed in the beginning of this section, it is clear that (i)–(iv) are satisfied by $\mathcal{R}$. Moreover, according to Székely, Rizzo and Bakirov (2007), if $\mathcal{R}_n(x;y)=1$, then there exists a vector α, a nonzero real number β and an orthogonal matrix C such that $Y=\alpha+\beta XC$, which is not quite the same as Rényi's requirement (v). The dcov measure, being a correlation based measure, in general depends on the distribution of the margins, and hence Rényi's invariance property (vi) does not hold in general; see also Berrett and Samworth (2019), Section 2.1. The final criterion (vii) of Rényi is that the dependent measure should reduce to the

absolute value of Pearson's ρ in the bivariate normal case. This is not quite the case for the dcov, but it comes close, as is seen from Theorem 6 of Székely and Rizzo (2009). In fact, if (X, Y) is bivariate normal with $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$ and with correlation ρ, then $\mathcal{R}(X, Y) \le |\rho|$ and

$$\inf_{\rho \neq 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \lim_{\rho \to 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \frac{1}{2(1+\pi/3-\sqrt{3})^{1/2}} \approx 0.891.$$

4.4 The HSIC Measure of Dependence

Recall the definition and formula for the maximal correlation. This, as stated in Section 4.1, gives rise to a statistic $S(X, Y)$, where $S(X, Y) = 0$ if and only if X and Y are independent. But it is difficult to compute since it requires the supremum of the correlation $\rho(f(X), g(Y))$ taken over Borel-measurable f and g. In the framework of reproducing kernel Hilbert spaces (RKHS) it is possible to pose this problem, or an analogous one, much more generally, and one can compute an analogue of S quite easily. This yields the so-called HSIC (Hilbert–Schmidt Independence Criterion).

Reproducing kernel Hilbert spaces are very important tools in mathematics as well as in statistics. A general reference to applications in statistics is Berlinet and Thomas-Agnan (2004). In the last decade or so, there has also been a number of uses of RKHS in dependence modeling. These have often, but not always, been published in the machine learning literature; see, for example, Gretton and Györfi (2010), Gretton and Györfi (2012), Sejdinovic et al. (2013) and Pfister et al. (2018).

We have found the quite early paper by Gretton et al. (2005) useful both for a glimpse of the general theory and for the HSIC criterion in particular.

A reproducing kernel Hilbert space is a separable Hilbert space $\mathcal{F}$ of functions f on a set $\mathcal{X}$, such that the evaluation functional $f \to f(x)$ is a continuous linear functional on $\mathcal{F}$ for every $x \in \mathcal{X}$. Then, from the Riesz representation theorem, Muscat (2014), Chapter 10, there exists an element $k_x \in \mathcal{F}$ such that $\langle f, k_x \rangle = f(x)$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{F}$. Applying this to $f = k_x$ and another point $y \in \mathcal{X}$, we have $\langle k_x, k_y \rangle = k_x(y)$. The function $(x, y) \to k_x(y)$ from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ is the kernel of the RKHS $\mathcal{F}$. It is symmetric and positive definite because of the symmetry and positive definiteness of the inner product in $\mathcal{F}$. We use the notation $k(x, y)$ for the kernel.

The next step is to introduce another set $\mathcal{Y}$ with a corresponding RKHS $\mathcal{G}$ and to introduce a probability structure and probability measures p_X, p_Y and $p_{X,Y}$ on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$, respectively. With these probability measures and function spaces $\mathcal{F}$ and $\mathcal{G}$, one can introduce correlation of functions of stochastic variables on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$. This is an analogy of the functions used in the definition of the maximal correlation. In the RKHS setting, the covariance (or cross covariance) is an *operator* on the function space $\mathcal{F}$. Note also that this has a clear analogy in functional statistics; see, for example, Ferraty and Vieu (2006).

It is time to introduce the Hilbert–Schmidt operator: A linear operator $C : \mathcal{G} \to \mathcal{F}$ is called a Hilbert–Schmidt operator if its Hilbert–Schmidt (HS) norm $\|C\|_{\mathrm{HS}}$

$$\|C\|_{\mathrm{HS}}^2 \doteq \sum_{i,j} \langle C v_j, u_i \rangle_{\mathcal{F}}^2 < \infty,$$

where u_i and v_j are orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$, respectively. The HS-norm generalizes the Froebenius norm $\|A\|_F = (\sum_i \sum_j a_{ij}^2)^{1/2}$ for a matrix $A = (a_{ij})$. Finally, we need to define the tensor product in this context: If $f \in \mathcal{F}$ and $g \in \mathcal{G}$, then the tensor product operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ is defined by

$$(f \otimes g)h \doteq f \langle g, h \rangle_{\mathcal{G}}, \quad h \in \mathcal{G}.$$

Moreover, by using the definition of the HS norm it is not difficult to show that

$$\|f \otimes g\|_{\mathrm{HS}}^2 = \|f\|_{\mathcal{F}}^2 \|g\|_{\mathcal{G}}^2.$$

We can now introduce an expectation and a covariance on these function spaces. Again, the analogy with corresponding quantities in functional statistics will be clear. We assume that $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Lambda)$ are furnished with probability measures p_X, p_Y, and with Γ and Λ being σ-algebras of sets on $\mathcal{X}$ and $\mathcal{Y}$. The expectations $\mu_X \in \mathcal{F}$ and $\mu_Y \in \mathcal{G}$ are defined by, X and Y are stochastic variables in $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Lambda)$, respectively,

$$\langle \mu_X, f \rangle_{\mathcal{F}} = \mathrm{E}_X[f(X)]$$

and

$$\langle \mu_Y, g \rangle_{\mathcal{G}} = \mathrm{E}_Y[g(Y)],$$

where μ_X and μ_Y are well-defined as elements in $\mathcal{F}$ and $\mathcal{G}$ because of the Riesz representation theorem. The norm is obtained by

$$\|\mu_X\|_{\mathcal{F}}^2 = \mathrm{E}_{X,X'}[k(X, X')],$$

where as before X and X' are independent but have the same distribution p_X, and where $\|\mu_Y\|_{\mathcal{F}}$ is defined in the same way. With given $\phi \in \mathcal{F}$, $\psi \in \mathcal{G}$, we can now define the cross covariance operator as

$$\begin{aligned} C_{X,Y} &\doteq \mathrm{E}_{X,Y}[(\phi(X) - \mu_X) \otimes (\psi(Y) - \mu_Y)] \\ &= \mathrm{E}_{X,Y}[\phi(X) \otimes \phi(Y)] - \mu_X \otimes \mu_Y. \end{aligned}$$

Now, take $\phi(X)$ to be identified with $k_X \in \mathcal{F}$ defined above as a result of the Riesz representation theorem, and $\psi(Y) \in \mathcal{G}$ defined in exactly the same way. The Hilbert–Schmidt Information Criterion (HSIC) is then defined as

the squared HS norm of the associated cross-covariance operator

$$\mathrm{HSIC}(p_{XY}, \mathcal{F}, \mathcal{G}) \doteq \|C_{XY}\|_{\mathrm{HS}}^2.$$

Let $k(x, x')$ and $l(y, y')$ be kernel functions on $\mathcal{F}$ and $\mathcal{G}$. Then (Gretton et al., 2005, Lemma 1), the HSIC criterion can be written in terms of these kernels as

$$\begin{aligned}&\mathrm{HSIC}(p_{XY}, \mathcal{F}, \mathcal{G})\\ &\quad= \mathrm{E}_{X,X',Y,Y'}\big[k(X, X')l(Y, Y')\big]\\ &\qquad+ \mathrm{E}_{X,X'}\big[k(X, X')\big]\mathrm{E}_{Y,Y'}\big[l(Y, Y')\big]\\ &\qquad- 2\mathrm{E}_{X,Y}\big\{\mathrm{E}_{X'}\big[k(X, X')\big]\mathrm{E}_{Y'}\big[l(Y, Y')\big]\big\}.\end{aligned} \tag{4.10}$$

Existence is guaranteed if the kernels are bounded. The similarity in structure to (4.9) for the distance covariance should be noted. Note that the kernel functions depend on the way the spaces $\mathcal{F}$ and $\mathcal{G}$ and their inner products are defined. In fact, it follows from a famous result by Moore–Aronszajn (see Aronszajn, 1950), that if k is a symmetric, positive definite kernel on a set $\mathcal{X}$, then there is a unique Hilbert space of functions on $\mathcal{X}$ for which k is a reproducing kernel. Hence as will be seen next, in practice when applying the HSIC criterion, the user has to choose a kernel.

With some restrictions, the HSIC measure is a proper measure of dependence in the sense of the Rényi (1959) criterion (iv): From Theorem 4 of Gretton et al. (2005), one has that if the kernels k and l are universal (universal kernel has a mild continuity requirement on the kernel) on compact domains $\mathcal{X}$ and $\mathcal{Y}$, then $\|C_{XY}\|_{\mathrm{HS}} = 0$ if and only if X and Y are independent. The compactness assumption results from the application of an equality for bounded random variables taken from Hoeffding (1963).

A big asset of the HSIC measure is that its empirical version is easily computable. In fact, if we have independent observations $X_1, \ldots, X_n$ and independent observations $Y_1, \ldots, Y_n$, then

$$\mathrm{HSIC}_n(X, Y, \mathcal{F}, \mathcal{G}) = (n-1)^{-2}\,\mathrm{tr}\{KHLH\}, \tag{4.11}$$

where tr is the trace operator and the $n \times n$ matrices H, K, L are defined by

$$K = \{K_{ij}\} = \big\{k(X_i, X_j)\big\}, \quad L = \{L_{ij}\} = \big\{l(Y_i, Y_j)\big\},$$
$$H = \{H_{ij}\} = \big\{\delta_{ij} - n^{-1}\big\},$$

where δ_{ij} is the Kronecker delta. It is shown in Gretton et al. (2005) that this estimator converges in probability toward $\|C_{XY}\|_{\mathrm{HS}}^2$. The convergence rate is $n^{-1/2}$. There is also a limit theorem for the asymptotic distribution, which under the null hypothesis of independence and scaled with n, converges in distribution to the random variable $Q = \sum_{i,j=1}^{\infty} \lambda_i \eta_j N_{ij}^2$, where the N_{ij} are independent standard normal variables, and λ_i and η_j are eigenvalues of integral operators associated with centralized kernels derived from k and l and integrating using the probability measures p_X and p_Y, respectively. Again, this could be compared to the limiting variable for the statistic in the Cramér–von Mises functional as stated in Theorem 2 by Skaug and Tjøstheim (1993a), or see equation (1.2) in the online supplement. Critical values can be obtained for Q, but as a rule one seems to rely more on resampling as is the case for most independence test functionals.

It is seen from (4.11) that computation of the empirical HSIC criterion requires the evaluation of $k(X_i, X_j)$ and $l(Y_i, Y_j)$. Then appropriate kernels have to be chosen. Two commonly used kernels are the Gaussian kernel given by

$$k(x, y) = e^{\frac{|x-y|^2}{2\sigma^2}}, \quad \sigma > 0$$

and the Laplace kernel

$$k(x, y) = e^{\frac{|x-y|}{\sigma}}, \quad \sigma > 0.$$

Pfister and Peters (2017) describe the recent R-package `dHSIC` involving HSIC. Gretton et al. (2005) use these kernels in comparing the HSIC test with several other tests, including the dcov test in, among other cases, an independent component setting. Both of these tests do well, and none of them decisively out-competes the other. This is perhaps not so unexpected because there is a strong relationship between these two tests. This is demonstrated by Sejdinovic et al. (2013). They look at both the dcov test and the HSIC test in a generalized setting of semimetric spaces, that is, with kernels and distances defined on such spaces $\mathcal{X}$ and $\mathcal{Y}$. For a given distance function, they introduce a distance-induced kernel, and under certain regularity conditions they establish a relationship between these two quantities.

For the distance covariance and the HSIC, the distribution under the null and under the alternative are generally different. The discrepancy between the two distributions has been analyzed by Zhang et al. (2018) and Yao, Zhang and Shao (2018)

Lately there have been other extensions of both the dcov and HSIC to conditional dependence, partial distance and to time series. A few references are Szekely and Rizzo (2014), Zhang et al. (2012) and Pfister et al. (2018). A recent tutorial on RKHS is Gretton (2019).

Further, the generalization of the distance covariance to more than two vectors have independently been shown by Bilodeau and Nangue (2017), building on Bilodeau and Lafaye de Micheaux (2009) and Böttcher, Keller-Ressel and Schilling (2019). More specifically, Bilodeau and Nangue (2017) use the Möbius transformation of characteristic functions to characterize independence, and a generalization to p vectors of distance covariance and Hilbert–Schmidt independence criterion (HSIC) is proposed. Consistency and weak convergence of both types of statistics are established.

4.5 Density Based Tests of Independence

Intuitively, one might think that knowing that the density exists should lead to increased power of the independence tests due to more information. This is true, at least for some examples (see, e.g., Teräsvirta, Tjøstheim and Granger, 2010, Chapter 7.7). As in the preceding sections, one can construct distance functionals between the joint density under dependence and the product density under independence. A number of authors have considered such an approach; both in the i.i.d. and time series case; see, for example, Rosenblatt (1975), Robinson (1991), Skaug and Tjøstheim (1993b, 1996), Granger, Maasoumi and Racine (2004), Hong and White (2005), Su and White (2007) and Berrett and Samworth (2019). For two random variables X and Y having joint density $f_{X,Y}$ and marginals f_X and f_Y the degree of dependence can be measured by $\Delta(f_{X,Y}, f_X f_Y)$, where Δ is now the distance measure between two bivariate density functions. The variables may be normalized with $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$. It is natural to consider the Rényi (1959) requirements again, in particular, the requirements (iv) and (vi).

All of the distance functionals considered will be of type

$$\Delta = \int B\{f_X(x), f_Y(y), f_{X,Y}(x,y)\} \times f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y, \tag{4.12}$$

where B is a real-valued function such that the integral exists. If B is of the form $B(z_1, z_2, z_3) = D(z_1 z_2 / z_3)$, we have

$$\Delta = \int D\left\{\frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)}\right\} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{4.13}$$

which by the change of variable formula for integrals is seen to have the Rényi property (vi). Moreover, if $D(w) = 0$ if and only if $w = 1$, then Rényi property (iv) is fulfilled. If $D(1) = 0$ and D is convex, then D is a so-called f-divergence (Csiszár, 1967) measure with $f = D$. Several well-known distance measures for density functions are of this type. For instance, letting $D(w) = 2(1 - w^{1/2})$, we obtain the Hellinger distance

$$\begin{aligned} H &= \int \{\sqrt{f_{X,Y}(x,y)} - \sqrt{f_X(x) f_Y(y)}\}^2\,\mathrm{d}x\,\mathrm{d}y \\ &= 2\int \left\{1 - \sqrt{\frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)}}\right\} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

between $f_{X,Y}$ and $f_X f_Y$. The Hellinger distance is a metric, and hence satisfies the Rényi property (iv).

The familiar Kullback–Leibler information (entropy) distance is obtained by taking $D(w) = -\ln w$,

$$I = \int \ln\left\{\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\right\} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y. \tag{4.14}$$

Since this distance is of type (4.13), it satisfies (vi). A very recent paper linking I with other recent approaches to independence testing is Berrett and Samworth (2019). Taking $D(w) = w^2 - 1$ yields the χ^2-divergence; see also the test of fit distance in Bickel and Rosenblatt (1973).

All of the above measures are trivially extended to two arbitrary multivariate densities. However, estimating such densities in high or moderate dimensions may be difficult due to the curse of dimensionality. A functional built up from pairwise dependencies can be considered instead.

For a given functional $\Delta = \Delta(f, g)$ depending on two densities f and g, Δ may be estimated by $\widehat{\Delta} = \Delta(\widehat{f}, \widehat{g})$. There are several ways of estimating the densities, for example, the kernel density estimator,

$$\widehat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)$$

for given observations $\{X_1, \ldots, X_n\}$. Here, $K_b(x - X_i) = b^{-p} K\{b^{-1}(x - X_i)\}$, where b is the bandwidth (generally a matrix), K is the kernel function and p is the dimension of X_i. It should be pointed out that there are often different estimators of $\Delta(f, g)$ that are much easier to calculate and have better theoretical properties. For example, in the case of $\Delta = I$, one can consider the KSG-estimator; see Kraskov, Stögbauer and Grassberger (2004).

Under regularity conditions (see, e.g., Skaug and Tjøstheim, 1996), consistency and asymptotic normality under the null hypothesis of independence can be obtained for the estimated test functionals. Berrett and Samworth (2019) have demonstrated that local asymptotic power properties can also sometimes be proved. It should be noted that the leading term in an asymptotic expansion of the standard deviation of $\widehat{\Delta}$ for the estimated Kullback–Leibler functional $\widehat{I}$ and the estimated Hellinger functional $\widehat{H}$ is of order $O(n^{-1/2})$. This is, of course, the same as for the standard deviation of a parametric estimate in a parametric estimation problem. In that situation, the next term of the Edgeworth expansion is of order $O(n^{-1})$, and for moderately large values of n the first-order term $n^{-1/2}$ will dominate. However, for the functionals considered above, using density estimates and due to the presence of an n-dependent bandwidth, the next terms in the Edgeworth expansion are much closer, being of order $O(n^{-1/2}b)$ and $O(\{nb\}^{-1})$, and since typically $b = O(n^{-1/6})$ or $O(n^{-1/5})$, n must be very large indeed to have the first term dominate in the asymptotic expansion. As a consequence, first-order asymptotics in terms of the normal approximation cannot be expected to work well unless n is exceedingly large. In this sense, the situation is quite different from the empirical functionals treated in the previous sections, where there is no bandwidth parameter involved. All of this suggests the use of the bootstrap or permutations as an alternative for constructing the null distribution.

4.6 Global Test Functionals Generated by Local Dependence Relationships

If one has bivariate normal data with standard normal marginals and $\rho = 0$, one gets observations scattered in a disc-like region around zero, and most test functionals will easily recognize this as a situation of independence. However, as pointed out by Heller, Heller and Gorfine (2013), if data are generated along a circle with radius r, for example, $X^2 + Y^2 = r^2 + \epsilon$ for some stochastic noise variable ϵ, then X and Y are dependent, but as reported by Heller, Heller and Gorfine (2013), in practice, the dcov, and some other nonlinear global test functionals, do not work well. Heller, Heller and Gorfine (2013) point a way out of this difficulty, namely by looking at dependence locally (along the circle) and then aggregate the dependence by integrating, or by other means, over the local regions. There are, of course, several ways of measuring local dependence and we will approach this problem more fundamentally in Section 5.

Another paper in this category, Reshef et al. (2011), is published in *Science*. The idea behind their MIC (Maximal Information Coefficient) statistic consists in computing the mutual information I as defined in (4.14) *locally* over a grid in the data set and then take as statistic the maximum value of these local information measures as obtained by maximizing over a suitable choice of grid. Some limitations of the method are identified in a later article by Reshef et al. (2013) and it should also be pointed out that Kinney and Atwal (2014) find serious problems with the paper. See also Gorfine, Heller and Heller (2012).

Finally, the so-called BDS test named after its originators Brock et al. (1996) should be mentioned. This test has a local flavor at its basis, but the philosophy is a bit different from the other tests presented here. The BDS test attracted much attention among econometricians in the 1990s, and it has since been improved by Genest, Ghoudi and Rémillard (2007).

5. BEYOND PEARSON'S ρ: LOCAL DEPENDENCE

The test functionals treated in Section 4 deal with the second aspect of modeling dependence stated in the beginning of that section, namely that of *testing* of independence. These functionals all do so by the computation of one nonnegative number, which is derived from local properties in Section 4.5. This number properly scaled may possibly be said to deal with the first aspect stated, namely that of *measuring* the strength of the dependence. But, as such, it may be faulted in several ways. Unlike the Pearson ρ, these functionals do not distinguish between positive and negative dependence, and they are not local.

A local dependence measure between two stochastic variables X and Y can be defined as a measure based on the joint cumulative distribution function (or the joint density function in case it exists) for X and Y restricted to a local region R. In finance, it is of special interest to look at the local dependence when R is the tail region. If the joint cumulative distribution $F_{X,Y}$ is very different from the product $F_X F_Y$ of the marginals in R, this can be taken as an indication of strong local dependence between X and Y in R. As is the case for a global dependence measure, there are many ways of defining a local dependence measure. The local Gaussian correlation defined in the next section is just one possibility. The local region R can be determined by a bandwidth parameter or by some other regional distance measure. Accumulating a local measure over the entire space leads to a global measure as in, for example, (4.3) and (4.4), or as in the distance functionals of Section 4.5. It is also possible to shrink the region R to a point (x, y), and get a local value of the measure at that point, as is done for the local Gaussian correlation $\rho_{X,Y}(x, y)$ in Section 6.1 or implicitly as in, for example, (4.3) and (4.4).

In Section 6, the main story will be the treatment of a local Gaussian correlation which in a sense returns to the Pearson ρ, but a local version of ρ, which satisfies many of the Rényi (1959) requirements, and which is signed. But first, in the present section, we go back to some earlier attempts. We start with a remarkable paper by Lehmann (1966), who manages to define positive and negative dependence in quite a general nonlinear situation.

5.1 Quadrant Dependence

Lehmann's theory is based on the concept of quadrant dependence. Consider two random variables X and Y with cumulative distribution $F_{X,Y}$. Then the pair (X, Y) or its distribution function $F_{X,Y}$ is said to be positively quadrant dependent if

$$\begin{aligned} &P(X \le x, Y \le y) \\ &\quad \ge P(X \le x)P(Y \le y) \quad \text{for all } (x, y). \end{aligned} \tag{5.1}$$

Similarly, (X, Y) or $F_{X,Y}$ is said to be negatively quadrant dependent if (5.1) holds with the central inequality sign reversed.

The connection between quadrant dependence and Pearson's ρ is secured through a lemma of Hoeffding (1940). The lemma is a general result and resembles the result by Székely (2002) in his treatment of the so-called Cramér functional, a forerunner of the Cramér–von Mises functional. If $F_{X,Y}$ denotes the joint and F_X and F_Y the marginals, then assuming that the necessary moments exist,

$$\begin{aligned} &E(XY) - E(X)E(Y) \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(F_{XY}(x, y) - F_X(x)F_Y(y)\big)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

It follows immediately from definitions that if (X, Y) is positively quadrant dependent (negatively quadrant dependent), then for Pearson's ρ, $\rho \ge 0$ ($\rho \le 0$). Similarly,

it is shown by Lehmann that if $F_{X,Y}$ is positively quadrant dependent, then Kendall's τ, Spearman's ρ_S and the quadrant measure q defined by Blomqvist (1950) are all nonnegative. An analogous result holds in the negatively quadrant dependent case.

Lehmann (1966) introduced two additional and stronger concepts of dependence, namely regression dependence and likelihood ratio dependence; see his paper for details.

5.2 Local Measures of Dependence

As mentioned already, econometricians have long looked for a formal statistical way of describing the shifting region-like dependence structure of financial markets. It is obvious that when the market is going down there is a stronger dependence between financial objects, and very strong in case of a panic. Similar effects, but perhaps not quite so strong, appear when the market is going up. But how should it be quantified and measured? This is important in finance, not the least in portfolio theory, where it is well known (see, e.g., Taleb, 2007), that ordinary Gaussian description does not work, and if used, may lead to catastrophic results. Mainly two approaches have been used among econometricians. The first is nonlocal and consists simply in using copula theory, but it may not always be so easy to implement in a time series and portfolio context. The other approach is local and is to use "conditional correlation" as in Silvapulle and Granger (2001) and Forbes and Rigobon (2002). One then computes an estimate as in (1.1) of Pearson's ρ but in various regions of the sample space, for example, in the tail of the distribution of two variables.

However, this estimate suffers from a serious bias, which is obvious by using the ergodic theorem or the law of large numbers, in the sense that for a Gaussian distribution it does not converge to ρ. This is unfortunate because if the data happen to be Gaussian, one would like estimated correlations to be close or identical to ρ in order to approximate the classic Gaussian portfolio theory of Markowitz (1952). This requirement is consistent with Rényi's property (vii).

Statisticians have also tried various other ways of describing local dependence. Bjerve and Doksum (1993) suggested a local measure of dependence, the correlation curve, based on localizing ρ by conditioning on X in a nonlinear regression model. The resulting correlation curve inherits many of the properties of ρ, and it succeeds in several of the cases where ρ fails to detect dependence, such as the parabola (2.2) in Section 2.3. However, unlike ρ, it is not symmetric in (X, Y). Conditioning and regression on Y would in general produce a different result. This brings out the difference between regression analysis and multivariate analysis, where ρ is a concept of the latter, which happens to enter into the first. Bjerve and Doksum do propose a solution to this dilemma, but it is an ad hoc one.

Heller, Heller and Gorfine (2013) used local contingency type arguments to construct a global test functional. Such reasoning goes further back in time. Holland and Wang (1987) used such arguments to obtain a local dependence function

$$\gamma(x, y) = \frac{\partial^2}{\partial x \partial y} \ln f(x, y),$$

where f is the density function of (X, Y). Implicitly it is assumed here that both mixed second-order partial derivatives exist and are continuous. For an alternative derivation based on limiting arguments of local covariance functions and for properties and extensions, we refer to Jones (1996), Jones and Koch (2003) and Inci, Li and McCarthy (2011).

The local dependence function γ does not take values between -1 and 1, and it does not reduce to ρ in the Gaussian bivariate case. Actually, in that case

$$\gamma(x, y) = \frac{\rho}{1 - \rho^2} \frac{1}{\sigma_X \sigma_Y}.$$

6. BEYOND PEARSON'S ρ: LOCAL GAUSSIAN CORRELATION

The Pearson ρ gives a complete characterization of dependence in a bivariate Gaussian distribution but, as has been seen, not for a general density $f(x, y)$ for two random variables X and Y. The idea of the Local Gaussian Correlation (LGC), introduced in Tjøstheim and Hufthammer (2013) is to approximate f locally in a neighborhood of a point (x, y) by a bivariate Gaussian distribution $\psi_{x,y}(u, v)$, where (u, v) are running variables. In this neighborhood, one gets close to a complete local characterization of dependence using the local correlation $\rho(x, y)$, which is the Pearson's ρ of the bivariate Gaussian $\psi_{x,y}(u, v)$. Its precision depends on the size of the neighborhood and, of course, on the properties of the density at the point (x, y). In practice, it has to be reasonably smooth. This section and the online supplement give a survey of some of the results obtained so far.

6.1 Definition and Examples

For notational convenience in this section, we write (x_1, x_2) instead of (x, y), and, by a slight inconsistency of notation, $x = (x_1, x_2)$. Similarly, (u, v) is replaced by $v = (v_1, v_2)$. Then, in this notation, letting $\mu(x) = (\mu_1(x), \mu_2(x))$ be the mean vector of ψ, $\sigma(x) = (\sigma_1(x), \sigma_2(x))$ the vector of standard deviations and $\rho(x)$ the correlation of ψ, the approximating density ψ is given by

$$\psi\big(v, \mu_1(x), \mu_2(x), \sigma_1^2(x), \sigma_2^2(x), \rho(x)\big) = \frac{1}{2\pi\sigma_1(x)\sigma_2(x)\sqrt{1 - \rho^2(x)}}$$

$$\times \exp\Big[-\frac{1}{2}\frac{1}{1-\rho^2(x)}\Big(\frac{(v_1-\mu_1(x))^2}{\sigma_1^2(x)} - 2\rho(x)\frac{(v_1-\mu_1(x))(v_2-\mu_2(x))}{\sigma_1(x)\sigma_2(x)} + \frac{(v_2-\mu_2(x))^2}{\sigma_2^2(x)}\Big)\Big].$$

Moving to another point $y=(y_1,y_2)$ in general gives another approximating normal distribution ψ_y depending on a new set of parameters $\{\mu_1(y),\mu_2(y),\sigma_1(y),\sigma_2(y),\rho(y)\}$. An exception is the case where f itself is Gaussian with parameters $\{\mu_1,\mu_2,\sigma_1,\sigma_2,\rho\}$, in which case $\{\mu_1(x),\mu_2(x),\sigma_1(x),\sigma_2(x),\rho(x)\}\equiv\{\mu_1,\mu_2,\sigma_1,\sigma_2,\rho\}$. This means that the bias of the conditional correlation described in Section 5 is avoided and it means that the property (vii) in Rényi (1959)'s scheme is satisfied (and indeed (vii′) as well).

To make this into a construction that can be used in practice, it is convenient to define the vector population parameter $\theta(x)\doteq\{\mu_1(x),\mu_2(x),\sigma_1(x),\sigma_2(x),\rho(x)\}$ and estimate it. Fortunately, this is a problem that has been treated in larger generality by Hjort and Jones (1996) and Loader (1996). They looked at the problem of approximating $f(x)$ with a general parametric family of densities, the Gaussian being one such family. Here, x in principle can have a dimension ranging from 1 to p, but with $p=1$ mostly covered in those publications. They were concerned with estimating f rather than the local parameters, one of which is the local Gaussian correlation (LGC) $\rho(x)$.

But first we need a more precise definition of $\theta(x)$. This can be done in two stages using a neighborhood defined by bandwidths $b=(b_1,b_2)$ in the (x_1,x_2) direction, and then letting $b\to 0$ componentwise.

A suitable function measuring the difference between f and ψ is defined by

$$q=\int K_b(v-x)\big[\psi\big(v,\theta(x)\big) - \ln\{\psi\big(v,\theta(x)\big)\}f(v)\big]\,\mathrm{d}v, \tag{6.1}$$

where $K_b(v-x)=(b_1b_2)^{-1}K_1(b_1^{-1}(v_1-x_1))\times K_2(b_2^{-1}(v_2-x_2))$ is a product kernel. As is seen in Hjort and Jones (1996), pages 1623–1624, the expression in (6.1) can be interpreted as a locally weighted Kullback–Leibler distance from f to $\psi(\cdot,\theta(x))$. We then obtain that the minimizer $\theta_b(x)$ (also depending on K) should satisfy

$$\int K_b(v-x)\frac{\partial}{\partial\theta_j}\big[\ln\{\psi\big(v,\theta(x)\big)\}f(v) - \psi\big(v,\theta(x)\big)\big]\,\mathrm{d}v=0,\quad j=1,\dots,5. \tag{6.2}$$

In the first stage, we define the population value $\theta_b(x)$ as the minimizer of (6.1), assuming that there is a unique solution to (6.2). The definition of $\theta_b(x)$ and the assumption of uniqueness are essentially identical to those used in Hjort and Jones (1996) for more general parametric families of densities.

In the next stage, we let $b\to 0$ and consider the limiting value $\theta(x)=\lim_{b\to 0}\theta_b(x)$. This is in fact considered indirectly by Hjort and Jones (1996) on pages 1627–1630 and more directly in Tjøstheim and Hufthammer (2013), both using Taylor expansion arguments. In the following, we will assume that a limiting value $\theta(x)$ independent of b and K exists. (It is possible to avoid the problem of a population value altogether if one takes the view of some of the publications cited in Section 4.6 by just estimating a suitable dependence function.)

In estimating $\theta(x)$ and $\theta_b(x)$, a neighborhood with a finite bandwidth has to be used in analogy with nonparametric density estimation. The estimate $\widehat{\theta}(x)=\widehat{\theta}_b(x)$ is obtained from maximizing a local likelihood. Given observations $X_1,\dots,X_n$, the local log likelihood is determined by

$$L\big(X_1,\dots,X_n,\theta(x)\big) = n^{-1}\sum_{i=1}^{n}K_b(X_i-x)\ln\psi\big(X_i,\theta(x)\big) - \int K_b(v-x)\psi\big(v,\theta(x)\big)\,\mathrm{d}v.$$

The last (and perhaps somewhat unexpected) term is essential, as it implies that $\psi(x,\theta_b(x))$ is not allowed to stray far away from $f(x)$ as $b\to 0$. It is also discussed at length in Hjort and Jones (1996). (When $b\to\infty$, the last term has 1 as its limiting value and the likelihood reduces to the ordinary global log-likelihood.) Using the notation,

$$u_j(\cdot,\theta)\doteq\frac{\partial}{\partial\theta_j}\ln\psi(\cdot,\theta),$$

by the law of large numbers, or by the ergodic theorem in the time series case, assuming $\mathrm{E}\{K_b(X_i-x)\ln\psi(X_i,\theta_b(x))\}<\infty$, we have almost surely

$$\frac{\partial L}{\partial\theta_j} = n^{-1}\sum_i K_b(X_i-x)u_j\big(X_i,\theta_b(x)\big) - \int K_b(v-x)u_j\big(v,\theta_b(x)\big)\psi\big(v,\theta_b(x)\big)\,\mathrm{d}v \to \int K_b(v-x)u_j\big(v,\theta_b(x)\big)\times\big[f(v)-\psi\big(v,\theta_b(x)\big)\big]\,\mathrm{d}v. \tag{6.3}$$

Putting the expression in the first line of (6.3) equal to zero yields the local maximum likelihood estimate $\widehat{\theta}_b(x)=\widehat{\theta}(x)$ of the population value $\theta_b(x)$, which satisfies (6.2).

We see the importance of the additional last term in the local likelihood by letting $b\to 0$, Taylor expanding and requiring $\partial L/\partial\theta_j=0$, which leads to

$$u_j\big(x,\theta_b(x)\big)\big[f(x)-\psi\big(x,\theta_b(x)\big)\big]+O\big(b^Tb\big)=0,$$

where b^T is the transposed of b. It is seen that ignoring solutions that yield $u_j(x, \theta_b(x)) = 0$ requires $\psi(x, \theta_b(x))$ to be close to $f(x)$.

An asymptotic theory has been developed in Tjøstheim and Hufthammer (2013) for $\widehat{\theta}_b(x)$ for the case that b is fixed and for $\widehat{\theta}(x)$ in the case that $b \to 0$. The first case is much easier to treat than the second one. In fact, for the first case the theory of Hjort and Jones (1996) can be taken over almost directly, although it is extended to the ergodic time series case in Tjøstheim and Hufthammer (2013). In the case that $b \to 0$, this leads to a slow convergence rate of $(n(b_1 b_2)^3)^{-1/2}$, which is the same convergence rate as for the the estimated dependence function treated in Jones (1996).

The local correlation is clearly dependent on the marginal distributions of X_1 and X_2 as is Pearson's ρ. This marginal dependence can be removed by scaling the observations to a standard normal scale. As mentioned in Section 3 about the copula, the dependence structure is disentangled from the marginals by Sklar's theorem. For the purpose of measuring local dependence in terms of the local Gaussian correlation, at least for a number of purposes it is advantageous to replace a scaling with uniform variables $U_i = F_i(X_i)$ by standard normal variables

$$Z = (Z_1, Z_2) = \big(\Phi^{-1}\big(F_1(X_1)\big), \Phi^{-1}\big(F_2(X_2)\big)\big), \tag{6.4}$$

where Φ is the cumulative distribution of the standard normal distribution. The local Gaussian correlation on the Z-scale will be denoted by $\rho_Z(z)$. Of course, the variable Z cannot be computed via the transformation (6.4) without knowledge of the margins F_1 and F_2, but these can be estimated by the empirical distribution function. Extensive use has been made of $\rho_Z(z_1, z_2)$, or rather $\rho_{\widehat{Z}}(z_1, z_2)$ with $\widehat{Z}_i = \Phi^{-1}(\widehat{F}_i)$. Under certain regularity conditions, as in the copula case, the difference between Z and $\widehat{Z}$ can be ignored in limit theorems. Using the sample of pairs of Gaussian pseudo observations $\{\Phi^{-1}(\widehat{F}_1(X_{1i}), \Phi^{-1}(\widehat{F}_2(X_{2i})\}$, $i = 1, \ldots, n$, one can estimate $\rho_Z(z_1, z_2)$ by local log likelihood as described above. Under regularity conditions, the asymptotic theory will be the same as in Tjøstheim and Hufthammer (2013). In Otneim and Tjøstheim (2017, 2018), a further simplification is made by taking $\mu_{Z_i}(z) \equiv 0$ and $\sigma_{Z_i}(z) \equiv 1$, in which case the asymptotic theory simplifies and one obtains the familiar nonparametric rate of $O((nb_1b_2)^{-1/2})$ for $\widehat{\rho}_{\widehat{Z}}(z)$.

The choice of Gaussian margins in the transformation (6.4) is not made without a purpose. It is natural since we are dealing with local Gaussian approximations. This leads to a more fundamental question: Why is a local Gaussian approximation and an associated local Gaussian correlation measure particularly useful?

6.2 Why Local Gaussian Approximation and Local Gaussian Correlation?

In principle, another parametric family could be used as a local approximation (as has been done by Hjort and Jones, 1996) in their consideration of locally parametric density estimation. The advantage of using the Gaussian distribution as an approximating family is its powerful and unique properties. Among them is the fact that the entire dependence structure of a multivariate Gaussian is determined by its set of pairwise correlations, and the fact that for a multivariate Gaussian the conditional distribution of one set of variables given another set of variables is again Gaussian. The idea, or the statistical modeling philosophy, of the local Gaussian approximation is that the unique properties of the Gaussian can be extended to non-Gaussian distributions, but *locally*. This can be shown to be useful in a number of different situations as follows.

Description of dependence and corresponding tests of independence have been given for pairs of i.i.d. variables, for single time series, and for pairs of time series, including the use of local Gaussian autocorrelation, in Berentsen and Tjøstheim (2014), and in Lacal and Tjøstheim (2017, 2019)

Applications to econometric data are given in Støve, Tjøstheim and Hufthammer (2014) and Støve and Tjøstheim (2014). In particular, for multivariate financial data, one can measure the increasing values of pairwise local Gaussian correlations in a market during an economic downturn. This describes quantitatively the well-known fact that financial objects perform similarly (stronger mutual positive dependence) in such a situation. In the extreme scenario of a panic, the local Gaussian correlations would approach 1; see also Nguyen et al. (2020).

A local Gaussian conditional distribution allows the introduction of a local Gaussian *partial* correlation, and density and conditional density estimation, as well as tests of conditional independence treated by Otneim and Tjøstheim (2017, 2018, 2021).

Locally Gaussian spectral estimation is contained in Jordanger and Tjøstheim (2020). They have shown that nonlinear and local oscillatory behavior can be detected in cases where it is missed in ordinary spectral analysis.

Finally, relationships to the copula concept have been investigated in Berentsen et al. (2014), and applications to discrimination using a local Fisher discriminant have been explored in Otneim, Jullum and Tjøstheim (2020). There are three R-packages; Berentsen, Kleppe and Tjøstheim (2014), Jordanger (2020) and Otneim (2019). All of these developments are being collected in a forthcoming book, Tjøstheim, Otneim and Støve (2021).

Due to lack of space, these developments cannot be described in more detail in the main part of this paper, but for the reader's convenience, we give a brief summary in the online supplementary material. Further, there are plots of

the local Gaussian correlation in simulation experiments, and for the financial return data of Figure 2. Finally, pointers are given to where the local Gaussian correlation is compared in testing situations with the dcov statistic from Section 4.3 and with the ordinary global Pearson's ρ.

ACKNOWLEDGMENTS

We are grateful to the Editor and to two anonymous referees for a number of valuable comments and suggestions.

FUNDING

This work has been partially supported by the Finance Market Fund (Norway).

SUPPLEMENTARY MATERIAL

Supplement to "Statistical Dependence: Beyond Pearson's ρ" (DOI: 10.1214/21-STS823SUPP; .pdf). The supplementary material consists of four sections. In Section 1 of the supplementary material we give some more details of Section 4.2 of the main article concerning the measuring and tests based on the distribution function in the time series case. Section 2 gives more details for the local Gaussian correlation in the time series and copula case, and it contains some simulation experiments and a real data example. Section 3 gives some additional symmetry properties of the local Gaussian correlation and a discussion of the Rényi criteria relative to this measure of dependence. Finally, Section 4 contains a brief overview of the use of the local Gaussian correlation in testing of independence.

REFERENCES

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884 https://doi.org/10.1016/j.insmatheco.2007.02.001

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404

BERENTSEN, G. D., KLEPPE, T. and TJØSTHEIM, D. (2014). Introducing localgauss, an R-package for estimating and visualizing local Gaussian corelation. *J. Stat. Softw.* **56** 1–18.

BERENTSEN, G. D. and TJØSTHEIM, D. (2014). Recognizing and visualizing departures from independence in bivariate data using local Gaussian correlation. *Stat. Comput.* **24** 785–801. MR3229697 https://doi.org/10.1007/s11222-013-9402-8

BERENTSEN, G. D., STØVE, B., TJØSTHEIM, D. and NORDBØ, T. (2014). Recognizing and visualizing copulas: An approach using local Gaussian approximation. *Insurance Math. Econom.* **57** 90–103. MR3225330 https://doi.org/10.1016/j.insmatheco.2014.04.005

BERGSMA, W. and DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli* **20** 1006–1028. MR3178526 https://doi.org/10.3150/13-BEJ514

BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. With a preface by Persi Diaconis. MR2239907 https://doi.org/10.1007/978-1-4419-9096-9

BERRETT, T. B. and SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. MR0348906

BILODEAU, M. and LAFAYE DE MICHEAUX, P. (2009). *A*-dependence statistics for mutual and serial independence of categorical variables. *J. Statist. Plann. Inference* **139** 2407–2419. MR2508002 https://doi.org/10.1016/j.jspi.2008.11.006

BILODEAU, M. and NANGUE, A. G. (2017). Tests of mutual or serial independence of random vectors with applications. *J. Mach. Learn. Res.* **18** Paper No. 74. MR3714237

BJERVE, S. and DOKSUM, K. (1993). Correlation curves: Measures of association as functions of covariate values. *Ann. Statist.* **21** 890–902. MR1232524 https://doi.org/10.1214/aos/1176349156

BLOMQVIST, N. (1950). On a measure of dependence between two random variables. *Ann. Math. Stat.* **21** 593–600. MR0039190 https://doi.org/10.1214/aoms/1177729754

BLUM, J. R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Stat.* **32** 485–498. MR0125690 https://doi.org/10.1214/aoms/1177705055

BÖTTCHER, B., KELLER-RESSEL, M. and SCHILLING, R. L. (2019). Distance multivariance: New dependence measures for random vectors. *Ann. Statist.* **47** 2757–2789. MR3988772 https://doi.org/10.1214/18-AOS1764

BROCK, W. A., SCHEINKMAN, J. A., DECHERT, W. D. and LEBARON, B. (1996). A test for independence based on the correlation dimension. *Econometric Rev.* **15** 197–235. MR1410877 https://doi.org/10.1080/07474939608800353

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345

CSÖRGŐ, S. (1985). Testing for independence by the empirical characteristic function. *J. Multivariate Anal.* **16** 290–299. MR0793494 https://doi.org/10.1016/0047-259X(85)90022-3

DATASTREAM (2018). Subscription service. Accessed June 2018.

DAVIS, R. A., MATSUI, M., MIKOSCH, T. and WAN, P. (2018). Applications of distance correlation to time series. *Bernoulli* **24** 3087–3116. MR3779711 https://doi.org/10.3150/17-BEJ955

DEHEUVELS, P. (1981a). A Kolmogorov–Smirnov type test for independence and multivariate samples. *Rev. Roumaine Math. Pures Appl.* **26** 213–226. MR0616038

DEHEUVELS, P. (1981b). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multivariate Anal.* **11** 102–113. MR0612295 https://doi.org/10.1016/0047-259X(81)90136-6

DUECK, J., EDELMANN, D., GNEITING, T. and RICHARDS, D. (2014). The affinely invariant distance correlation. *Bernoulli* **20** 2305–2330. MR3263106 https://doi.org/10.3150/13-BEJ558

ENGLE, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50** 987–1007. MR0666121 https://doi.org/10.2307/1912773

ESCANCIANO, J. C. and HUALDE, J. (2019). Measuring asset market linkages: Nonlinear dependence and tail risk. *J. Bus. Econom. Statist.* 1–25.

ESCANCIANO, J. C. and VELASCO, C. (2006). Generalized spectral tests for the martingale difference hypothesis. *J. Econometrics* **134**

151–185. MR2328319 https://doi.org/10.1016/j.jeconom.2005.06.019

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. Springer, New York. MR2229687

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

FOKIANOS, K. and PITSILLOU, M. (2017). Consistent testing for pairwise dependence in time series. *Technometrics* **59** 262–270. MR3635048 https://doi.org/10.1080/00401706.2016.1156024

FORBES, K. J. and RIGOBON, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *J. Finance* **57** 2223–2261.

FRANCQ, C. and ZAKOÏAN, J.-M. (2011). *GARCH Models*: *Structure*, *Statistical Inference and Financial Applications*. Wiley, Chichester. MR3186556 https://doi.org/10.1002/9780470670057

GALTON, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proc. Roy. Soc. Lond.* **45** 135–145.

GALTON, F. (1890). Kinship and correlation. *N. Amer. Rev.* **150** 419–431.

GEBELEIN, H. (1941). Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM Z. Angew. Math. Mech.* **21** 364–379. MR0007220 https://doi.org/10.1002/zamm.19410210604

GENEST, C., GHOUDI, K. and RÉMILLARD, B. (2007). Rank-based extensions of the Brock, Dechert, and Scheinkman test. *J. Amer. Statist. Assoc.* **102** 1363–1376. MR2372539 https://doi.org/10.1198/016214507000001076

GENEST, C. and NEŠLEHOVÁ, J. (2007). A primer on copulas for count data. *Astin Bull.* **37** 475–515. MR2422797 https://doi.org/10.2143/AST.37.2.2024077

GENEST, C., KOJADINOVIC, I., NEŠLEHOVÁ, J. and YAN, J. (2011). A goodness-of-fit test for bivariate extreme-value copulas. *Bernoulli* **17** 253–275. MR2797991 https://doi.org/10.3150/10-BEJ279

GHOUDI, K. and RÉMILLARD, B. (2018). Serial independence tests for innovations of conditional mean and variance models. *TEST* **27** 3–26. MR3764021 https://doi.org/10.1007/s11749-016-0521-3

GÓMEZ, E., GÓMEZ-VILLEGAS, M. A. and MARÍN, J. M. (2003). A survey on continuous elliptical vector distributions. *Rev. Mat. Complut.* **16** 345–361. MR2031887 https://doi.org/10.5209/rev_REMA.2003.v16.n1.16889

GORFINE, M., HELLER, R. and HELLER, Y. (2012). Comment on "Detecting novel associations in large data sets" by Reshef et al., Science Dec 16, 2011.

GRANGER, C. W., MAASOUMI, E. and RACINE, J. (2004). A dependence metric for possibly nonlinear processes. *J. Time Series Anal.* **25** 649–669. MR2086354 https://doi.org/10.1111/j.1467-9892.2004.01866.x

GRETTON, A. (2019). Introduction to RKHS, and some simple kernel algorithms. Unpublished manuscript, Lecture Notes Gatsby Computational Neuroscience Unit.

GRETTON, A. and GYÖRFI, L. (2010). Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* **11** 1391–1423. MR2645456

GRETTON, A. and GYÖRFI, L. (2012). Strongly consistent nonparametric test of conditional independence. *J. Multivariate Anal.* **82** 1145–1150.

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory* (S. Jain, U. Simon and E. Tomita, eds.). *Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. MR2255909 https://doi.org/10.1007/11564089_7

HELLER, R., HELLER, Y. and GORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510. MR3068450 https://doi.org/10.1093/biomet/ass070

HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647. MR1416653 https://doi.org/10.1214/aos/1032298288

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. MR0029139 https://doi.org/10.1214/aoms/1177730150

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

HÖFFDING, W. (1940). Maszstabinvariante Korrelationstheorie. *Schr. Math. Inst. U. Inst. Angew. Math. Univ. Berlin* **5** 181–233. MR0004426

HOLLAND, P. W. and WANG, Y. J. (1987). Dependence function for continuous bivariate densities. *Comm. Statist. Theory Methods* **16** 863–876. MR0886560 https://doi.org/10.1080/03610928708829408

HONG, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: A generalized spectral density approach. *J. Amer. Statist. Assoc.* **94** 1201–1220. MR1731483 https://doi.org/10.2307/2669935

HONG, Y. (2000). Generalized spectral tests for serial dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 557–574. MR1772415 https://doi.org/10.1111/1467-9868.00250

HONG, Y. and WHITE, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* **73** 837–901. MR2135144 https://doi.org/10.1111/j.1468-0262.2005.00597.x

HUANG, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* **38** 2047–2091. MR2676883 https://doi.org/10.1214/09-AOS770

INCI, A. C., LI, H.-C. and MCCARTHY, J. (2011). Financial contagion: A local correlation analysis. *Res. Int. Bus. Finance* **25** 11–25.

JENTSCH, C., LEUCHT, A., MEYER, M. and BEERING, C. (2020). Empirical characteristic functions-based estimation and distance correlation for locally stationary processes. *J. Time Series Anal.* **41** 110–133. MR4048684 https://doi.org/10.1111/jtsa.12497

JOE, H. (2014). *Dependence Modeling with Copulas*. Chapman & Hall, London.

JONES, M. C. (1996). The local dependence function. *Biometrika* **83** 899–904. MR1440052 https://doi.org/10.1093/biomet/83.4.899

JONES, M. C. and KOCH, I. (2003). Dependence maps: Local dependence in practice. *Stat. Comput.* **13** 241–255. MR1982478 https://doi.org/10.1023/A:1024270700807

JORDANGER, L. A. (2020). LocalgaussSpec. Available at https://github.com/LAJordanger/localgaussSpec.

JORDANGER, L. A. and TJØSTHEIM, D. (2020). Nonlinear spectral analysis: A local Gaussian approach. *J. Amer. Statist. Assoc.* 1–55.

KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–89.

KING, M. L. (1987). Testing for autocorrelation in linear regression models: A survey. In *Specification Analysis in the Linear Model* (M. L. King and D. E. A. Giles, eds.). *Internat. Lib. Econom.* 19–73. Routledge, London. MR0899966

KINNEY, J. B. and ATWAL, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **111** 3354–3359. MR3200177 https://doi.org/10.1073/pnas.1309933111

KLAASSEN, C. A. J. and WELLNER, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3** 55–77. MR1466545 https://doi.org/10.2307/3318652

KOJADINOVIC, I. and HOLMES, M. (2009). Tests of independence among continuous random vectors based on Cramér–von Mises functionals of the empirical copula process. *J. Multivariate Anal.* **100** 1137–1154. MR2508377 https://doi.org/10.1016/j.jmva.2008.10.013

KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* (3) **69** 066138. MR2096503 https://doi.org/10.1103/PhysRevE.69.066138

LACAL, V. and TJØSTHEIM, D. (2017). Local Gaussian autocorrelation and tests for serial independence. *J. Time Series Anal.* **38** 51–71. MR3601314 https://doi.org/10.1111/jtsa.12195

LACAL, V. and TJØSTHEIM, D. (2019). Estimating and testing nonlinear local dependence between two time series. *J. Bus. Econom. Statist.* **37** 648–660. MR4016160 https://doi.org/10.1080/07350015.2017.1407777

LANCASTER, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **44** 289–292.

LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Stat.* **37** 1137–1153. MR0202228 https://doi.org/10.1214/aoms/1177699260

LOADER, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24** 1602–1618. MR1416652 https://doi.org/10.1214/aos/1032298287

MARKOWITZ, H. M. (1952). Portfolio selection. *J. Finance* **7** 77–91.

MUSCAT, J. (2014). *Functional Analysis*: *An Introduction to Metric Spaces*, *Hilbert Spaces*, *and Banach Algebras*. Springer, Cham. MR3308576 https://doi.org/10.1007/978-3-319-06728-5

NELSEN, R. B. (1999). *An Introduction to Copulas*. *Lecture Notes in Statistics* **139**. Springer, New York. MR1653203 https://doi.org/10.1007/978-1-4757-3076-0

NGUYEN, Q. N., ABOURA, S., CHEVALLIER, J., ZHANG, L. and ZHU, B. (2020). Local Gaussian correlations in financial and commodity markets. *European J. Oper. Res.* **285** 306–323. MR4083070 https://doi.org/10.1016/j.ejor.2020.01.023

OTNEIM, H. (2019). `lg`: Locally gaussian distributions: Estimation and methods. Available at https://CRAN.R-project.org/package=lg.

OTNEIM, H., JULLUM, M. and TJØSTHEIM, D. (2020). Pairwise local Fisher and naive Bayes: Improving two standard discriminants. *J. Econometrics* **216** 284–304. MR4077395 https://doi.org/10.1016/j.jeconom.2020.01.019

OTNEIM, H. and TJØSTHEIM, D. (2017). The locally Gaussian density estimator for multivariate data. *Stat. Comput.* **27** 1595–1616. MR3687328 https://doi.org/10.1007/s11222-016-9706-6

OTNEIM, H. and TJØSTHEIM, D. (2018). Conditional density estimation using the local Gaussian correlation. *Stat. Comput.* **28** 303–321. MR3747565 https://doi.org/10.1007/s11222-017-9732-z

OTNEIM, H. and TJØSTHEIM, D. (2021). The locally Gaussian partial correlation. *J. Bus. Econom. Statist.* 1–33. To appear.

PEARSON, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. R. Soc. Lond.* **187** 253–318.

PEARSON, K. (1922). *Francis Galton*: *A Centenary Appreciation*. Cambridge Univ. Press, Cambridge.

PEARSON, K. (1930). *The Life*, *Letters and Labors of Francis Galton*. Cambridge Univ. Press, Cambridge.

PFISTER, N. and PETERS, J. (2017). dHSIC: Independence testing via Hilbert Schmidt independence criterion. Available at https://CRAN.R-project.org/package=dHSIC.

PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. and PETERS, J. (2018). Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 5–31. MR3744710 https://doi.org/10.1111/rssb.12235

PINKSE, J. (1998). A consistent nonparametric test for serial independence. *J. Econometrics* **84** 205–231. MR1630190 https://doi.org/10.1016/S0304-4076(97)00084-5

PRUDNIKOV, A. P., BRYCHKOV, Y. A. and MARICHEV, O. I. (1986). *Integrals and Series*. Gordon & Breach, New York.

RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10** 441–451. MR0115203 https://doi.org/10.1007/BF02024507

RESHEF, D. N., RESHEF, Y. A., FINUCANE, H. K., GROSSMAN, S. R., MCVEAN, G., TURNBAUGH, P. J., LANDER, E. S., MITZENMACHER, M. and SABETI, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524. https://doi.org/10.1126/science.1205438

RESHEF, D., RESHEF, Y., MITZENMACHER, M. and SABETI, P. (2013). Equitability analysis of the maximal information coefficient, with comparisons.

RIZZO, M. L. and SZEKELY, G. J. (2018). Energy: E-statistics: Multivariate inference via the energy of data. Available at https://CRAN.R-project.org/package=energy.

ROBINSON, P. M. (1991). Consistent nonparametric entropy-based testing. *Rev. Econ. Stud.* **58** 437–453. MR1108130 https://doi.org/10.2307/2298005

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1–14. MR0428579

SEJDINOVIC, D., SRIPERUMBUDUR, B., GRETTON, A. and FUKUMIZU, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866 https://doi.org/10.1214/13-AOS1140

SILVAPULLE, P. and GRANGER, C. W. J. (2001). Large returns, conditional correlation and portfolio diversification: A value-at-risk approach. *Quant. Finance* **1** 542–551. MR1863876 https://doi.org/10.1088/1469-7688/1/5/306

SKAUG, H. J. and TJØSTHEIM, D. (1993a). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika* **80** 591–602. MR1248024 https://doi.org/10.1093/biomet/80.3.591

SKAUG, H. J. and TJØSTHEIM, D. (1993b). Nonparametric tests of serial independence. In *Developments in Time Series Analysis* (T. S. Rao, ed.) 207–229. CRC Press, London. MR1292268

SKAUG, H. J. and TJØSTHEIM, D. (1996). Testing for serial independence using measures of distance between densities. In *Athens Conference on Applied Probability and Time Series Analysis*, *Vol. II* (P. M. Robinson and M. Rosenblatt, eds.). *Lect. Notes Stat.* **115** 363–377. Springer, New York. MR1466759 https://doi.org/10.1007/978-1-4612-2412-9_27

SKLAR, M. (1959). Fonctions de Répartition à N Dimensions et Leurs Marges. Université Paris 8.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* **15** 72–101.

STANTON, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *J. Stat. Educ.* **9** 1–13.

STIGLER, S. M. (1989). Francis Galton's account of the invention of correlation. *Statist. Sci.* **4** 73–79. MR1007556

STØVE, B. and TJØSTHEIM, D. (2014). Asymmetric dependence patterns in financial returns: An empirical investigation using local Gaussian correlation. In *Essays in Nonlinear Time Series Econometrics* (M. Meitz N. Haldrup and P. Saikkonen, eds.) 307–329. Oxford Univ. Press, Oxford. MR3288225 https://doi.org/10.1093/acprof:oso/9780199679959.003.0013

STØVE, B. TJØSTHEIM, D. and HUFTHAMMER, K. (2014). Using local Gaussian correlation in a nonlinear re-examination of financial contagion. *J. Empir. Finance* **25** 785–801.

SU, L. and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *J. Econometrics* **141** 807–834. MR2413488 https://doi.org/10.1016/j.jeconom.2006.11.006

SZÉKELY, G. J. (2002). $\mathcal{E}$-statistics: The energy of statistical samples. Technical report 02-16, Bowling Green State Univ., Bowling Green, OH.

SZÉKELY, G. J. and RIZZO, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *J. Classification* **22** 151–183. MR2231170 https://doi.org/10.1007/s00357-005-0012-9

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 https://doi.org/10.1214/09-AOAS312

SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745 https://doi.org/10.1016/j.jspi.2013.03.018

SZÉKELY, G. J. and RIZZO, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann. Statist.* **42** 2382–2412. MR3269983 https://doi.org/10.1214/14-AOS1255

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

TALEB, N. N. (2007). *The Black Swan*: *The Impact of the Highly Improbable*. Random House, New York.

TERÄSVIRTA, T., TJØSTHEIM, D. and GRANGER, C. W. J. (2010). *Modelling Nonlinear Economic Time Series*. *Advanced Texts in Econometrics*. Oxford Univ. Press, Oxford. MR3185399 https://doi.org/10.1093/acprof:oso/9780199587148.001.0001

TJØSTHEIM, D. and HUFTHAMMER, K. O. (2013). Local Gaussian correlation: A new measure of dependence. *J. Econometrics* **172** 33–48. MR2997128 https://doi.org/10.1016/j.jeconom.2012.08.001

TJØSTHEIM, D., OTNEIM, H. and STØVE, B. (2021). *Statistical Modeling Using Local Gaussian Approximation*. Elsevier, Amsterdam. To appear.

TJØSTHEIM, D., OTNEIM, H. and STØVE, B. (2022). Supplement to "Statistical Dependence: Beyond Pearson's ρ." https://doi.org/10.1214/21-STS823SUPP

VON NEUMANN, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **12** 367–395. MR0006656 https://doi.org/10.1214/aoms/1177731677

VON NEUMANN, J. (1942). A further remark concerning the distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Stat.* **13** 86–88. MR0006657 https://doi.org/10.1214/aoms/1177731645

YAO, S., ZHANG, X. and SHAO, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 455–480. MR3798874 https://doi.org/10.1111/rssb.12259

YENIGÜN, C. D., SZÉKELY, G. J. and RIZZO, M. L. (2011). A test of independence in two-way contingency tables based on maximal correlation. *Comm. Statist. Theory Methods* **40** 2225–2242. MR2862708 https://doi.org/10.1080/03610921003764274

ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2012). Kernel-based conditional independence test and applications in causal discovery. In *Proceedings of the Uncertainty in Artificial Intelligence* 804–813. AUAI Press, Corvallis, OR.

ZHANG, Q., FILIPPI, S., GRETTON, A. and SEJDINOVIC, D. (2018). Large-scale kernel methods for independence testing. *Stat. Comput.* **28** 113–130. MR3741641 https://doi.org/10.1007/s11222-016-9721-7

ZHOU, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *J. Time Series Anal.* **33** 438–457. MR2915095 https://doi.org/10.1111/j.1467-9892.2011.00780.x