Abstract. Pearson’s $\rho$ 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 $\rho$. 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 $\rho$, nonlinear dependence, distance covariance, HSIC, mutual information, local Gaussian correlation.

1. INTRODUCTION

Pearson’s $\rho$, 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 $\rho$ 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 $\rho$. Galton died in 1911, and Karl Pearson became his biographer, resulting in a massive 4-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other relevant historical references are Fisher (1915, 1921), von Neumann (1941, 1942) and the survey paper by King (1987).

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

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

with \( \sigma_X = \sqrt{\sigma_X^2} = \sqrt{E(X - E(X))^2} \) being the standard deviation of \( X \) and similarly for \( \sigma_Y \). 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 \( \rho \) is given by

\[
r = \hat{\rho} = \frac{\sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2} \sqrt{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}}
\]

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

The correlation coefficient \( \rho \) 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, \( \rho \) is proportional to the slope of the regression line. This also means that \( \rho \) and its estimate \( \hat{\rho} \) appears naturally in a linear least squares analysis.

(iii) In a bivariate Gaussian density

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

the dependence between \( X \) and \( Y \) is completely characterized by \( \rho \). 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 \( X_1, \ldots, X_p \), their joint dependencies can simply (but not always accurately) be characterized by their covariance matrix \( \Sigma = \{\sigma_{ij}\} \), with \( \sigma_{ij} \) being the covariance between \( X_i \) and \( X_j \). Similarly the correlation matrix is defined by \( \Lambda = \{\rho_{ij}\} \), with \( \rho_{ij} \) being the correlation
between $X_i$ and $X_j$. Again, for a column vector $x = (x_1, \ldots, x_p)^T$, the joint normality density is defined by

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

where $|\Sigma|$ is the determinant of the covariance matrix $\Sigma$ (whose inverse $\Sigma^{-1}$ is assumed to exist), and $\mu = E(X)$. Then the complete dependence structure of the Gaussian vector is given by the pairwise covariances $\sigma_{ij}$, or equivalently the pairwise correlations $\rho_{ij}$. This is remarkable: the entire dependence structure is determined by pairwise dependencies. We will make good use of this fact later when we get to the local Gaussian dependence measure in Section 6.

(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 $\rho(t)$. For linear (say ARMA) type series the analysis as a rule is based on the autocovariance function, even though the entire joint probability structure cannot be captured by this in the non-Gaussian case. 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. Similar considerations have been made in spatial models such as in linear Kriging models, see Stein (1999).

In spite of these assets, there are several serious weaknesses of Pearson’s $\rho$. 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 $\rho$ 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 modelling. 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 document.
2. WEAKNESSES OF PEARSON’S $\rho$

We have subsumed, somewhat arbitrarily, the problems of Pearson’s $\rho$ 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 e.g. Gómez, Gómez-Villegas, and Mari’in (2003).

Unfortunately, the equivalence between uncorrelatedness and independence is generally not true for elliptical distributions. Consider for instance the multivariate $t$-distribution with $\nu$ degrees of freedom

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

Unlike the multinormal distribution where the exponential form of the distribution forces the distribution to factor if $\Sigma$ is a diagonal matrix (uncorrelatedness), this is not true for the $t$ distribution defined in equation (2.1) if $\Sigma$ 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 $\rho$ in measuring dependence in $t$ distributions, and indeed in general elliptical (and of course non-elliptical) 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 $\rho$, 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

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

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’s $\rho$ to certain nonlinearities of the data which are treated in the next subsection.

Another way of using the ranks is the Kendall’s $\tau$ 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...
Fig 1: Illustration of some problems related to the Pearson correlation coefficient

\[ X_i > X_j \text{ and } Y_i > Y_j \text{ or if both } X_i < X_j \text{ and } Y_i < Y_j. \] Similarly, they are said to be discordant if \( X_i > X_j \text{ and } Y_i < Y_j \) or if \( X_i < X_j \text{ and } Y_i > Y_j. \) 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 \( \tau \) is then given by

\[
\hat{\tau} = \frac{\text{(number of concordant pairs)} - \text{(number of discordant pairs)}}{n(n-1)/2}.
\]

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

2.3 The nonlinearity issue

This is probably the most serious issue with Pearson’s \( \rho, \) 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

\[
Y = X^2.
\]

Here, \( Y \) is uniquely determined once \( X \) is given; i.e., basically the strongest form of dependence one can have. Nevertheless, if \( X \) has a symmetric distribution on the real line, and if sufficient moments exist in the case of \( \rho, \) then the population quantities corresponding to these three test statistics are all zero. A version of this situation is illustrated in Figure 1c, where we have generated 500 observations of the standard normally distributed independent variables \( X \) and \( \epsilon, \) and calculated \( Y \) as \( Y = X^2 + \epsilon. \) Still, \( \rho(X,Y) = 0. \) The sample values for Pearson’s \( \rho, \) Spearman’s \( \rho_S \) and Kendall’s \( \tau \) are \( \hat{\rho} = -0.001, \hat{\rho}_S = -0.03 \) and \( \hat{\tau} = -0.02 \) respectively, and
none of them are significantly different from zero. It may be noted that there is a consistent non-negative modification of Kendall’s $\tau$; 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 $E(U) = E(V) = 0$, 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 iid 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

\begin{equation}
X_t = \varepsilon_t h_t^{1/2}, \quad h_t = \alpha + \beta h_{t-1} + \gamma X_{t-1}^2,
\end{equation}

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 $\beta$ and $\gamma$ are non-negative 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 Zakoian (2011). The work on these kind 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 $\rho$ 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 1d we see some simulated data from a GARCH(1,1)-model with $\varepsilon_t \sim$ iid $N(0, 1)$, $\alpha = 0.1$, $\beta = 0.7$ and $\gamma = 0.2$, with $X_t$ on the horizontal axis, and $X_{t-1}$ on the vertical axis. In this particular case, $\hat{\rho}(X_t, X_{t-1}) = 0.018$, 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 $\rho$.

3. BEYOND PEARSON’S $\rho$: 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. An exception is the multivariate normal distribution of course, but even for elliptical distributions the «correlation» parameter $\rho$ is not, as we have seen, necessarily a good measure of 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

\begin{equation}
F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))
\end{equation}
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). For simplicity and in keeping with the assumptions in the rest of this paper we will mostly limit ourselves to the continuous case.

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

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

in the two-dimensional case. It can be extended to higher dimensions, still with one scalar parameter $\theta$, but with $\theta \in (0, \infty)$. Generally $\theta \to 0$ leads to independence, while $\theta \to \infty$ leads to stronger dependence. For a connection between Kendall’s $\tau$ and the copula parameter $\theta$, 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

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

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

Figure 2a 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 2b displays the uniform scores of the same data, i.e. $(\hat{U}_{1i}, \hat{U}_{2i}) = (\hat{F}_1(X_{1i}), \hat{F}_2(X_{2i}))$ where $\hat{F}_1$ and $\hat{F}_2$ 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 2b, then, is a scatter diagram of the copula dependence function $C$ in the formula (3.1) when $p = 2$. In Figure 2c the observations have been transformed to normal scores given by $\tilde{Z}_i = \Phi^{-1}(\hat{F}(X_i)), i = 1, 2$, where $\Phi^{-1}$ 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 than is the case of Figure 2b. Finally, Figure 2d shows the scatter plot of 784 simulated pairs of variables, on uniform scale, from a Clayton copula fitted to the return data. The plot resembles Figure 2b, 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 2a, and perhaps more clearly in Figure 2c, 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 $\theta$ of the Clayton copula in terms of such type of dependence. In fact, in this particular case, $\hat{\theta} = 0.96$. 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, e.g. 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 non-negative. This means that they cannot distinguish between negative and positive dependence, whatever this may mean in the general nonlinear case. We will return to this later in the paper.

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. This is not so easily done for parametric densities, since the dependence on parameters in the test functional may be very complex, and tests of independence may be more sensibly stated in terms of the parameters themselves, as is certainly the case for the Gaussian distribution. Therefore 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 7 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) $0 \leq \delta(X, Y) \leq 1$.
(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 $\rho$.

The product moment correlation $\rho$ satisfies only (i), (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’) : $-1 \leq \delta(X, Y) \leq 1$, (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 $\rho$. It is constructed to avoid the problem demonstrated in Section 2.3 that $\rho$ 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

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

where $\rho$ is Pearson’s $\rho$. 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 $\rho$. 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
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 $F_X F_Y$ 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 $\Delta$ is a metric, and it is natural to require, see e.g. Skaug and Tjøstheim (1996), that

$$\Delta(F_{X,Y}, F_X F_Y) \geq 0$$

and

$$\Delta(F_{X,Y}, F_X F_Y) = 0 \text{ if and only if } F_{X,Y} = F_X F_Y.$$  

(4.1)

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

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

$$\hat{\Delta}(F_{X,Y}, F_X F_Y) = \Delta(\hat{F}_{X,Y}, \hat{F}_X \hat{F}_Y),$$

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

$$\hat{F}_X(x) = \frac{1}{n} \sum_{j=1}^{n} 1(X_j \leq x) \quad \hat{F}_Y(y) = \frac{1}{n} \sum_{j=1}^{n} 1(Y_j \leq y)$$

and

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

or a normalized version with $n^{-1}$ replaced by $(n+1)^{-1}$ for given observations $\{(X_1,Y_1), \ldots, (X_n,Y_n)\}$.

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

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

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

$$\Delta_2(F,G) = \int \left\{ F(x,y) - G(x,y) \right\}^2 \ dF(x,y)$$

Here both $\Delta_1$ and $\Delta_2$ 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 iid 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 iid 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 cumulative distribution function for all subsets of \( X_1, \ldots, X_p \). 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

\[
\phi_{X,Y}(u,v) = \mathbb{E}(e^{iuX+ivY}), \quad \phi_X(u) = \mathbb{E}(e^{iuX}), \quad \phi_Y(v) = \mathbb{E}(e^{ivY}).
\]

This was exploited by Csörgö (1985) and Pinkse (1998) to construct tests for independence based on the characteristic function in the iid 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

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

By taking Fourier transform of this quantity one obtains

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

Hong (1999) called (4.2) the generalized spectral density function. Here \( f(\omega, u, v) \) can be estimated by

\[
\hat{f}_n(\omega, u, v) = \frac{1}{2\pi} \sum_{k=-n+1}^{n-1} (1 - |k|/n)^{1/2}w(k/b)\hat{\sigma}_k(u, v)e^{-ik\omega},
\]

where \( w \) is a kernel weight function, \( b \) is a bandwidth or lag order, and

\[
\hat{\sigma}_k(u, v) = \hat{\phi}_k(u, v) - \hat{\phi}_k(u, 0)\hat{\phi}_k(0, v)
\]

with

\[
\hat{\phi}_k(u, v) = (n - |k|)^{-1} \sum_{t=|k|+1}^{n} e^{iuX_t + ivX_{t-|k|}}.
\]

Under the null hypothesis of serial independence \( f(\omega, u, v) \) becomes a constant function of frequency \( \omega \):

\[
f_0(\omega, u, v) = \frac{1}{2\pi} \sigma_0(u, v)
\]

with \( \sigma_0(u, v) = \phi(u + v) - \phi(u)\phi(v) \), where \( \phi(\cdot) = \phi_{X_1}(\cdot) \). In order to test for independence one can compare \( \hat{f}_n(\omega, u, v) \) and \( f_0(\omega, u, v) \) using e.g. an \( L^2 \)-functional. 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 iid 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. Thus, Pinkse uses a weight function $g$ in the functional

$$\int g(u)g(v)|\varphi(u,v)|^2\,du\,dv$$

where for a pair of two random variables $(X, Y)$

$$\varphi(u, v) = E\left(e^{iuX+ivY}\right) - E\left(e^{iuX}\right)E\left(e^{ivY}\right).$$

Let $h(x) = \int e^{iux}g(u)\,du$. Pinkse in his simulation experiments chose the weight functions $h(x) = \exp(-\frac{1}{2}x^2)$ and $h(x) = 1/(1 + x^2)$.

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 non-time 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 iid 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 $\phi_{X,Y}(u, v) = E\left(e^{i\langle X,u\rangle + \langle Y,v\rangle}\right)$, $\phi_X(u) = E\left(e^{i\langle X,u\rangle}\right)$ and $\phi_Y(v) = E\left(e^{i\langle Y,v\rangle}\right)$ 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

$$V^2(X, Y; w) = \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(u, v) - \phi_X(u)\phi_Y(v)|^2 w(u, v)\,du\,dv,$$

where $w$ is a weight function to be chosen. Note that it is easy to choose $w$ so that $V^2(X, Y) = 0$ if and only if $X$ and $Y$ are independent. Similarly, one defines

$$V^2(X; w) = \int_{\mathbb{R}^p} |\phi_{X,X}(u, v) - \phi_X(u)\phi_X(v)|^2 w(u, v)\,du\,dv.$$
and $\mathcal{V}^2(Y; w)$. The distance correlation, $dcor$, is next defined by, assuming $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$,

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

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

$$V^2_n(X, Y; w) = \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}^n(u, v) - \phi_X^n(u)\phi_Y^n(v)|^2 w(u, v) \, du \, dv,$$

where, for a set of observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ 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

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

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

$$(4.6)\ w(u, v) = (c_pc_q|u|^{1+p}|v|^{1+q})^{-1}$$

is a good choice. Here $|\cdot|$ is the Euclidean norm in $\mathbb{R}^p$ and similarly for $|\cdot|$.

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 $V_n$ converges to the theoretical functionals $V$ 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, in particular the odd symmetry of products of cosines and sines which makes corresponding integrals disappear. 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

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

for $0 < \alpha < 2$ with

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

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

$$w(u, v; \alpha) = (C(p, \alpha)C(q, \alpha)|u|^{p+\alpha}_p|v|^{q+\alpha}_q)^{-1}.$$
With the simplification $\alpha = 1$ all of this implies that $V^2_n$ as defined in (4.5), can be computed as
\[ V^2_n(u, v) = S_1 + S_2 - 2S_3 \]
where
\[ S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q, \]
\[ S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2} \sum_{k,l=1}^{n} |Y_k - Y_l|_q, \]
(4.8)
\[ S_3 = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q \]
which explains the appellation distance covariance. In fact, it is possible to further simplify this by introducing
\[ a_{kl} = |X_k - X_l|_p, \quad \bar{a}_k = \frac{1}{n} \sum_{l=1}^{n} a_{kl}, \quad \bar{a}_l = \frac{1}{n} \sum_{k=1}^{n} a_{kl}, \]
\[ \bar{a}_. = \frac{1}{n^2} \sum_{k,l=1}^{n} a_{kl}, \quad A_{kl} = a_l - \bar{a}_k - \bar{a}_l + \bar{a}_., \]
for $k, l = 1, \ldots, n$. Similarly, one can define $b_{kl} = |Y_k - Y_l|_q$ and $B_{kl} = b_{kl} - \bar{b}_k - \bar{b}_l + \bar{b}_.$ and
\[ V^2_n(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl} \]
and
\[ V^2_n(X) = V^2_n(X, X) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2 \]
and similarly for $V^2_n(Y)$. From this one can easily compute $\mathcal{R}^2_n(X, Y)$. The computations are available in an R-package: 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 $V(X, Y)$ is given by (assuming $E|X|_p < \infty$ and $E|Y|_q < \infty$)
\[ V^2(X, Y) = E_{X, X', Y, Y'} \{|X - X'|_p |Y - Y'|_q \}
+ E_{X, X'} \{|X - X'|_p E_{Y, Y'} \{|Y - Y'|_q \}
- 2E_{X, Y} \{E_{X'}|X - X'|_p E_{Y'}|Y - Y'|_q \} \]
(4.9)
where $(X, Y), (X', Y')$ are iid. This expression will be useful later in Section 4.4 in a comparison with the HSIC statistic. Properly scaled $V^2_n$ has a limiting behavior under independence somewhat similar to that described in Theorem 2 of Skaug.
and Tjøstheim (1993a), see also eq. (1.2) in Section 1 in the online supplement. Namely, under the condition of existence of first moment, \( Q_n = n\gamma^2_n/S_2 \), converges in distribution to a quadratic form

\[
Q_n \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j Z_j^2
\]

where \( \{Z_j\} \) are independent standard normal variables, \( \{\lambda_j\} \) are non-negative constants that depend on the distribution of \((X,Y)\). One can also obtain an empirical process limit theorem, Theorem 5 of Székely, Rizzo, and Bakirov (2007).

In the R-package, 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 \( \alpha \), a non-zero real number \( \beta \) and an orthogonal matrix \( C \) such that \( Y = \alpha + \beta X C \), 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 \( \rho \) 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 \( E(X) = E(Y) = 0 \) and \( \text{Var}(X) = \text{Var}(Y) = 1 \) and with correlation \( \rho \), then \( \mathcal{R}(X,Y) \leq |\rho| \) 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 e.g. 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 $F$ of functions $f$ on a set $X$, such that the evaluation functional $f \to f(x)$ is a continuous linear functional on $F$ for every $x \in X$. Then, from the Riesz representation theorem, Muscat (2014), chapter 10, there exists an element $k_x \in F$ such that $\langle f, k_x \rangle = f(x)$, where $\langle \cdot, \cdot \rangle$ is the inner product in $F$. Applying this to $f = k_x$ and another point $y \in X$ we have $\langle k_x, k_y \rangle = k_x(y)$. The function $(x, y) \to k_x(y)$ from $X \times X$ to $\mathbb{R}$ is the kernel of the RKHS $F$. It is symmetric and positive definite because of the symmetry and positive definiteness of the inner product in $F$. We use the notation $k(x, y)$ for the kernel.

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

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

$$||C||_{HS}^2 = \sum_{i,j} (Cv_j, u_i)_F^2 < \infty$$

where $u_i$ and $v_j$ are orthonormal bases of $F$ and $G$, respectively. The HS-norm generalizes the Frobenius 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 $f \in F$ and $g \in G$, then the tensor product operator $f \otimes g : G \to F$ is defined by

$$(f \otimes g)h = f(g, h)_G, \quad h \in G.$$ 

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

$$||f \otimes g||_{HS}^2 = ||f||_F^2 ||g||_G^2.$$ 

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 $(X, \Gamma)$ and $(Y, \Lambda)$ are furnished with probability measures $p_X, p_Y$, and with $\Gamma$ and $\Lambda$ being $\sigma$-algebras of sets on $X$ and $Y$. The expectations $\mu_X \in F$ and $\mu_Y \in G$ are defined by, $X$ and $Y$ are stochastic variables in $(X, \Gamma)$ and $(Y, \Lambda)$, respectively,

$$\langle \mu_X, f \rangle_F = E_X[f(X)]$$

and

$$\langle \mu_Y, g \rangle_G = E_Y[g(Y)]$$

where $\mu_X$ and $\mu_Y$ are well-defined as elements in $F$ and $G$ because of the Riesz representation theorem. The norm is obtained by

$$||\mu_X||_F^2 = E_{X,X'}[k(X, X')]$$.
where as before \( X \) and \( X' \) are independent but have the same distribution \( p_X \), and where \( ||\mu_Y|| \) 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

\[
C_{X,Y} = \mathbb{E}_{x,y}[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)] = \mathbb{E}_{x,y}[(\phi(x) \otimes \psi(y)] - \mu_x \otimes \mu_y.
\]

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

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

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

\[
\begin{align*}
\text{HSIC}(p_{XY}, \mathcal{F}, \mathcal{G}) &= \mathbb{E}_{x,x',y,y'}[k(X, X') l(Y, Y')] + \mathbb{E}_{x,x'}[k(X, X')] \mathbb{E}_{y,y'}[l(Y, Y')] \noalign{\hfill (4.10)\noalign{\hfill}
- 2 \mathbb{E}_{x,y}[k(X, X') \mathbb{E}_{y'}[l(Y, Y')]].
\end{align*}
\]

Existence is guaranteed if the kernels are bounded. The similarity in structure to (4.9) for the distance covariance should be noted (partly due to the identity \( (a - b)^2 = a^2 + b^2 - 2ab \) but going deeper as will be seen below when HSIC is compared to dcov). 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 is a mild continuity requirement on the kernel) on compact domains \( \mathcal{X} \) and \( \mathcal{Y} \), then \( ||C_{XY}||_{HS} = 0 \) 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 Höffding (1963), that is being used actively in the proof.

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

\[
(4.11) \quad \text{HSIC}_n(X, Y, \mathcal{F}, \mathcal{G}) = (n - 1)^{-2} \text{tr}\{KHLH\}
\]

where \( \text{tr} \) is the trace operator and the \( n \times n \) matrices \( H, K, L \) are defined by

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

where \( \delta_{ij} \) is the Kronecker delta. It is shown in Gretton et al. (2005) that this estimator converges towards \( ||C_{XY}||_{HS} \). 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 $\lambda_i$ and $\eta_j$ 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 a recent R-package 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, i.e 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. There is a related paper by Lyons (2013) which obtains similar results but not in an explicit RKHS context, in fact in a general dcov context.

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 Micheaux (2009), and Böttcher et al. (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. We also refer to the note on p. 2787 in Böttcher et al. (2019), explaining the relationship between the results from these two papers.

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 iid and time series case, see e.g. Rosenblatt (1975), Robinson (1991), Skauge 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 $\Delta$ is now the distance measure between two bivariate density functions. The variables may be normalized with $E(X) = E(Y) = 0$ and $\text{Var}(X) = \text{Var}(Y) = 1$. 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
\[
\Delta = \int B(f_X(x), f_Y(y), f_{X,Y}(x,y)) f_{X,Y}(x,y) \, dx \, dy
\]
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
\[
\Delta = \int D \left\{ \frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)} \right\} f_{X,Y}(x,y) \, dx \, dy
\]
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
\[
H = \int \left\{ \sqrt{f_{X,Y}(x,y)} - \sqrt{f_X(x) f_Y(y)} \right\}^2 \, dx \, dy
\]
\[
= 2 \int \left\{ 1 - \sqrt{\frac{f_X(x) f_Y(y)}{f_{X,Y}(x,y)}} \right\} f_{X,Y}(x,y) \, dx \, dy
\]
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 $D(w) = -\ln w$,
\[
I = \int \ln \left\{ \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right\} f_{X,Y}(x,y) \, dx \, dy.
\]
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 $\chi^2$-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 $g$, $\Delta$ may be estimated by $\hat{\Delta} = \Delta(\hat{f}, \hat{g})$. There are several ways of estimating the densities, e.g. the kernel density estimator,

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

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. Skog and Tjostheim (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 $\hat{\Delta}$ for the estimated Kullback-Leibler functional $\hat{I}$ and the estimated Hellinger functional $\hat{H}$ 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. One may anticipate that it picks up higher-order terms of the Edgeworth expansion (Hall 1992, Chapters 3 and 4), although no rigorous analysis to confirm this has been carried out for the functionals discussed here.

### 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 $r$, e.g. $X^2 + Y^2 = r^2 + \epsilon$ for some stochastic noise variable $\epsilon$, 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).

Two additional papers in this category are Wang et al. (2015) and Wang et al.
(2017). In both papers the authors define the locality by means of a neighborhood
of $X$ and then consider suitable $Y$-values. Consequently, as remarked by Wang et
al. (2015) their test may be best suited to nonlinear regression alternatives of the
form $Y_i = f(X_i) + \varepsilon_i$.

Finally, the so-called BDS test named after its originators Brock, Dechert and
Scheinkman (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 $\rho$: LOCAL DEPENDENCE

The test functionals treated in Section 4 deal with the second aspect of model-
ing dependence stated in the beginning of that section, namely that of testing of
independence. These functionals all do so by the computation of one non-negative
number, which is derived from local properties in Section 4.5. This number proper-
ly 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 $\rho$, 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 e.g. (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 e.g. (4.3) and (4.4).

In Section 6 the main story will be the treatment of a local Gaussian corre-
lation 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

$$P(X \leq x, Y \leq y) \geq P(X \leq x)P(Y \leq y) \text{ for all } (x,y).$$

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,

$$E(XY) - E(X)E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F_{XY}(x,y) - F_X(x)F_Y(y)) \, dx \, dy.$$ 

It follows immediately from definitions that if $(X,Y)$ is positively quadrant dependent (negatively quadrant dependent), then for Pearson’s ρ, $\rho \geq 0$ ($\rho \leq 0$). Similarly, it is shown by Lehmann that if $F_{X,Y}$ is positively quadrant dependent, then Kendall’s $\tau$, Spearman’s $\rho_S$, and the quadrant measure $q$ defined by Blomqvist (1950) are all non-negative. The paper by Blomqvist is an even earlier paper where positive and negative dependence were considered in a nonlinear case, and using quadrants centered at the median. 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 non-local 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 \( \rho \) but in various regions of the sample space, e.g. 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 \( \rho \). This is unfortunate because if the data happen to be Gaussian, one would like estimated correlations to be close or identical to \( \rho \) 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 \( \rho \) by conditioning on \( X \) in a nonlinear regression model. The resulting correlation curve inherits many of the properties of \( \rho \), and it succeeds in several of the cases where \( \rho \) fails to detect dependence, such as the parabola (2.2) in Section 2.3. However, unlike \( \rho \), it is not symmetric in \((X,Y)\). This is of course due to the conditioning on and regression on \( X \). Conditioning and regression on \( Y \) would in general produce a different result. This brings out the difference between regression analysis and multivariate analysis, where \( \rho \) 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 \( \gamma \) does not take values between -1 and 1, and it does not reduce to \( \rho \) in the Gaussian bivariate case. Actually, in that case

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

As for the correlation curve, the local dependence function works well for the example \( Y = X^2 + \varepsilon \), where \( \rho = 0 \), see (2.2), producing dependence proportional to \( x \) and producing a sign of the dependence in accordance with intuition.

6. BEYOND PEARSON’S \( \rho \): LOCAL GAUSSIAN CORRELATION

The Pearson \( \rho \) 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 we get close to a complete local characterization of dependence using the local correlation \( \rho(x, y) \),
which is the Pearson’s $\rho$ 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 $\psi$, $\sigma(x) = (\sigma_1(x), \sigma_2(x))$ the vector of standard deviations and $\rho(x)$ the correlation of $\psi$, the approximating density $\psi$ is given by

$$
\psi(v, \mu_1(x), \mu_2(x), \sigma_1^2(x), \sigma_2^2(x), \rho(x)) = \frac{1}{2\pi\sigma_1(x)\sigma_2(x)\sqrt{1 - \rho^2(x)}} \times \exp \left[-\frac{1}{2} \frac{1}{1 - \rho^2(x)} \left( \frac{(v_1 - \mu_1(x))^2}{\sigma_1^2(x)} - 2\rho(x) \frac{(v_1 - \mu_1(x))(v_2 - \mu_2(x))}{\sigma_1(x)\sigma_2(x)} + \frac{(v_2 - \mu_2(x))^2}{\sigma_2^2(x)} \right) \right].
$$

Moving to another point $y = (y_1, y_2)$ in general gives another approximating normal distribution $\psi_y$ 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 their estimation method using local likelihood is applicable also for estimating local parameters, and will be followed here.

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 $b \to 0$ component-wise.

A suitable function measuring the difference between $f$ and $\psi$ is defined by

$$
q = \int K_b(v - x)[\psi(v, \theta(x)) - \ln\{\psi(v, \theta(x))\}f(v)] \text{dv}
$$

where $K_b(v - x) = (b_1b_2)^{-1}K_1(b_1^{-1}(v_1 - x_1))K_2(b_2^{-1}(v_2 - x_2))$ is a product kernel. As is seen in Hjort and Jones (1996, pp 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

$$
\int K_b(v-x) \frac{\partial}{\partial \theta_j} \ln \{\psi(v, \theta(x))\} f(v) - \psi(v, \theta(x)) \, dv = 0, \quad j = 1, \ldots, 5.
$$

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 pp. 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 $\hat{\theta}(x) = \theta_b(x)$ is obtained from maximizing a local likelihood. Given observations $X_1, \ldots, X_n$ the local log likelihood is determined by

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

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

$$
u_j(x, \theta) = \frac{\partial}{\partial \theta_j} \ln \psi(x, \theta),
$$

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

$$
\frac{\partial L}{\partial \theta_j} = n^{-1} \sum_i K_b(X_i - x) \nu_j(X_i, \theta_b(x)) - \int K_b(v-x) \nu_j(v, \theta_b(x)) \psi(v, \theta_b(x)) \, dv
$$

$$
\to \int K_b(v-x) \nu_j(v, \theta_b(x)) [f(v) - \psi(v, \theta_b(x))] \, dv.
$$

Putting the expression in the first line of (6.3) equal to zero yields the local maximum likelihood estimate $\hat{\theta}_b(x) = \hat{\theta}(x)$ 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

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

where $b^T$ is the transposed of $b$. It is seen that ignoring solutions that yield $\nu_j(x, \theta_b(x)) = 0$ requires $\psi(x, \theta_b(x))$ to be close to $f(x)$. 
An asymptotic theory has been developed in Tjøstheim and Hufthammer (2013) for \( \hat{\theta}_b(x) \) for the case that \( b \) is fixed and for \( \hat{\theta}(x) \) in the case that \( b \rightarrow 0 \). 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 \( b \rightarrow 0 \), this leads to a slow convergence rate of \( (n(b_1b_2)^3)^{-1/2} \), 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 \( \rho \). 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 LGC, at least for a number of purposes it is advantageous to replace a scaling with uniform

\[
Z = (Z_1, Z_2) = (\Phi^{-1}(F_1(X_1)), \Phi^{-1}(F_2(X_2))),
\]

where \( \Phi \) 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 \( \rho_Z(z_1, z_2) \) with \( \hat{Z}_i = \Phi^{-1}(\hat{F}_i) \). Under certain regularity conditions, as in the copula case, the difference between \( Z \) and \( \hat{Z} \) can be ignored in limit theorems. Using the sample of pairs of Gaussian pseudo observations \( \{\Phi^{-1}(\hat{F}_1(X_{1i})), \Phi^{-1}(\hat{F}_2(X_{2i}))\} \), \( i = 1, \ldots, n \) 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 \( \mu_{Z_i}(z) = 0 \) and \( \sigma_{Z_i}(z) = 1 \), in which case the asymptotic theory simplifies and one obtains the familiar nonparametric rate of \( O((nb_1b_2)^{-1/2}) \) for \( \hat{\rho}_Z(z) \).

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 are 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 \textit{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 iid 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; 2019).

Locally Gaussian spectral estimation is contained in Jordanger and Tjøstheim (2020). They have shown that nonlinear and local oscillatory behaviour 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 (2018) 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 in 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 $\rho$.

ACKNOWLEDGEMENTS

We are grateful to the editor and to two anonymous referees for a number of valuable comments and suggestions. This work has been partially supported by the Finance Market Fund (Norway), grant number 261570.

REFERENCES


Berentsen, Geir Drage, Bård Støve, Dag Tjøstheim, and Tommy Nordbø. 2014.


Online Library: 110–33.


Reshef, David, Yakir Reshef, Michael Mitzenmacher, and Pardis Sabeti. 2013. “Equitability Analysis of the Maximal Information Coefficient, with Comparisons.”


