Statistics 28 (1996) pp 249-284
Reprints available directly from the publisher
Photocopying permitted by license only

MEASURES OF DEPENDENCE AND TESTS OF INDEPENDENCE

DAG TJØSTHEIM

University of Bergen, Norway

(Received in final form 20 June 1995)

Measures of dependence and resulting tests of independence are surveyed. Measures arising both from linear and nonlinear modeling are examined. Tests based on chaos theory are briefly discussed. The main emphasis, however, is on some recently developed nonparametric tests using estimated distribution and density functions. Most of the paper is phrased in terms of serial dependence for a univariate stationary time series, but it is indicated how more general situations can be analysed. The bootstrap is an essential tool for determining the critical value of the new tests.

Keywords: Hellinger distance; Kernel estimation; Kullbach-Leibler information; tests of independence

1. INTRODUCTION

In virtually every field of statistics there is a need for measuring dependence between stochastic variables and for constructing accompanying tests of independence. A traditional measure of dependence is the correlation function. It is suitable for measuring dependence in Gaussian models and to some degree in general linear models. With the strong recent emphasis on nonlinear and non-Gaussian models it is not surprising that there has been a search for alternative dependence measures and tests of independence. In this paper I will give a brief and somewhat subjective survey of parts of these developments. Having myself mainly worked in a time series setting, I will stress the aspect of serial dependence, but many of the results are relevant for looking at dependencies between arbitrary stochastic variables or stochastic processes. Also the emphasis is on the testing aspect, and measures of dependence are often only considered as an intermediate step in obtaining the tests.

Roughly, applications can be subdivided into two groups: First, the need for tests of independence may originate from the model formulation in a particular problem. For example in many physical systems the noise is assumed to consist of independent identically distributed (iid) random variables, and such a hypothesis should be testable. Another example can be found in econometrics, where it may be postulated that certain variables are exogenous, whereas others are endogenous resulting in one sided dependence structures which one would like to test. Moreover, for certain financial time series economic theories have been put forward which predict that logarithmic differences should be iid (cf. Fama 1965, 1970).

The other main area of application is that of statistical model fitting. Often it is a goal to obtain a model where there is no structural information left in the residuals, i.e. where these are iid. Once more a test of independence is needed, but here the situation is more complex as the estimated residuals are influenced by errors in the parameter estimates. Only to a very little degree have these effects been incorporated in recent theory.

Finally, it should be mentioned that there are connections between independence tests and linearity tests. We can test linearity by testing for independence of the residuals from a linear model fit. Vice versa linearity tests involving the conditional mean and the conditional variance can be exploited to test implications of an iid hypothesis.

A short summary of the paper is as follows: In section 2 we look at correlation based measures and generalizations to higher order moment measures and tests. The correlation integral of chaos theory and the so-called BDS test are discussed in Section 3. Independence between stochastic variables is defined in terms of distribution functions, or, if they exists, in terms of density functions. It is natural to try to exploit this in constructing tests. Both historic and some more recent work in this area are summarized in Sections 4 and 5. The main thrust of the paper lies here, and many of the results are from Skaug and Tjøstheim (1993 a, b and 1994). The tests we look at are nonparametric, and they are designed to have fairly decent power properties against a wide range of alternatives. If one wants to test against a specific alternative, one should use a test designed for such an alternative, e.g. an ARCH test (Engle 1982) for ARCH type dependence. Some such tests are discussed by Pagan and Hall (1983) and White (1987). General asymptotic theory for functionals appearing in nonparametric testing and estimation is briefly outlined in the Appendix.

2. CORRELATION BASED MEASURES AND THEIR EXTENSIONS

2.1. The Correlation Function

Consider two random variables X and Y having second order moments. The correlation function is defined by

$$\operatorname{corr}(X, Y) = \rho(X, Y) = \frac{\mathrm{E}[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}]}{(\mathrm{E}[\{X - \mathrm{E}(X)\}^2])^{1/2}(\mathrm{E}[\{Y - \mathrm{E}(Y)\}^2])^{1/2}}$$

and $-1 \leqslant \rho \leqslant 1$ with $\rho = 0$ if X and Y are independent. If X and Y are Gaussian, independence is equivalent to X and Y being uncorrelated.

For a vector $\underline{X} = [X_1, \ldots, X_d]$ of random variables the dependencies as expressed in terms of correlations are summed up by the correlation matrix $\rho = (\rho_{ij})$. The dependencies are completely described by *pairwise* relationships, as are all partial and multiple correlations of interest. Also, of course, for Gaussian random variables the entire dependence structure is described by the pairwise covariances and the individual variances. Unfortunately, the reduction to pairwise relations cannot in general be attained for other dependence measures and non-Gaussian variables.

The correlation is estimated by

$$\hat{\rho}(X, Y) = \frac{n^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\{n^{-1}\Sigma_{i=1}^{n}(X_i - \bar{X})\}^{1/2}\{n^{-1}\Sigma_{i=1}^{n}(Y_i - \bar{Y})\}^{1/2}}$$

where $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are pairs of observations having the same distribution as (X, Y).

For a stationary process $\{X_t\}$ whose second moment exists, the autocorrelation function at lag k is defined by

$$\rho(k) = \operatorname{corr}(X_t, X_{t-k})$$

and is estimated by

$$\hat{\rho}(k) = \frac{n^{-1}\Sigma_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})}{n^{-1}\Sigma_{t=1}^{n}(X_t - \bar{X})^2}$$

Tests of independence can be based on the asymptotic distribution of $\hat{\rho}$. In the case of two stationary processes $\{X_t\}$ and $\{Y_t\}$ assume that

$$X_t - \mu_x = \sum_{j=-\infty}^{\infty} \alpha_j Z_{t-j,1}$$

$$Y_t - \mu_y = \sum_{j=-\infty}^{\infty} \beta_j Z_{t-j,2}$$

where $\Sigma|\alpha_j| < \infty$, $\Sigma|\beta_j| < \infty$, where $\mu_x = E(X_t)$, $\mu_y = E(Y_t)$, and where $\{Z_{t,1}\}$ and $\{Z_{t,2}\}$ are independent sequences of iid random variables having second moments. Then $\hat{\rho}_{x,y}(k) \xrightarrow{a.s} \rho_{x,y}(k) = 0$ as $n \to \infty$. Here $\rho_{x,y}(k)$ is the cross correlation between $\{X_t\}$ and $\{Y_t\}$ at lag k. Moreover, (Brockwell and Davis 1987, p. 400) $\hat{\rho}_{xy}(k)$ is asymptotically normal with zero mean and variance

$$n^{-1} \sum_{j=-\infty}^{\infty} \rho_{xx}(j)\rho_{yy}(j)$$

where ρ_{xx} and ρ_{yy} are the autocorrelation functions of $\{X_t\}$ and $\{Y_t\}$, respectively. If the pairs $\{(X_t, Y_t), t = 1,\ldots, n\}$ are iid, then it is a classical result of multivariate analysis (Anderson 1958, p. 77) that

$$\sqrt{n}(1-\rho^2)^{1/2}(\hat{\rho}_{xy} - \rho_{xy}) \xrightarrow{\mathscr{L}} \mathcal{N}(0, 1).$$

However, it is well known that the sampling distribution is appreciably skewed for quite substantial sample sizes, and a better approximation to normality is obtained by looking at the ratio

$$\hat{U} = \frac{1}{2}\log\frac{1+\hat{\rho}_{xy}}{1-\hat{\rho}_{xy}}$$

where now $\sqrt{n}(\hat{U} - U) \xrightarrow{\mathscr{L}} \mathcal{N}(0, 1)$. It does not seem, though, that this transformation is much used in general time series analysis.

One might think that an alternative way of picking up the skewness of the sampling distribution is to use the bootstrap since it typically includes the second order term of the Edgeworth expansion (Hall 1992, p. 83). But it is known that the correlation function is problematic to handle with the ordinary bootstrap (Efron 1982, Hall 1992), as pivoting is difficult due to large variance in the variance estimate of $\hat{\rho}$ (Hall 1992, p. 141, p. 150) and rather careful analysis involving an iterated bootstrap and/or a bias correction should be done in the construction of confidence intervals.

In the case of serial correlation similar results are: For a linear process $\{X_t\}$ given by

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \alpha_j Z_{t-j}$$

with $\Sigma|\alpha_j| < \infty$ and $\Sigma|j|\alpha_j^2 < \infty$, and with $\{Z_t\}$ being iid and having a second moment, the estimated autocorrelation $\hat{\rho}(k)$ at lag k is a strongly consistent estimator of $\rho(k)$, and $\{\hat{\rho}(k), k = 1,\dots m\}$ are jointly asymptotically normal with asymptotic means $\{\rho(k), k = 1,\dots m\}$ and covariance matrix $W = (w_{ij})$ with

$$w_{ij} = n^{-1} \sum_{k=-\infty}^{\infty} \{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k)$$

$$- 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+j)\},$$

which is Barlett's formula. Alternatively such a result can be proved under mixing assumptions.

One of the prime uses of the autocorrelation function is in the test of fit. Traditionally this is done in the context of ARMA models, i.e. $\{X_t\}$ is given by

$$X_t - \mu - \sum_{i=1}^{p} a_i(X_{t-i} - \mu) = e_t + \sum_{i=1}^{q} b_i e_{t-i}$$

where the series $\{e_t\}$ consists of iid zero-mean random variables having a second moment. If we denote by $\hat{e}_t$ the estimated residuals and let

$$\hat{\rho}_e(k) = \frac{n^{-1}\Sigma_{t=k+1}^{n}(\hat{e}_t - \bar{\hat{e}})(\hat{e}_{t-k} - \bar{\hat{e}})}{n^{-1}\Sigma_{t=1}^{n}(\hat{e}_t - \bar{\hat{e}})^2},$$

then the Box-Pierce statistic for testing of fit is given as

$$\hat{S} = n \sum_{k=1}^{h} \hat{\rho}_e^2(k).$$

Under the assumption that $h = h_n \to \infty$ as $n \to \infty$ and the conditions of Box-Pierce (1970): a) $\phi_j = O(n^{-1/2})$ for $j \geqslant h_n$, where ϕ_j are the coefficients in the expansion $X_t = \Sigma_{j=0}^{\infty}\phi_j e_{t-j}$ and b) $h_n = O(n^{1/2})$ as $n \to \infty$, $\hat{S}$ is asymptotically distributed as a χ^2-variable with $h - p - q$ degrees of freedom. One should note that there is a modification of the Box-Pierce statistic due to Ljung and Box (1978) and further one due to Dufour and Roy (1986). The first is discussed in Brockwell and Davis (1987, p. 301). If $\hat{S}$ (or its Box-Ljung modified version) exceeds the upper χ^2 critical value, one concludes that the model is not well fitted, and specific deviations in the residuals can be used to suggest changes in the model.

In recent years this way of model fitting has come under rather heavy criticism, the core of the critique being that the statistic $\hat{S}$ is a measure of correlation rather than dependence. The test has a too large tendency to let through models with interesting structure in the residuals. It suffices to mention one such example: For financial series it is of interest to model the risk structure as well as the expected future value of the series. The risk is modeled in terms of the conditional variance of the residuals $\{e_t\}$. For the simplest (pure ARCH) case this is done with the model

$$e_t = \sqrt{(a + be_{t-1}^2)}\eta_t$$

where a and b are nonnegative constants, and $\{\eta_t\}$ is another series of iid zero-mean variables with variance 1. Then

$$\mathrm{var}(e_t|e_{t-1} = x) = (a + bx^2)$$

and hence if large fluctuations (large e_t's) have occurred, then large fluctuations (high risk) can be expected in the future. However, it is easy to see that $\mathrm{corr}(e_t, e_{t-k}) = 0$ for all $k \neq 0$, so that the Box-Pierce-Ljung test will fail to pick up this structure. The ARCH model was introduced by Engle (1982) and is currently extensively used by econometricians. The fact that the Box-Pierce-Ljung statistic does not detect this structure has undermined its credibility. There emerges the need for alternative statistics.

2.2. The Rank Correlation

Replacing observations by ranks generally robustifies the analysis to deviations from normality. The idea of rank correlation goes back at least to Spearman (1904). Given observations $\{X_1, \ldots, X_n\}$ we denote by $R_t^{(n)}$ the rank of X_t among $X_1, \ldots, X_n$. The estimated rank correlation function given n pairwise observations of two random variables X and Y is given by

$$r = \frac{n^{-1}\Sigma_{t=1}^n R_{t,X}^{(n)} R_{t,Y}^{(n)} - (n+1)^2/4}{(n^2-1)/12}.$$

The rank correlation is thought to be especially effective in picking up linear trends in the data. It is obvious how it can be modified to a rank autocorrelation measure for time series (Knoke 1977, Bartels 1982 and

Hallin and Mélard 1988). The asymptotic normality was established early. Some of the theory is reviewed in Kendall (1970). It can now be viewed as a special case of much more general results established by Hallin *et al.* (1985).

Although not much applied in a time series context, rank procedures have recently been systematically pursued by Hallin and Puri and their coworkers in a series of papers. See Hallin and Puri (1993) and the references therein. In particular they have looked at a number of theoretical properties of rank tests for testing iid against ARMA type alternatives. They have shown that these tests possess optimality properties using Pitman efficiency as a measure of optimality and using Le Cam's theory of local asymptotic normality (LAN). Moreover, Hallin (1994) shows, extending a classical result by Chernoff and Savage (1958), that except for the normal distribution case, where they are asymptotically equivalent, traditional correlation tests will have lower power asymptotically than the optimal normal score rank test. In this sense the correlation test is inadmissible.

It will lead too far from the main topic if I were to give a detailed account of this work, and I will only sketch a few points. Hallin and Puri typically look at linear serial rank statistics of the form

$$S^{(n)} = (n-k)^{-1} \sum_{t=k+1}^{n} a^{(n)}(R_t^{(n)}, R_{t-1}^{(n)}, \ldots, R_{t-k}^{(n)})$$

where $a^{(n)}(i_1, \ldots, i_{k+1})$ is a collection of scores defined over the set of all $(k+1)$-tuples of distinct integers in $\{1, \ldots, n\}$. Taking $a(i_1, \ldots i_{k+1}) = i_1 i_{k+1}$ essentially gives the Spearman rank correlation function. Note that Wald-Wolfowitz (1943) proposed a circular version of this. To handle time series problems satisfactorily it is necessary to introduce the linear serial signed rank statistics

$$S_+^{(n)} = (n-k)^{-1} \sum_{t=k+1}^{n} a_+^{(n)} (s_t R_{+,t}^{(n)}, s_{t-1} R_{+,t-1}^{(n)}, \ldots, s_{t-k} R_{+,t-k}^{(n)})$$

where $R_{+,t}^{(n)}$ is the rank of $|X_t|$ among $|X_1|, \ldots, |X_n|$ and s_t the sign of X_t, and $a_+^{(n)}$ is again a score function. An asymptotic theory can be derived in the general case, but as far as the rank correlation is concerned, it is sufficient to look at

$$S^{(n)} = (n-k)^{-1} \sum_{t=k+1}^{n} a^{(n)}(R_t^{(n)})\, b^{(n)}(R_{t-k}^{(n)})$$

and

$$S_{+}^{(n)} = (n-k)^{-1} \sum_{t=k+1}^{n} a_{+}^{(n)}(s_t R_{+,t}^{(n)})\, b_{+}^{(n)}(s_{t-k} R_{+,t-k}^{(n)}).$$

Under conditions essentially involving a) smoothness, b) strong mixing on $\{X_t\}$, and c) symmetry and absolute continuity with respect to Lebesgue measure for the distribution of X_t, consistency and a symptotic normality of $S^{(n)}$ and $S_{+}^{(n)}$ can be established. Moreover, using LeCam type LAN theory locally most powerful tests can be obtained for testing randomness against ARMA models, with recent extensions to bilinear models (Benghabrit and Hallin 1992). It is not yet clear to which degree these tests remain optimal and practical to use against a wider class of nonlinear alternatives such as those considered in Skaug and Tjøstheim (1994).

Related to rank tests are permutation tests. They have been considered in Dufour and Roy (1985, 1986) and Dufour and Hallin (1991).

2.3. Higher Moments and Derived Measures

An ad-hoc procedure for increasing the power of the correlation test in situations where it has little or no power, the ARCH process of Section 2.1 being one example, is to square the observations and compute the correlations of the squares. This is a well known device, and in the time series case it results in the McLeod-Li (1983) test based on

$$\hat{\rho}_{X^2}(k) = \frac{\Sigma_{t=k+1}^{n}(X_t^2 - X^2)(X_{t-k}^2 - X^2)}{\Sigma_{t=1}^{n}(X_t^2 - X^2)^2}.$$

If fourth order moments exist, and if $\{X_t\} = \{\hat{e}_t\}$ are the estimated residuals from an ARMA model of known order, then it can be shown (McLeod and Li 1983) that the asymptotic distribution of $\hat{\rho}_{X^2}(k)$ is normal and a χ^2-statistic analogous to the Box-Pierce-Ljung statistic can be constructed.

For the ARCH example of Section 2.1, $e_t^2 = (a + be_{t-1}^2)\eta_t^2$. The e_t^2's are correlated variables, and hence the McLeod-Li test will be able to pick up this type of dependence, which is one reason this test has gained popularity among econometricians. It should be noted, however, that the higher moments involved incur larger estimation errors, and that for linear models the squaring operation generally leads to a loss of power compared to the ordinary correlation based test. Moreover, its usefulness is of course limited to those models where the dependence structure can be expressed as a detectable correlation between squares of variables.

Lawrance and Lewis (1987) argue specifically that in case one has residuals from non-linear models, the autocorrelation of the squared residuals may be difficult to handle since it involves fourth order moments, and they suggest computing correlations between e_s and e_t^2, which should be one-sided for the ARCH model, or between e_s and $(X_t - \mu)^2$, or e_s^2 and $X_t - \mu$. They illustrate the use of these quantities for a few classes of nonlinear models, mainly random coefficient models. Lawrance's (1991) work on reversibility may also be relevant, as independence implies reversibility, and a good test of reversibility would presumably also constitute a good test of independence for a wide class of process.

2.4. Frequency Based Tests

If $\{X_t\}$ consists of uncorrelated random variables, then its spectral density

$$S_x(\lambda) = \sum_k e^{-i2\pi k\lambda} \operatorname{cov}(X_t, X_{t-k})$$

is a constant independent of the frequency λ. Hence, the estimated spectral density, or rather the cumulative spectral distribution function, can be used as a test of independence and as a test of fit statistic (cf. Brockwell and Davis 1987, Ch. 9.4). Roughly speaking, if properly estimated, it will have many of the same weaknesses and strengths as the Box-Pierce-Ljung test.

For non-Gaussian and nonlinear models one can look at the bispectrum $B_x(\lambda_1, \lambda_2)$, which is the Fourier transform of the third order cumulant function. For a zero mean process $\{X_t\}$,

$$B_x(\lambda_1, \lambda_2) = \sum_k \sum_m \mathrm{E}(X_t X_{t-k} X_{t-m}) \exp\{-i2\pi(\lambda_1 k + \lambda_2 m)\}.$$

For a linear process $X_t = \Sigma_i \alpha_i Z_{t-i}$ it can be shown that

$$\frac{|B_x(\lambda_1, \lambda_2)|^2}{S_x(\lambda_1) S_x(\lambda_2) S_x(\lambda_1 + \lambda_2)} = \frac{\mu_{3,z}^2}{\sigma_z^6}$$

where S_x is the spectral density of $\{X_t\}$, and $\mu_{3,z}$ and σ_z^2 are the third moment and the variance of $\{Z_t\}$, respectively. The above relationship is used primarily as a basis for testing of linearity (Hinich 1982, Subba Rao and Gabr 1980). If $\{X_t\}$ consists of iid random variables, then the bispectrum itself is a constant ($= \mathrm{E}(X_t^3)$), and it is zero if $\mu_3 = 0$. This is testable, and will in general produce information in addition to the test based on the cumulative spectral distribution. Hinich (1993) prefers to look at the bicovariances $G(k, m) = \mathrm{E}(X_t X_{t-k} X_{t-m})$, $k < m$, themselves and uses a test statistic, which is related to the Lawrence and Lewis (1987) approach, with

$$\hat{G}(k, m) = (n - m)^{-1/2} \sum_{t=m+1}^{n} X_t X_{t-k} X_{t-m},$$

and where under the null hypothesis of independence, for k and $m \neq 0$, $\mathrm{E}\{(\hat{G}(k, m)\} = 0$. A test based on this quantity would have zero power for processes for which $G(k,m) = 0$, e.g. the pure ARCH process. Hinich also proposes a test based on $\hat{G}^2(k, m)$. The test is based on sixth order moments, and, not surprisingly, requires clipping in order to work. The asymptotic theory requires moments of order 12. A number of simulations as well as experiments on a large real data set are given in Hinich (1993).

3. THE CORRELATION INTEGRAL OF CHAOS THEORY

In Grassberger and Procaccia (1983) the correlation integral was introduced as a means of measuring the fractal dimension of deterministic data. It measures serial dependence patterns in the sense that it keeps track of the frequency with which temporal patterns are repeated in a data sequence. Let $\{x_1, \ldots, x_n\}$ be a sequence of numbers and let

$$x_t^{(k)} = [x_t, x_{t-1}, \ldots, x_{t-k+1}] \quad k \leqslant t \leqslant n.$$

Then the correlation integral for embedding dimension k is given by

$$C_{k,n}(\varepsilon) = \frac{2}{n(n-1)} \sum_{1 \leqslant s < t \leqslant n} 1(\|\underline{x}_t^{(k)} - \underline{x}_s^{(k)}\| < \varepsilon)$$

where $\|\underline{x}\| = \max_{1 \leqslant i \leqslant m} |x_i|$, where $1(\cdot)$ is the indicator function and $\varepsilon > 0$ is a cut off threshold which could be a multiple of the standard deviation in the case of a stationary process. The parameter ε might also be considered to be a "fudge" or a "tuning" parameter. Let

$$C_k(\varepsilon) = \lim_{n\to\infty} C_{k,n}(\varepsilon).$$

If $\{X_t\}$ is an absolutely regular (Bradley 1986, p. 169) stationary process, the above limit exists and is given by

$$C_k(\varepsilon) = \int 1(\|\underline{x}^{(k)} - \underline{y}^{(k)}\| < \varepsilon)\, dF_k(\underline{x}^{(k)})\, dF_k(\underline{y}^{(k)}), \tag{3.1}$$

where F_k is the joint cumulative distribution function of $\underline{X}_t^{(k)}$. Since $1(\|\underline{x}^{(k)} - \underline{y}^{(k)}\| < \varepsilon) = \Pi_{i=1}^k 1(|x_i - y_i| < \varepsilon)$ it is easily seen that if $\{X_t\}$ consists of iid random variables, then

$$C_k(\varepsilon) = \{C_1(\varepsilon)\}^k,$$

and this expression can be used as a basis for a test of independence. Note that no moments of $\{X_t\}$ need exist.

The sampling properties of $C_{k,n}$ can be derived by exploiting that $C_{k,n}$ is a generalized U-statistic (Serfling 1980, Ch. 5, and Denker and Keller 1983) with a symmetric kernel $1(\|\underline{x} - \underline{y}\| < \varepsilon)$. See also Brock *et al.* (1991). Under the hypothesis of independence, and excluding the case of uniformly distributed random variables, Brock *et al.* (1991) have established that

$$\frac{\sqrt{n}\,[C_{k,n}(\varepsilon) - \{C_{1,n}(\varepsilon)\}^k]}{V_{k,n}} \to_{\mathscr{D}} \mathcal{N}(0,1)$$

where

$$\frac{1}{4}V_{k,n}^2 = k(k-2)C_n^{2k-2}(K_n - C_n^2) + K_n^k + C_n^{2k} +$$

$$2\sum_{j=1}^{k-1}[C_n^{2j}(K_n^{k-j} - C_n^{2k-2j}) - kC_n^{2k-2}(K_n - C_n^2)]$$

with

$$C_n = \frac{1}{n^2}\sum_s\sum_t 1(|X_s - X_t| < \varepsilon)$$

and

$$K_n = \frac{1}{n^3} \sum_{r=1}^{n} \sum_{s=1}^{n} \sum_{t=1}^{n} 1(|X_r - X_s| < \varepsilon) 1(|X_s - X_t| < \varepsilon).$$

Here we have stated the test in terms of a data series $\{X_t\}$. In Brock *et al.* (1991) it is shown how this can be adapted in principle and asymptotically to a test of fit situation involving estimated residuals.

The test is called the BDS test after its originators (Brock, Dechert and Scheinkman 1987). It is now reasonably well-tried. It is effective against quite a wide range of alternatives including ARCH and its generalization GARCH (Bollerslev 1986). There are problems with obtaining the right level though. For small, moderate and even quite large sample sizes cases of large over-estimation of the level have been reported giving the test a high false alarm rate (Brock *et al.* 1991 and Mizrach 1991). In power experiments on simulated data as a general rule the level has been determined by simulations, so that is does not influence the power comparisons of those experiments. For real data this cannot be done, and one has to be careful in applying the test in the form given above. Its small sample properties have been discussed in Hsieh and LeBaron (1988). In the light of the results to be reported in Section 5 the bootstrap should be a natural alternative to using the asymptotic distributions in determining critical values.

In a recent contribution Wolff (1994) obtains a Poisson law for the correlation integral under the null hypothesis of independence, and he uses nonparametric methods to specify the test precisely. His paper also incluudes a numerical comparison with the methods to be presented in Sections 4 and 5. Howell Tong and his coworkers have recently worked intensively on various aspects of chaos theory and nonlinear time series. The reader is referred to Tong (1995) and references therein.

4. MEASURES AND TESTS BASED ON THE DISTRIBUTION FUNCTION

4.1. Classes of Measures

Let X and Y be stochastic variables with distribution functions F_X and F_Y, respectively. The problem of measuring the dependence between X and Y can be formulated as a problem of measuring the distance between the two bivariate distribution functions $F_{X,Y}$ and $F_X \otimes F_Y$, where $F_{X,Y}$ is the joint distribution function of (X, Y) and $F_X \otimes F_Y(x, y) = F_X(x)F_Y(y)$. (An alternative would be to use characteristic functions. Pinkse (1993) has explored this possibility in an interesting contribution.) Let $\Delta(\cdot,\cdot)$ be a candidate for such a distance functional. It will not be assumed that Δ is a metric, but it is natural to require (Skaug and Tjϕstheim 1994) that

$$\Delta(F_{X,Y}, F_X \otimes F_Y) \geqslant 0 \text{ and } \Delta(F_{X,Y}, F_X \otimes F_Y) = 0 \text{ iff } F_{X,Y} = F_X \otimes F_Y. \quad (4.1)$$

Moreover, one may require invariance under transformations, or more precisely

$$\Delta(F^*_{X,Y}, F^*_X \otimes F^*_Y) = \Delta(F_{X,Y}, F_X \otimes F_Y) \quad (4.2)$$

where $F^*_X(x) = F_X\{h^{-1}(x)\}$ and $F^*_{X,Y}(x,y) = F_{X,Y}\{h^{-1}(x), h^{-1}(y)\}$. Here h is an increasing function, and F^*_X, F^*_Y, $F^*_{X,Y}$ are the marginal and bivariate distribution functions of the random variables $h(X)$, $h(Y)$ and $\{h(X), h(Y)\}$, respectively.

For distance functionals not satisfying (4.2) we can at least obtain scale and location invariance by standardizing such that $\mathrm{E}(X) = \mathrm{E}(Y) = 0$ and $\mathrm{var}(X) = \mathrm{var}(Y) = 1$, assuming that the second moments exist. In practice empirical averages and variances must be employed, but asymptotically the difference between using empirical and theoretical quantities is a second order effect. We have used such a standardization for all of our functionals.

The measures of dependence introduced in Sections 2 and 3 can be expressed as functionals on F_X, F_Y and $F_{X,Y}$ although not generally as distance functionals depending on $F_{X,Y}$ and $F_X \otimes F_Y$. For example, with X and Y standardized, the correlation squared can be written.

$$\rho^2 = \{ \int xy \, \mathrm{d}F_{X,Y}(x, y) - \int x \, \mathrm{d}F(x) \int y \, \mathrm{d}F(y) \}^2 \quad (4.3)$$

Similarly, the Spearman rank correlation is based on an estimate of

$$\rho^2_{RC} = [\int \{2F_X(x) - 1\}\{2F_Y(y) - 1\} \mathrm{d}F_{X,Y}(x.y)]^2. \quad (4.4)$$

Neither ρ^2 nor ρ^2_{RC} satisfy either of the conditions (4.1), (4.2). Similarly, from (3.1) it is clear that $C_2(\varepsilon)$ of the BDS statistic can be written

$$C_2(\varepsilon) = \int 1(|x_1 - x_2| < \varepsilon) 1(|y_1 - y_2| < \varepsilon) \, \mathrm{d}F_{X,Y}(x_1, y_1) \, \mathrm{d}F_{X,Y}(x_2, y_2) \quad (4.5)$$

and likewise for $C_1(\varepsilon)$. A distance functional can be constructed based on $C_2(\varepsilon) - \{C_1(\varepsilon)\}^2$. The relationships (4.1) and (4.2) will in general not be satisfied for such a functional.

Conventional distance measures between two distribution functions F_1 and F_2 are the Kolmogorov-Smirnov distance

$$\Delta_1(F_1, F_2) = \sup_{\underline{x}} |F_1(\underline{x}) - F_2(\underline{x})|$$

and the Cramer–von Mises type distance

$$\Delta_2(F_1, F_2) = \int \{F_1(\underline{x}) - F_2(\underline{x})\}^2 \mathrm{d}F_1(\underline{x}). \tag{4.6}$$

Here Δ_1 satisfies (4.1) and (4.2), whereas Δ_2 satisfies (4.1) but not (4.2). A third distance measure is the Mallows measure

$$\Delta_3(F_1, F_2) = \inf\{\mathrm{E}(X_1 - X_2)^2\}^{1/2}$$

where the infimum is taken over all pair of random variables X_1 and X_2 having marginal distributions F_1 and F_2, respectively. This measure is much used in bootstrap asymptotics.

By letting $F_1 = F_{X,Y}$ and $F_2 = F_X \otimes F_Y$ these distance functions can be taken as measures of dependence. An and Cheng (1990) have used the Kolmogorov-Smirnov distance in connection with a linearity test of theirs. It could be converted into an independence test, but apart from this, as far as I know, all the work pertaining to measuring dependence and testing of independence has been done in terms of the Cramer–Von Mises distance (Hoeffding 1948, Blum *et al.* 1961, Rosenblatt 1975, Rosenblatt and Wahlen 1992, and Skaug and Tjøstheim 1993b).

I have only considered a pair of random variables. In principles one can extend the procedure to the general vector case $[X_1, \ldots, X_k]$ with $F_1 = F_{X_1,\ldots,X_k}$ and $F_2 = \Pi_{i=1}^k \otimes F_{X_i}$. In practice, however, $F_{X_1,\ldots,X_k}$ will be difficult to estimate due to the curse of dimensionality. Instead, for a stationary process $\{X_t\}$ Skaug and Tjøstheim (1993b) have considered functionals $\Delta^{(k)}$ based on pairwise measures

$$\Delta^{(k)} = \sum_{i=1}^{k} \Delta(F_{X_t, X_{t-i}}, F_{X_t} \otimes F_{X_{t-i}})$$

but, obviously, unlike the correlation structure, the full joint dependence structure is not described by all pairwise relationships.

4.2. Estimation of Distance Functionals and Tests of Independence

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

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

where $\hat{F}$ is the empirical distribution function given by

$$\hat{F}_X = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leqslant x) \quad \hat{F}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leqslant x)1(Y_i \leqslant y)$$

for given observations $\{(X_1, Y_1),\ldots, (X_n, Y_n)\}$. Similarly for a stationary times series $\{X_t\}$,

$$\hat{F}_k(x, y) = \hat{F}_{X_t, X_{t-k}}(x, y) = \frac{1}{n-k}\sum_{t=k+1}^{n} 1(X_t \leqslant x)1(X_{t-k} \leqslant y).$$

It is noted that if this method is used in (4.3) (4.4) (4.5), it leads to the classical estimates of the correlation function and to the familiar estimate of the correlation distance used in the BDS statistic. Again, we standardize with $\mu = \mathrm{E}(X_t)$ and $\sigma = SD(X_t)$, or rather $\bar{X} = n^{-1}\Sigma X_t$ and $\hat{\sigma} = \{(n-1)^{-1}\Sigma(X_t - \bar{X})^2\}^{1/2}$.

The work centered around the Cramer–von Mises type statistic (4.6) was started already by Hoeffding (1948), who studied finite sample distributions in some special cases. An asymptotic theory was provided by Blum, Kiefer and Rosenblatt (1961) in the case of having observations (X_i, Y_i) where the pairs $\{(X_1, Y_1\},\ldots, (X_n, Y_n)\}$ are iid. This was extended to the time series case with a resulting test of serial independence in Skaug and Tjøstheim (1993b). I now give a brief review of that work.

In the time series case the Cramer–von Mises distance at lag k is given by

$$D_{F_k} = \int \{F_k(x, y) - F(x)F(y)\}^2 \mathrm{d}F_k(x, y)$$

where F_k and F are the joint and marginal distributions of (X_t, X_{t-k}) and X_t, respectively. Replacing theoretical distributions by empirical ones leads to the estimate

$$\hat{D}_{F_k} = \frac{1}{n-k} \sum_{t=k+1}^{n} \{\hat{F}_k(X_t, X_{t-k}) - \hat{F}(X_t)\hat{F}(X_{t-k}\}^2.$$

Assuming $\{X_t\}$ to be ergodic, we have Skaug and Tjøstheim (1993b, Th.1,) that $\hat{D}_{F_k} \xrightarrow{a.s.} D_{F_k}$ as $n \to \infty$.

To construct a test of serial independence we need the distribution of $\hat{D}_{F_k}$ under the assumption of $\{X_t\}$ being iid. Let $\underline{Z}_t = (Z_t^{(1)}, Z_t^{(2)}) \doteq (X_t, X_{t-k})$. Then

$$\hat{D}_{F_k} = \frac{1}{n^2} \sum_{s,t=1}^{n} h(\underline{Z}_s, \underline{Z}_t) + O_p(n^{-3/2})$$

where $V_n = n^{-2}\Sigma_{s,t=1}^{n} h(\underline{Z}_s, \underline{Z}_t)$ is a von Mises statistic in the technical sense of Denker and Keller (1983) with a degenerate symmetric kernel function. Using asymptotic theory (Carlstein 1988, Denker and Keller 1983 and Skaug 1993a) of this statistic or the related U-statistic we have (Skaug and Tjøstheim 1993b, Th.2) the convergence in distribution

$$n\hat{D}_{F_k} \xrightarrow{\mathscr{L}} \sum_{i,j=1}^{\infty} \eta_i \eta_j W_{ij}^2 \text{ as } n \to \infty, \tag{4.7}$$

where $\{W_{ij}\}$ is an independent identically distributed sequence of $\mathcal{N}(0,1)$ variables, and where the $\{\eta_m\}$ are the eigenvalues of the eigenvalue problem

$$\int g(x, y)\, h(y)\, \mathrm{d}F(y) = \eta h(x)$$

with

$$g(x, y) = \int [1\{x \leqslant w\} - F(w)]\{1(y \leqslant w) - F(w)\} \mathrm{d}F(w).$$

If the distribution of each X_t is continuous, then D_{F_k} is distribution free, i.e., its distribution does not depend on F. Then all calculations can be carried out with F being the uniform [0,1] distribution in which case $g(x, y) = -\max(x, y) + 1/2(x^2 + y^2) + 1/3$ and $\eta_m = (m\pi)^{-2}$, and the distribution in (4.7) can be tabulated by truncating it for a large value of the summation index.

When the distribution of X_t contains a discrete component, the distribution free property is lost, and the complexity increases considerably (Skaug and Tjøstheim 1993b).

A test of the null hypothesis of independence, or rather pairwise independence at lag k, can now be constructed. In the light of our previous results it is reasonable to reject H_0 if large values of $\hat{D}_{F_k}$ is observed. Thus a test of level ε is:

$$\text{reject } H_0 \text{ if } n\hat{D}_{F_k} > u_{n,\varepsilon}$$

where $u_{n,\varepsilon}$ is the upper ε-point in the null distribution of $n\hat{D}_{F_k}$. Since the exact distribution of $\hat{D}_{F_k}$ is unknown, we can use the asymptotic approximation furnished by (4.7). For $n = 50, 100$ and k small this works well. However, as k increases, in general (Skaug and Tjøstheim 1993b) the level is seriously overestimated.

Under the hypothesis of $\{X_t\}$ being iid the bootstrap is a natural tool to use for constructing the null distribution and critical values. For moderate and large k's boot strapping gives a much better approximation to the level and is to be recommended.

Under the alternative hypothesis that X_t and X_{t-k} are dependent, the test statistic $\hat{D}_{F_k}$ will in general be asymptotically normal with a different rate from that in (4.7), but the power function will be very complicated, and we have not tried to obtain an asymptotic expression for it.

To extend the scope to testing of serial independence among $[X_t, \ldots, X_{t-k}]$ or alternatively between several random variables for which there are iid observations for each of them, one might use a functional

$$\hat{D}_{1,k} = \frac{1}{n} \sum_{t=k+1}^{n} \{\hat{F}_{1,k}(X_t, X_{t-1}, \ldots, X_{t-k}) - \hat{F}(X_t)\hat{F}(X_{t-1}) \cdots \hat{F}(X_{t-k})\}^2.$$

The asymptotic theory under the null hypothesis of independence for such a test has been examined by Delgado (1996), but due to the curse of dimensionality, problems can be expected for moderately large k's. As an alternative Skaug and Tjøstheim (1993b) used a "Box-Pierce-Ljung analogy", testing for pairwise independence in all of the pairs $(X_t, X_{t-1}), (X_t, X_{t-2}), \ldots, (X_t, X_{t-k})$ using the statistic

$$\hat{D}_F^{(k)} = \sum_{i=1}^{k} \hat{D}_{F_i}.$$

Of course examples can be found with pairwise independence in all of the pairs, but dependence among $[X_t, X_{t-1}, \ldots, X_{t-k}]$, but I do not think that such examples are very important in practice. The asymptotic theory of such a test is given in Skaug (1993a), and corresponding to (4.7), under the null hypothesis of independence, we obtain

$$n\hat{D}_F^{(k)} \to \sum_{i,j=1}^{\prime} \eta_i \eta_j W_{ij}(k) \text{ as } n \to \infty$$

where the $\{\eta_m, m \geqslant 1\}$ are as in (4.7) and $\{W_{ij}(k)\}$ is an iid sequence of χ^2-variables with k degrees of freedom. Again, better approximations to the nominal level of the test is obtained by bootstrapping.

Since our test is an omnibus test against a non-specified alternative of dependence, it will obviously have lower power for a fully specified alternative than a test designed explicitly for that alternative. To give an idea of the power properties of the test we have considered $k = 1$ and the processes

a) Moving average (MA): $X_t = e_t + \alpha e_{t-1}$
b) Nonlinear moving average (NLMA): $X_t = e_{t-1}(\alpha + e_t)$
c) Bilinear (BL): $X_t = (\alpha + \beta e_{t-1})X_{t-1} + e_t$, $\alpha^2 + \beta^2 < 1$

where $\{e_t\}$ consists of iid $\mathcal{N}(0,1)$ random variables. The bilinear model is from Chan and Tran (1992), whereas the NLMA model is from Skaug and Tjøstheim (1993a). In Figure 1 is shown the simulated power for the tests as a function of α for $n = 100$. We have used 8000 independent realizations for each model. The test is compared to the correlation test based on

$$\hat{\rho}(1) = \frac{1}{n}\sum X_t X_{t-1}$$

and two tests based on estimated densities to be described in the next section. The parameter β is held fixed at $\beta = 0.4$. To get a fair comparison between the tests, critical values found by simulation were used for all of them.

For the MA model it is seen that $\hat{D}_F = \hat{D}_{F_1}$ comes surprisingly close in performance to the correlation functional, which is optimal in this situation. The same was the case for an AR(1) model in Skaug and Tjøstheim (1993b). Significantly better power than for the tests based on densities is obtained. For the BDS test for $\alpha = 0.5$ Mizrach (1991) reports about a power of 0.63 for the MA(1) case, whereas the power of the $\hat{D}_F$-test is approximately 0.96. For the nonlinear MA, $\hat{D}_F$ does not perform well in the middle region of α, but otherwise it outperforms the correlation test. But it is itself clearly

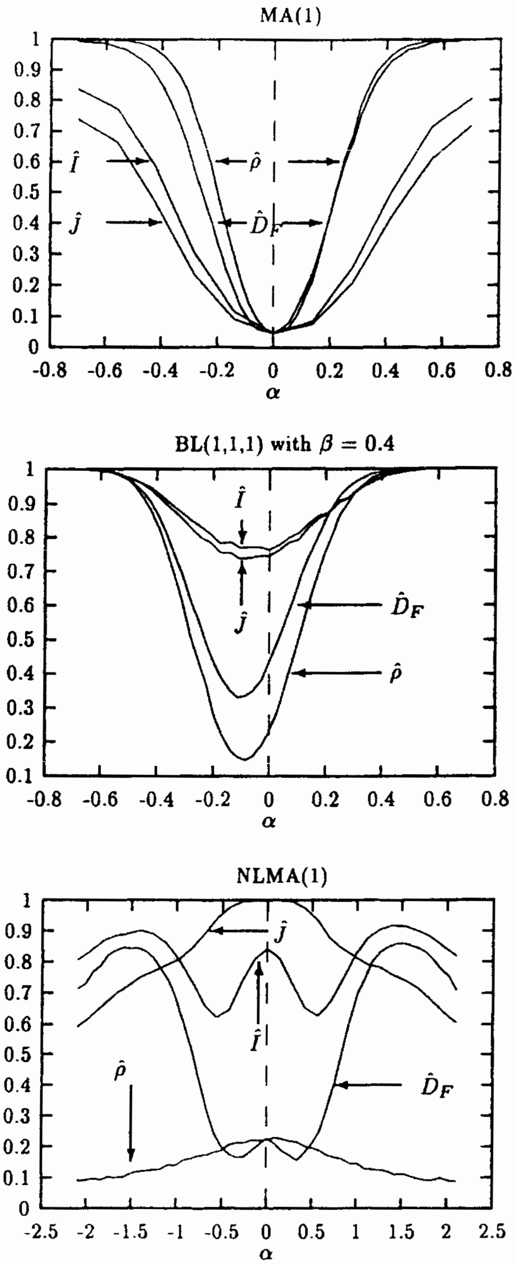

FIGURE 1 Power of tests (significance level is 0.05) as a function of α for the models a), b) and c).

beaten by the two tests based on the density functions. The results for the bilinear model are similar. As explained in the introduction econometricians are especially concerned with ARCH-like deviations from independence, but unfortunately the $\hat{D}_F$ statistic did very poorly for this type of dependence with virtually no power. An ARCH example as well as other examples are given in Skaug and Tjøstheim (1993b).

5. MEASURES AND TESTS BASED ON DENSITY FUNCTIONS

In this section density functions of all variables concerned will be assumed to exist. Most of the material is taken from Skaug and Tjøstheim (1994), to which we refer for more details.

5.1. Measures of Dependence

We use the same principle as in the preceding section. For two random variables X and Y having a joint density $f_{X,Y}$ and marginals f_X and f_Y we measure the degree of dependence by $\Delta(f_{X,Y}, f_X \otimes f_Y)$ where Δ now is a distance measure between two bivariate density functions. The requirements (4.1) and (4.2) discussed in Section 4.1 will be natural to consider here too. Again, the variables are normalized so that the mean and the standard deviation are equal to 0 and 1, respectively.

All of the functionals that I consider will be of the type

$$\Delta = \int B\{f_{X,Y}(x, y), f_X(x), f_Y(y)\} f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y \tag{5.1}$$

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

$$\Delta = \int D\left\{\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\right\} f_{X,Y}(x, y)\,\mathrm{d}x\mathrm{d}y \tag{5.2}$$

which by the change of variable formula for integrals is seen to have the property (4.2). Moreover, If $D(u) \geqslant 0$ and $D(u) = 0$ iff $u = 1$, then (4.1) is fulfilled. Several well known distance measures for density functions are of this type. For instance letting $D(u)=(1-u^{-1/2})$ we obtain the Hellinger distance

$$H=\iint\left\{\sqrt{f_{X,Y}(x,y)}-\sqrt{f_X(x)f_Y(y)}\right\}^2\mathrm{d}x\mathrm{d}y$$

$$=\int 2\left\{1-\sqrt{\frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)}}\right\}f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y \tag{5.3}$$

between $f_{X,Y}$ and $f_X\otimes f_Y$. The Hellinger distance is a metric and thus satisfies (4.1). It is seen to be a special case, for $\gamma=1/2$, of the so-called directed divergence of degree γ ($\gamma\neq 0$, 1) (Chung *et al.* 1987):

$$\Delta_\gamma(f_{X,Y},f_X\otimes f_Y)=\frac{1}{\gamma-1}\left[\int\left\{\frac{f_X(x)f_Y(y)}{f_{X,Y}(x,y)}\right\}^\gamma-1\right]f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y.$$

Clearly (4.2) is fulfilled and for $0<\gamma<1$, using Hölder's inequality

$$(\gamma-1)\,\Delta_\gamma(f_{X,Y},f_X\otimes f_Y)=\int\{f_X(x)f_Y(y)\}^\gamma\{f_{X,Y}(x,y)\}^{1-\gamma}\mathrm{d}x\mathrm{d}y-1\leqslant 0$$

with equality iff $f_{X,Y}=f_X\otimes f_Y$. Hence Δ_γ satisfies (4.1) for $0<\gamma<1$.

The familiar Kullback-Leibler information (entropy) distance is obtainable as a limiting case as $\gamma\to 1$:

$$I=\int\log\left\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right\}f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y. \tag{5.4}$$

Since it is of type (5.2), it satisfies (4.2), and it can be shown to satisfy (4.1) (see Robinson 1991). For $\gamma=2$ the Bickel and Rosenblatt (1973) test of fit distance emerges.

A generalization of the entropy measure is the Renyi entropy measure defined for $\lambda\neq 0,\ -1$ by

$$\mathrm{IR}_\lambda=\frac{2}{\lambda(\lambda+1)}\log\iint\left[\left\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right\}^{1+\lambda}\right.$$

$$\left.-(1+\lambda)\left\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}-1\right\}\right]f_X(x)f_Y(y)\,\mathrm{d}x\mathrm{d}y.$$

The Renyi distance has recently been considered by Parzen (1993) and Zhang and Taniguchi (1994), who obtained higher robustness than with the ordinary Kullback-Leibler statistic when measuring goodness of fit between spectral densities. As far as I know, it has not been tried for the kind of problems considered in this paper.

Other distance functionals, not of the form (5.2), have also been suggested. Thus Chan and Tran (1992) have looked at a distance functional

$$\int |f_{X,Y}(x,y) - f_X(x)f_Y(y)|\,\mathrm{d}x\mathrm{d}y$$

and examined it by simulations in the serial dependence case. A squared distance functional has been considered by Rosenblatt (1975) and Rosenblatt and Wahlen (1992)

$$\Delta_{\mathrm{RW}} = \int \{f_{X,Y} - f_X(x)f_Y(y)\}^2 \mathrm{d}x\mathrm{d}y. \tag{5.5}$$

Both of these satisfy (4.1) but not (4.2), i.e they are not transformation invariant. The latter is not even scale invariant, and it seems essential that some sort of normalization is introduced.

A weighted difference functional was introduced in Skaug and Tjøstheim (1993a). It is related to a series expansion of I, and it is given by

$$J = \int \{f_{X,Y}(x,y) - f_X(x)f_Y(y)\}f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y. \tag{5.6}$$

The distance measure J has neither of the properties (4.1) or (4.2). Thus, there are obvious arguments for dismissing J offhand. The reason this has not been done, is that it has worked consistently well on examples used in the literature for comparing seriel independence tests. Moreover, its simple structure is ideal for exemplifying the asymptotic structure for the estimated functionals.

All of the above measures are conceptually based on the distance between two arbitrary multivariate densities $f_1(\underset{\sim}{x})$ and $f_2(\underset{\sim}{x})$. It is therefore, as was the case for the distribution functions, obvious how one can extend the above formulae to the case of measuring distance between $f_{X_1,\ldots,X_k}$, and $f_{X_1} \otimes \cdots \otimes f_{X_k}$. However, estimating such a measure would involve estimating higher dimensional densities which is difficult due to the curse of dimensionality. Therefore we again look at functionals built up from measuring pairwise dependencies. For a stationary time series and serial dependence we introduce

$$\Delta^{(k)} = \sum_{i=1}^{k} \Delta(f_i, f \otimes f)$$

with $f_i = f_{X_t, X_{t-i}}$ and $f = f_{X_t}$, and for a set of random variables $[X_1, \ldots, X_k]$

$$\Delta^{(k)} = \sum_{i<j} \Delta(f_{X_i, X_j}, f_{X_i} \otimes f_{X_j}).$$

The first construction has been implemented for tests of serial independence in Skaug and Tjøstheim (1994), whereas to my knowledge the second one has not been tried.

5.2. Estimation and Sampling Properties. Tests of Independence

For a given functional $\Delta = \Delta(f, g)$ depending on two densities f and g we estimate Δ by $\hat{\Delta} = \Delta(\hat{f}, \hat{g})$. There are several ways of estimating the densities f and g, but I will only consider nonparametric kernel estimates, i.e.

$$\hat{f}(\underline{x}) = \frac{1}{n} \sum_{t=1}^{n} K_{b_n}(\underline{x} - \underline{X}_t)$$

for given observations $\{\underline{X}_1, \ldots, \underline{X}_n\}$. Here $K_{b_n}(\underline{x} - \underline{X}_t) = b_n^{-d} K\{b_n^{-1}(\underline{x} - \underline{X}_t)\}$, where b_n is the bandwidth, K is the kernel function, and d is the dimension of $\underline{X}$. The kernel function was taken to be a product of one-dimensional kernels in Skaug and Tjøstheim (1994), i.e. $K(\underline{x}) = \Pi K_i(x_i)$, where each K_i generally is non-negative and satisfies

$$\int K_i(x)\, \mathrm{d}x = 1, \quad \int x^2 K_i(x) \mathrm{d}x < \infty.$$

The product form is convenient but not necessary. In Skaug and Tjøstheim (1993a and 1994) we have used a Gaussian kernel, i.e. $K(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$, but there are many other possibilities, e.g. $K(x) = 15/16\ (1 - x^2)^2 1(|x| \leqslant 1)$ and similar kernels having a compact support.

In optimality theory of density estimation (see e.g. Silverman 1986) b_n will depend on both the sample size n, the dimension d and the covariance matrix or a similar scatter measure for $\underset{\sim}{X}$. Such an optimality theory has not been extended to the type of functionals we are considering, and in our experiments involving pairs of random variables (X_t, X_{t-k}) we have taken a rather simple minded approach: We have scaled all variables involved so that they have $\mathrm{E}(X_t)=0$ and $\mathrm{var}(X_t)=1$ (or rather the corresponding empirical quantities satisfy these identities), and we have chosen $b_n = n^{-1/6}$ which is roughly optimal asymptotically in the sense of Silverman (1985, p. 86) in case of the bivariate estimate $\hat{f}_{X,Y}(x, y)$ but not for the univariate density estimate $\hat{f}_X(x)$. Using the same bandwidth b_n for the bivariate and univariate density estimates simplifies the asymptotic analysis (cf. Skaug and Tjøstheim 1994) and is consistent with Rosenblatt's approach (1975). Robinson (1991) uses different bandwidths and introduces a kernel function which is allowed to take on negative values to get the asymptotic theory to go through in his case.

Once estimates for $f_{X,Y}$, f_X and f_Y have been obtained in the integral expression (5.1) for Δ, the integral could have been computed by numerical integration, but we have opted for taking empirical averages. Then in effect we replace $f(x, y)\,\mathrm{d}x\mathrm{d}y = \mathrm{d}F(x, y)$ by $\mathrm{d}\hat{F}(x, y)$ so that corresponding to (5.1)

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} B\{\hat{f}_{X,Y}(X_i, Y_i), \hat{f}_X(X_i), \hat{f}_Y(Y_i)\}$$

for given pairs of observations $\{(X_1, Y_1),\dots,(X_n, Y_n)\}$, and

$$\hat{\Delta}_k = \frac{1}{n-k}\sum_{t=k+1}^{n} B\{\hat{f}_k(X_t, X_{t-k}), \hat{f}(X_t), \hat{f}(X_{t-k})\}$$

in case we are considering serial dependence for a stationary process with observations $\{X_1, \dots, X_n\}$. From now on we will concentrate on the latter, the former being similar and simpler in case of iid series $\{X_t\}$ and $\{Y_t\}$, and more difficult when one has serial dependence in the individual series $\{X_t\}$ and $\{Y_t\}$. We have found it convenient to introduce a weight function $w(X_t, X_{t-k})$ such that

$$\hat{\Delta}_k = \hat{\Delta}_{k,w} = \frac{1}{n-k}\sum_{t=k-1}^{n} B\{\hat{f}_k(X_t, X_{t-k}), \hat{f}(X_t), \hat{f}(X_{t-k})\}w(X_t, X_{t-k}).$$

Typically $w(x, y) = 1\{|x| \leqslant \lambda\,\mathrm{sd}(X)\}1\{|y| \leqslant \lambda\,\mathrm{sd}(Y)\}$ where λ usually is between 2 and 3. The purpose of the weighting is twofold: We want to screen off outliers, and the asymptotic theory simplifies with this device. In effect it means that we are only measuring the dependence within the support of w. We could let the support of w tend to infinity as $n \to \infty$, but such an n-dependence would lead to complications in the asymptotic analysis (cf. Robinson 1991 for a related case). Moreover, it is not clear how the n-dependence should be chosen in practice for a given finite sample.

When a weight function w is included, then the theoretical counterpart is

$$\Delta_{k,w} = \int B\{f_k(x, y), f(x), f(y)\} w(x, y) f_k(x, y) \, dx dy$$

which is what the estimate $\hat{\Delta}_k$ should be measured against. In Skaug and Tjøstheim (1994) an asymptotic theory is derived for the weighted functionals $\hat{I}_{k,w}$, $\hat{J}_{k,w}$ and $\hat{H}_{k,w}$ corresponding to the functionals (5.3), (5.4) and (5.6), the weighting being as above. Then (4.1) and (4.2) are only approximately fulfilled. Similar derivations can be done for more general functionals, and it is indicated how an asymptotic theory can be developed in these more general cases in the Appendix.

Under a precise set of regularity conditions stated in Skaug and Tjøstheim (1994) consistency and asymptotic normality can be obtained for the test functionals. It should be noted that the leading term in an asymptotic expansion of the standard deviation is of order $n^{-1/2}$ for all of the three functionals $H^{(k)}$, $I^{(k)}$ and $J^{(k)}$. This is the same as the standard deviation of parametric estimates in a regular parametric estimation problem, whereas the standard deviation for the estimators $\hat{f}(x)$ and $\hat{f}_k(x, y)$ are of order $(nb_n)^{-1/2}$ or $(nb_n^2)^{-1/2}$, respectively. In the regular parametric case the next term of the Edgeworth expansion is of order n^{-1}, and for moderately large values of n the first order term of order $n^{-1/2}$ will dominate. Similarly (Hall, 1992 Ch. 4.4)) for $\hat{f}(x)$ and $\hat{f}_k(x, y)$ the next terms are typically of order $(nb_n)^{-1}$ and $(nb_n^2)^{-1}$. However, for the functionals constructed above the next terms are generally much closer in order. For $\hat{J}$ it is shown in Skaug and Tjøstheim (1993a) that the next terms are of order $n^{-1/2}b_n$ and $(nb_n)^{-1}$, and since $b_n = O(n^{-1/6})$ or $O(n^{-1/5})$, n must be very large in order for the first order term to dominate in the asymptotic expansion. Hence first order asymptotic theory based on the normal approximation can be expected to be inaccurate except when n is very large. Basing a test of independence directly on such a theory may be hazardous as the real level will typically deviate substantially from the nominal level. This is amply demonstrated in Skaug and Tjøstheim (1993a and 1994) where sometimes the level was twice that of the nominal level for a sample size of 100 observations. Similar problems has been reported for the other functionals, and in linearity testing with nonparametric functionals (Hjellvik and Tjøstheim 1995).

An obvious remedy is to include higher order terms in the asymptotic expansion, but this is problematic, as they are difficult to compute analytically and would involve complicated expressions which have to be estimated. This suggests 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 (cf. Hall 1992, Ch. 3 and Ch. 4) although no rigorous analysis to confirm this has been carried through for the present functionals. Simulations in Skaug and Tjøstheim (1993a, 1994) do indicate that much better appromixation to the level is obtained. It is my belief that the potential for bootstrap methods is larger here than in the purely parametric or nonparametric case, since in those cases the first order asymptotics work quite well for modest sample sizes.

As in Section 4 it is quite difficult to do asymptotic power studies. Such studies would also be unreliable unless n is rather large for the reasons just mentioned. For a linear AR(1) alternative analytic expressions are obtainable, and it can be shown that the Pitman efficiency of $\hat{J}$ has a lower rate than the correlation function ($n^{-1/4}$ against $n^{-1/2}$). This type of result can also be expected for the other functionals and for general classes of alternatives, as it is closely tied to the nonparametric approach. For a finite sample size ($n = 100$), the test functionals $\hat{H}$, $\hat{I}$ and $\hat{J}$ are inferior to the correlation functional and to the functional derived from the empirical distribution function for the MA model of Figure 1 and for an AR(1) process considered in Skaug and Tjøstheim (1994). However, for a number of other examples of nonlinear character, including the ARCH model, the functionals $\hat{H}$, $\hat{I}$ and $\hat{J}$ are far superior to the correlation functional, and comparable, and in some cases superior, to the BDS functional. This is illustrated in a few special cases on Figure 1. A much more thorough documentation and a real data example involving exchange rates are given in Skaug and Tjøstheim (1994).

6. DISCUSSION AND WORK IN PROGRESS

In a test of fit situation one is interested in subjecting the residuals to an independence test. One should then compensate for the fact that the residuals only are estimates of the true residuals, which in turn are independent if the model is correct. In linear model checking one adjusts for this in the Box-Pierce-Ljung statistic via the reduction of the degrees of freedom of the χ^2-test according to the number of estimated parameters.

In Brock *et al.* (1991) it is discussed how the fitting aspect can be handled asymptotically for the BDS test. But as $n \to \infty$ under the assumption that the model is correct, the estimated residuals will converge towards the theoretical residuals, and the peculiar features of the fitting process tend to disappear.

I believe that a bootstrap approach could give better results in a finite sample situation, but the bootstrap procedure should be adjusted for the fitting process with the probable loss of some power.

Another aspect I have not touched much on, is the possibility of constructing tests along the lines suggested in Sections 4 and 5, when one wants to test independence between processes $\{X_t\}$ and $\{Y_t\}$. This can be done pairwise by imposing the assumption of joint stationarity on $\{X_t, Y_t\}$ and by looking at pairs of random variables $\{X_t, Y_{t-k}\}$. An asymptotic theory can be built in analogy to the one discussed in Sections 4 and 5. The bootstrap can be used in case $\{X_t\}$ and $\{Y_t\}$ are each iid. If there is serial dependence, the ordinary bootstrap will fail, but possibly a block bootstrap (Künsch 1989, Bühlmann 1993) may work. This is under investigation. So far the results are preliminary, but the block bootstrap seems to be rather more unstable than the ordinary bootstrap.

Another set of problems are those connected with conditional independence and causality tests. A difficulty is that in more complicated dependence relationships, one need to estimate conditional probabilities $Pr\ (Y_1,\dots,Y_k \mid X_1,\dots X_\ell)$ and unless the sample size is very large this estimation problem could be rendered impossible by the curse of dimensionality if $k + \ell$ is larger than 3–4. A semi-parametric approach to the problem may be better in this case, but it is not clear how one should proceed most effectively.

Acknowledgement

I am grateful to Hans J. Skaug for providing the plots of Figure 1.

Appendix: A Framework for the Asymptotic Analysis of a Class of Nonparametric Functionals

The asymptotic analysis for the functionals of Section 5 is of limited practical applicability. This appendix is therefore perhaps mostly of theoretical interest. In practice we do not often expect to have at our disposal the very large sample sizes required for the first order asymptotic expansions to be accurate. Nevertheless, I will outline a framework for an asymptotic analysis of somewhat larger generality than that discussed in Section 5.

I write in the language of a univariate stationary series $\{X_t\}$, but it will be clear how it can be generalized to other situations. I assume that the quantity we want to estimate is of the form

$$\Delta = \int B\{\underline{u}(\underline{x}),\ \underline{v}(\underline{x},\underline{\theta})\}\, w(\underline{x})\mathrm{d}F(\underline{x}).$$

Here B and $\underline{v}$ are known functions, but $\underline{v}$ depends on an unknown vector parameter $\underline{\theta}\in\Theta$. The vector function $\underline{u}$ is unknown and is estimated non-parametrically. In Section 5, $\underline{u}(\underline{x}) = \{u_1(x_1,x_2),\ u_2(x_1,x_2),\ u_3(x_1,x_2)\} = \{f_{X_1X_2}(x_1,x_2), f_{X_1}(x_1),\ f_{X_2}(x_2)\}$ and $\underline{v}$ is absent. Finally, $w(\underline{x})$ is a weight function having support on a compact set S and $F(\underline{x}) = F(x_1,\ldots,x_k)$ is the cumulative distribution function of $\{X_t,\ X_{t-i_1},\ldots,\ X_{t-i_{k-1}}\}$. In this manner we could discuss a test of fit problem examining if the density $f(\underline{x})$ belongs to a certain parametric class (Skaug 1993b), or test the difference between the non-parametric nonlinear predictor and a linear parametric predictor (Hjellvik and Tjøstheim 1995)(see also Härdle and Mammen 1993), or perform a test of independence. The function B can be thought of as measuring the distance between the functions $\underline{u}$ and $\underline{v}$ or, if $\underline{v}$ is missing, between the components of $\underline{u}$. In Section 5 $B\{\underline{u}(x,y)\} = \ln\ \{u_1(x,y)/u_2(x)u_3(y)\} = \ln\{f_{X,Y}(x,y)/f_X(x)f_Y(y)\}$ in case the information measure is used. In a univariate test of fit problem $\underline{u}(\underline{x}) = f(x)$, where $f(x)$ is the true unknown density and $\underline{v}(\underline{x},\underline{\theta}) = g(x,\ \underline{\theta})$ where g is a known candidate family of densities depending on a parameter $\underline{\theta}$.

We can estimate Δ by

$$\hat{\Delta} = \int B\{\underline{\hat{u}}(\underline{x}),\ \underline{v}(\underline{x},\underline{\hat{\theta}})\}w(\underline{x})\,\mathrm{d}\hat{F}(\underline{x}) \qquad (A.1)$$

where $\hat{F}$ is the empirical distribution function of $\{X_t, X_{t-i_1},\ldots,X_{t-i_{k-1}}\}$, $\underline{\hat{u}}(\underline{x})$ is a non-parametric estimate of $\underline{u}(\underline{x})$, and $\underline{\hat{\theta}}$ is an (e.g. maximum likelihood) estimate of $\underline{\theta}$.

One can obtain consistency by essentially requiring:

A0) $\{X_t\}$ is ergodic.

A1) i) $\sup_{x\in S}\|\underline{\hat{u}}(\underline{x}) - \underline{u}(\underline{x})\| \xrightarrow{a.s.} 0$

ii) $\sup_{x\in S}\|\underline{v}(\underline{x},\underline{\hat{\theta}}) - v(\underline{x},\ \underline{\theta})\| \leqslant M|\underline{\hat{\theta}} - \underline{\theta}|$ and $|\underline{\hat{\theta}} - \underline{\theta}| \xrightarrow{a.s.} 0$.

A2) There exists an open set $A \supset \mathrm{Range}_{x\in S}\{\underline{u}(\underline{x}),\ \underline{v}(\underline{x},\underline{\theta})\}$ such that for $i = 1,\ldots,\dim(\underline{u})$, $j = 1,\ldots,\dim(\underline{\theta})$ and each $\underline{\theta}\in\Theta$

$$\sup_{u,v\in A}\left|\frac{\partial B}{\partial u_i}(\underline{u},\underline{v})\right| \leqslant C_1 \quad\text{and}\quad \sup_{u,v\in A}\left|\frac{\partial B}{\partial \theta_j}(\underline{u},\underline{v})\right| \leqslant C_2$$

where C_1 and C_2 are constants. There are a number of sufficient conditions guaranteeing the uniform almost sure convergence of nonparametric estimates on a compact set needed under A1 i). To satisfy A1 ii) it will in general suffice to require that $\underline{v}$ is differentiable with respect to $\underline{\theta}$ such that $\sup_{\underline{x}\in S} \partial \underline{v}(\underline{x}, \underline{\theta})/\partial\theta_i \leqslant M$ and to establish $\hat{\underline{\theta}} \to \underline{\theta}$ a.s., and many conditions are known to imply the latter. Usually more than A0 is required to satisfy A1, so that in practice this condition must be strengthened to yield some form of mixing (strong mixing or absolute regularity). Again, since S is assumed to be a compact set, A2 will in general be a mild condition when specified jointly with A1.

Under A0–A2 strong consistency can be obtained as follows:

$$\hat{\Delta} - \Delta = \int B\{\underline{u}(\underline{x}), \underline{v}(\underline{x},\underline{\theta})\}w(\underline{x})\{\mathrm{d}\hat{F}(\underline{x}) - \mathrm{d}F(\underline{x})\}$$

$$+ \int [B\{\hat{\underline{u}}(\underline{x}), v(\underline{x},\hat{\underline{\theta}})\} - B\{\underline{u}(\underline{x}), v(\underline{x},\underline{\theta}\}]w(\underline{x})\mathrm{d}\hat{F}(\underline{x}) = I + II.$$

Here $I \xrightarrow{a.s.} 0$ due to the ergodic theorem. By A1 there exists for each realization an integer N such that

$$K_n \doteq 1[\{\hat{\underline{u}}(\underline{x}), \underline{v}(\underline{x}, \hat{\underline{\theta}})\}\in A] = 1$$

for $n \geqslant N$. Therefore, showing $II \xrightarrow{a.s.} 0$ is equivalent to showing $K_n II \xrightarrow{a.s.} 0$. By the mean value theorem there exists a random function $\underline{u}'(\underline{x})$ and a random vector parameter $\underline{\theta}' \in \Theta$ such that

$$K_n|B\{\hat{\underline{u}}(\underline{x}), v(\underline{x}, \hat{\underline{\theta}})\} - B\{\underline{u}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}\}|$$

$$\leqslant \sum_{i=1}^{\dim(u)} K_n\left|\frac{\partial B\{\underline{u}'(\underline{x}), v(\underline{x},\underline{\theta}')\}}{\partial u_i}\right||\hat{u}_i(\underline{x}) - u_i(\underline{x})|$$

$$+ \sum_{i=1}^{\dim(\theta)} K_n\left|\frac{\partial B\{\underline{u}'(\underline{x}), v(\underline{x}, \underline{\theta}')\}}{\partial\theta_i}\right||\hat{\theta}_i - \theta_i|$$

and the result follows from A1 and A2.

It is rather more difficult to establish the asymptotic distribution. If $\partial B/\partial u_i\{\underline{u}(\underline{x}), \underline{v}(\underline{x},\underline{\theta})\}$ and $\partial B/\partial\theta_j\{\underline{u}(\underline{x}), \underline{v}(\underline{x},\underline{\theta})\} \neq 0$, for at least one i and one j, then asymptotic normality can be obtained under a set of regularity conditions to be discussed. If this is not true, then higher order terms contribute, and typically U-statistic theory must be employed. In special cases this situation has been considered in papers by Hall (1984) and Rosenblatt (1975). If there is n-dependence in the kernel, one can still get asymptotic normality. Otherwise one usually obtains a χ^2-type distribution or so-called Kiefer processes. (see e.g. Denker and Keller 1983). Here I will only treat the case where the first order terms are non-zero covering most of the cases mentioned in Section 5, and I will indicate the type of regularity conditions which can produce asymptotic normality then.

The following lemma is useful in such an analysis. I use the notation $\hat{U}(\underline{x}) = n^{1/2}\{\hat{F}(\underline{x}) - F(\underline{x})\}$.

LEMMA A1 (Skaug and Tjøstheim 1994)
Let $\{D_n(\underline{x}, \underline{\eta}),\ n \geqslant 1\}$ be a sequence of possibly complex-valued, uniformly bounded functions. i.e. there exists a constant C such that

$$\sup_n \sup_{x,\eta} |D_n(\underline{x}, \underline{\eta})| \leqslant C.$$

Further, let $\{X_t\}$ be a strongly mixing process with mixing coefficient $\alpha(j)$. If $\Sigma_j \alpha(j) < \infty$, then

$$\sup_\eta \mathrm{E}\left|\int D_n(\underline{x}, \underline{\eta})\,\mathrm{d}\hat{U}(\underline{x})\right|^2 = O(1)$$

and if $\Sigma_j j\alpha(j) < \infty$, then

$$\sup_\eta \mathrm{E}\left|\int D_n(\underline{x}, \underline{\eta})\,\mathrm{d}\hat{U}(\underline{x})\right|^4 = O(1).$$

This is only proved for $k = 2$ in Skaug and Tjøstheim (1994), but it can be extended.

We now make an assumption on rates

A3) i) $\sup_{x \in S} |\underline{\hat{u}}(\underline{x}) - \underline{u}_{b_n}(\underline{x})| \overset{a.s.}{=} o(n^{-1/4})$
ii) $|\underline{\hat{\theta}} - \underline{\theta}_n| \overset{a.s.}{=} o(n^{-1/4})$

where $\underline{u}_{b_n}(x) = \mathrm{E}\{\underline{\hat{u}}(\underline{x})\}$ and $\underline{\theta}_n = \mathrm{E}(\underline{\hat{\theta}})$. These assumptions are not very restrictive, and one can quite easily find conditions under which they are satisfied. For example for nonparametric kernel estimates, there are a number of sufficient conditions resulting in "optimal" almost sure rates

$$\sup_{\underline{x}\in S}|\hat{\underline{u}}(\underline{x}) - \underline{u}_{b_n}(\underline{x})| \overset{a.s.}{=} O(n^{-1/2}(\ln n)^{1/2} b_n^{-k/2})$$

and $(\ln n)^{1/2} b_n^{-k/2} = o(n^{1/4})$ would then suffice. For example the familiar rate $b_n = O\{n^{-1/(k+4)}\}$ in the standard situation (Silverman 1986, p. 86) would be enough for $k \leqslant 3$.
Concerning A3 ii), under wide restrictions $|\hat{\underline{\theta}} - \underline{\theta}_n| \overset{a.s.}{=} O\{n^{-1/2}(\ln n)^{1/2}\}$.
We also impose the quite mild condition

A4) : for each θ there exists an open set $A \supset \text{Range}_{\underline{x}\in S}\{\underline{u}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta})\}$ such that

$$\sup_{\underline{u},\underline{v}\in A} \frac{\partial^2 B}{\partial u_i \partial u_j} \frac{\partial^2 B}{\partial u_i \partial \theta_j}, \frac{\partial^2 B}{\partial \theta_i \partial \theta_j} \leqslant C.$$

Under A0–A4 and Lemma A1 we can obtain the following linearization:

$$\begin{aligned}\hat{\Delta} = &\int B\{\underline{u}_{b_n}(\underline{x}),\ \underline{v}(\underline{x}, \underline{\theta}_n)\} w(\underline{x})\,\mathrm{d}F(\underline{x}) \\ &+ \int B\{\underline{u}_{b_n}(\underline{x}),\ \underline{v}(\underline{x}, \underline{\theta}_n)\} w(\underline{x})\,\{\mathrm{d}\hat{F}_n(\underline{x}) - \mathrm{d}F(\underline{x})\} \\ &+ \sum_i \int \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}(\underline{x}),\ \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{u}_i - u_{i,b_n})\, w(\underline{x})\,\mathrm{d}F(\underline{x}) \\ &+ \sum_i \int \frac{\partial B}{\partial \theta_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n}) w(\underline{x})\mathrm{d}F(\underline{x}) + o_p(n^{-1/2})\end{aligned}$$

I will indicate a very rough proof of this result: Assumption A4 means that we can Taylor expand up to second order: We write (omitting the $\underline{x}$-dependence)

$$\begin{aligned}B\{\hat{\underline{u}}, \underline{v}(\hat{\underline{\theta}})\} = B\{\underline{u}_{b_n}, \underline{v}(\underline{\theta}_n)\} &+ \sum_i \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}, \underline{v}(\underline{\theta}_n)\}(\hat{u}_i - u_{i,b_n}) \\ &+ \sum_i \frac{\partial B}{\partial \theta_i}\{\underline{u}_{b_n}, \underline{v}(\underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n})\end{aligned}$$

$$+\frac{1}{2}\sum_i \sum_j \frac{\partial^2 B}{\partial u_i \partial u_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{u}_i - u_{i,b_n})(\hat{u}_j - u_{j,b_n})$$

$$+\frac{1}{2}\sum_i \sum_j \frac{\partial^2 B}{\partial u_i \partial \theta_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{u}_i - u_{i,b_n})(\hat{\theta}_{j,n} - \theta_{j,n})$$

$$+\frac{1}{2}\sum_i \sum_j \frac{\partial^2 B}{\partial \theta_i \partial \theta_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{\theta}_i - \theta_{i,n})(\hat{\theta}_j - \theta_{j,n})$$

where $\underline{u}', \underline{\theta}'$ are stochastic quantities depending on n and determined by the mean value theorem. Subtracting and adding terms in the expression (A.1) for $\hat{\Delta}$, using the Taylor expansion and the notation $\hat{U}(\underline{x}) = n^{1/2}\{\hat{F}(\underline{x}) - F(\underline{x})\}$ we obtain

$$\begin{aligned}
\hat{\Delta} = {} & \int B\{\underline{u}_{b_n}(\underline{x}),\ v(\underline{x}, \underline{\theta}_n)\}\, w(\underline{x})\, \mathrm{d}F(\underline{x}) + \int B\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\} w(\underline{x})\{\mathrm{d}\hat{F}(\underline{x}) - \mathrm{d}F(\underline{x})\} \\
& + \int\Bigg[\sum_i \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{u}_i - u_{i,b_n}) \\
& + \sum_i \frac{\partial \mathrm{B}}{\partial \theta_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n})\Bigg] w(\underline{x})\, \mathrm{d}F(\underline{x}) \\
& + n^{-1/2}\int\Bigg[\sum_i \frac{\partial B}{\partial u_i}\{(\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{u}_i - u_{i,b_n}) \\
& + \sum_i \frac{\partial B}{\partial \theta_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n})\Bigg] w(\underline{x})\, \mathrm{d}\hat{U}(\underline{x}) \\
& + \int \frac{1}{2}\sum \sum \frac{\partial^2 B}{\partial u_i \partial u_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{u}_i - u_{i,b_n})(\hat{u}_j - u_{j,b_n}) w(\underline{x})\, \mathrm{d}\hat{F}(\underline{x}) \\
& + \int \frac{1}{2}\sum \sum \frac{\partial^2 B}{\partial u_i \partial \theta_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{u}_i - u_{i,b_n})(\hat{\theta}_j - \theta_{j,n}) w(\underline{x})\, \mathrm{d}\hat{F}(\underline{x}) \\
& + \int \frac{1}{2}\sum \sum \frac{\partial^2 B}{\partial \theta_i \partial \theta_j}\{\underline{u}', \underline{v}(\underline{\theta}')\}(\hat{\theta}_i - \theta_{i,n})(\hat{\theta}_j - \theta_{j,n}) w(\underline{x})\, \mathrm{d}\hat{F}(\underline{x}).
\end{aligned} \tag{A.2}$$

Using A3 and A4 we have that the term involving second order derivatives of B is of order $o(n^{1/2})$ almost surely, and therefore the same is true in probability. Lemma A1 and Condition A3 can be used to show that the terms involving the first order derivative of B in the fourth and fifth line of (A.2) are of order $o_p(n^{-3/4})$, and we then obtain the desired linearization.

The linearization can be used to prove asymptotic normality under the null situation of $B\{\underline{u}(\underline{x}), \underline{v}(\underline{x}, \theta)\} = 0$ if $B\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\} = 0$. In the test of independence this is the case ($\underline{v}$ missing). Otherwise, under rather weak conditions $B\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\} \to B\{\underline{u}(\underline{x}), \underline{v}(\underline{x}, \theta)\} = 0$ uniformly in $\underline{x}$, and Lemma A1 will then yield that the term

$$\int B\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\} w(\underline{x})\{\mathrm{d}\hat{F}(\underline{x})\} - \mathrm{d}F(\underline{x})\} = o_p(n^{-1/2}),$$

and the asymptotic distribution is then determined by

$$\int\left[\sum_i \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{u}_i - u_{i,b_n})\right.$$

$$\left. + \sum_i \frac{\partial B}{\partial \theta_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n})\right] w(\underline{x})\mathrm{d}F(\underline{x}) \tag{A.3}$$

which is generally of order $O_p(n^{-1/2})$. In the kernel estimation case for example, with $k = \dim(\underline{x})$, eventhough $\hat{u}_i - u_{i,b_n} = O_p\{(n^{-1}b_n^{-k})^{1/2}\}$, we obtain using Lemma A1, Bochner's Lemma and representations of type

$$\hat{\underline{u}}(\underline{x}) - \underline{u}_{b_n}(\underline{x}) = n^{-1/2}\int g(\underline{x}'')K_{b_n}(\underline{x}' - \underline{x})\,\mathrm{d}\hat{U}([\underline{x}', \underline{x}''])$$

(cf. Skaug and Tjøstheim 1994 for an example) a term of order $O_p(n^{-1/2})$. Here g is a function determined by the representation. Moreover, a mixing argument can be used to show asymptotic normality of

$$n^{-1/2}\int \sum \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\} g(\underline{x}'')K_{b_n}(\underline{x}' - \underline{x}) w(\underline{x})\,\mathrm{d}F(\underline{x})\mathrm{d}\hat{U}([\underline{x}', \underline{x}'']).$$

Similarly, asymptotic normality of the term

$$\int \sum \frac{\partial B}{\partial u_i}\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}(\hat{\theta}_i - \theta_{i,n}) w(\underline{x})\mathrm{d}F(\underline{x})$$

will in general follow from asymptotic normality of $\hat{\underline{\theta}}$, and the arguments can be extended to obtain joint asymptotic normality of the two terms in (A.3).

Finally, in the non-null situation the term

$$\int B\{\underline{u}_{b_n}(\underline{x}), \underline{v}(\underline{x}, \underline{\theta}_n)\}\, w(\underline{x})\{\mathrm{d}\hat{F}(\underline{x}) - \mathrm{d}F(\underline{x})\}$$

will (cf. Lemma A1) be of order $O_p(n^{-1/2})$ and will contribute to the asymptotic distribution, which, again using a mixing theorem, can be shown to be asymptotically normal under quite weak conditions. The outline given in this appendix is exemplified on special models with rigorous proofs and precise conditions in Skaug and Tjøstheim (1994), and related work is presented in Masry and Tjøstheim (1995 and 1997).

References

An, Hong-Zhi and Cheng, Bing (1990). A Kolmogorov-Smirnov type statistic with appplication to test for normality in time series. *Inter. Stat. Rev.*, **59**, 287–307.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* Wiley: New York.

Bartels, R. (1982). The rank version of von Neumann's ratio test for randomness. *J. Amer. Stat. Assoc.*, **77**, 40–46.

Benghabrit, Y. and Hallin, M. (1992). Optimal rank-based test against first-order superdiagonal bilinear dependence, *J. Stat. Plan. and Inf.*, **32**, 45–61.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Stat.*, **1**, 1071–1095.

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.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. of Economics*, **31**, 307–327.

Box, G. E. P. and Pierce, D. (1970). Distribution of residual autocorrelation in autoregressive integrated moving average time series models. *J. Amer. Stat. Assoc.*, **65**, 1509–1526.

Bradley, R. (1986). *Basic properties of strong mixing conditions.* In Dependence in Probability and Statistics. A Survey of Recent Results. Eberlein, E. and Taqqu, M. S. Eds., Birkhäuser: Boston.

Brock, W. A., Dechert, W. D. and Scheinkman, J. A. (1987). A test for independence based on the correlation dimension. Department of Economics, University of Wisconsin.

Brock, W. A., Dechert, W. D., Scheinkman, J. A. and LeBaron, B. (1991). A test of independence based on the correlation dimension. Department of Economics, University of Wisconsin.

Brockwell, P. and Davis, R. (1987). *Time Series: Theory and Methods.* Springer: New York.

Bühlman, P. L. (1993). The blockwise bootstrap in time series and empirical processes. Doctoral thesis, Swiss Federal Institute of Technology, Zürich (ETH).

Carlstein, E. (1988). Degenerate U-statistics based on non-independent observations. *Calcutta Stat. Assoc. Bulletin*, **37**, 55–65.

Chan, N. H. and Tran, L. T. (1992). Nonparametric tests for serial dependence. *J. of Time Ser. Anal.*, **13**, 19–28.

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric tests. *Ann. Math. Stat.*, **29**, 972–994.

Chung, J. K., Kannapau, P. L., Ng, C. T. and Sohoo, P. K. (1989). Measures of distance between probability distributions, *J. Math. Anal. and Appl.*, **138**, 280–292.

Delgado, M. (1996). Testing serial independence using the sample distribution function. *J. of Time Ser. Anal.*, **17**, 271–286.

Denker, M. and Keller, G. (1983). On U-statistics and von Mises statistics for weakly dependent processes. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, **64**, 505–522.

Dufour, J. M. and Hallin, M. (1991). An exponential bound for the permutational distribution of a first order autocorrelation coefficient. Statistique et Analyse des Données, **15**, 45–56.

Dufour, J. M. and Roy, R. (1985). Some robust exact results on sample auto correlations and tests of randomness. *J. of Econometrics*, **29**, 257–273, **41**, 279–281.

Dufour, J. M. and Roy, R. (1986). Generalized portmanteau statistics and tests of randomness, Communications in Statistics. A. Theory and Methods, **15**, 2953–2972.

Efron, B. (1982). The Jackknife, the Bootstrap and other Resampling Plans. SIAM Regional Conference Series in Applied Mathematics. Arrowsmith: Bristol.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of the U.K. inflation. *Econometrica*, **50**, 987–1008.

Fama, E. (1965). The behavior of stock market prices. *J. of Business*, **38**, 34–105.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *J. of Finance*, **25**, 383–417.

Grossberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica*, **9D**, 189.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate non-parametric density estimators, *J. of Multivariate Analysis*, **14**, 1–16.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer: New York.

Hallin, M. (1994). On the Pitman nonadmissibility of correlation based testing methods in time series analysis. *J. of Time Ser. Anal.*, **15**, 607–612.

Hallin, M. and Mélard, G. (1988). Rank-based test for randomness against first order serial dependence. *J. Amer. Stat. Assoc.*, **83**, 1117–1129.

Hallin, M. and Puri, M. L. (1993). Rank tests for time series analysis. A survey. preprint. To appear in the IMA Volumes in Mathematics and its Applications, Proceedings on Time Series.

Hallin, M., Ingenbleek, J.-Fr. and Puri, M. L. (1985). Linear serial rank tests for randomness against ARMA alternatives. *Ann. Stat.*, **13**, 1156–1181.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 1926–1948.

Hinich, M. (1982). Testing for Gaussianity and linearity of a stationary time series. *J. of Time Ser. Anal.*, **3**, 169–176.

Hinich, M. (1993). Testing for dependence in the input to a linear time series model. Preprint. Department of Government, University of Texas, Austin.

Hjellvik, V. and Tjøstheim, D. (1995). Nonparametric test of linearity for time series. *Biometrika*, **82**, 351–368.

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

Hsieh, D. and LeBaron, B. (1988). Small sample properties of the BDS statistics I and II. Graduate School of Business, University of Chicago and University of Wisconsin.

Kendall, M. G. (1970). *Rank Correlation Methods*, 4th edition. Griffin: London.

Knoke, D. J. (1977). Testing for randomness against autocorrelation: alternative tests. *Biometrika*, **64**, 523–529.

Künsch, H. R. (1989). The Jacknife and the bootstrap for general stationary observations. *Ann. Stat.*, **17**, 1217–1241.

Lawrance, A. J. (1991). Directionality and reversibility in time series. *Intern. Stat. Review*, **59**, 67–79.

Lawrance, A. J. and Lewis, P. A. W. (1987). Higher-order residual analysis for nonlinear time series with autoregressive correlation structures. *Intern. Stat. Review*, **55**, 21–35.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. Strong convergence and asymptotic normality. *Econometric Theory*, **11**, 258–289.

Masry, E. and Tjøstheim, D. (1997). Additive nonlinear ARX time series and projection estimates. Econometric Theory, to appear.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residuals and autocorrelations. *J. of Time Ser. Anal.*, **4**, 269–273.

Mizrach, B. (1991). A simple nonparametric test of independence of order (p). Department of Finance. The Wharton School, Philadelphia.

Pagan, A. and Hall, A. (1983). Diagnostic tests of residual analysis. *Econometric Reviews*, **2**, 159–218.

Parzen, E. (1993). Stationary time series analysis using information and spectral analysis. In Developments in Time Series Analysis. The Priestley Birthday Volume. T. Subba Rao editor, 139–148.

Pinkse, C. A. P. (1993). A general characteristic function based measure applied to serial independence testing. Preprint. *London School of Economics*.

Robinson, P. (1991). Consistent nonparametric entropy-based testing. *Review of Economic Studies*, **58**, 437–453.

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

Rosenblatt, M. and Wahlen, B. (1992). A nonparametric measure of independence under a hypothesis of independent components. *Statistics and Probab. Letters*, **15**, 245–252.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman Hall: London.

Skaug, H. J. (1993a). The limit distribution of the Hoeffding statistic for test of serial independence. Preprint. Department of Mathematics, University of Bergen.

Skaug, H. J. (1993b). A class of tests for the simple goodness of fit hypothesis based on estimated densities. Preprint. Department of Mathematics, University of Bergen.

Skaug, H. J. and Tjøstheim, D. (1993a). *Nonparametric tests of serial independence*. In Developments in Time Series Analysis. The Priestley Birthday Volume, T. Subba Rao editor. Chapman and Hall: London, 207–230.

Skaug, H. J. and Tjøstheim, D. (1993b) A nonparametric test of serial independence based on the empirical distribution function. *Biometrika*, **80**, 591–602.

Skaug, H. J. and Tjøstheim, D. (1994). Measures of distance between densities with application to testing for serial independence. Preprint. Department of Mathematics, University of Bergen.

Spearman, C. (1904). The proof and measurement of association between two things. *Amer. J. of Psychology*, **15**, 88.

Subba Rao, T. and Gabr, M. M. (1980). A test for linearity of stationary time series. *J. of Time Ser. Anal.*, **3**, 169–176.

Tong, H. (1995). A personal overview of nonlinear time series from a chaos perspective (with discussion). *Scand. J. of Stat.*, **22**, 399–446.

Wald, A. and Wolfowitz, J. (1943). An exact test for randomness in the nonparametric case based on serial correlation. *Ann. Math. Stat.*, **14**, 378–388.

White, H. (1987). *Specification testing in dynamic models, in Advances in Econometrics*, Bewley, T. editor. University Press: New York.

Wolff, R. C. L. (1994). Independence in time series: Another look at the BDS test. Philosophical Transactions of the Royal Society of London, Series A, **348**, 383–395.

Zhang, G. and Taniguchi, M. (1994). Discrimination analysis for stationary time series. *J. of Time Ser. Anal.*, **15**, 117–126.

Dag Tjøstheim
Department of Mathematics
University of Bergen
5007 Bergen
NORWAY