J. R. Statist. Soc. B (1998)
60, *Part* 2, *pp.* 429–453

Testing for pairwise serial independence via the empirical distribution function

Yongmiao Hong†

Cornell University, Ithaca, USA

[Received July 1996. Revised May 1997]

Summary. Built on Skaug and Tjøstheim's approach, this paper proposes a new test for serial independence by comparing the pairwise empirical distribution functions of a time series with the products of its marginals for various lags, where the number of lags increases with the sample size and different lags are assigned different weights. Typically, the more recent information receives a larger weight. The test has some appealing attributes. It is consistent against all pairwise dependences and is powerful against alternatives whose dependence decays to zero as the lag increases. Although the test statistic is a weighted sum of degenerate Cramér–von Mises statistics, it has a null asymptotic $N(0, 1)$ distribution. The test statistic and its limit distribution are invariant to any order preserving transformation. The test applies to time series whose distributions can be discrete or continuous, with possibly infinite moments. Finally, the test statistic only involves ranking the observations and is computationally simple. It has the advantage of avoiding smoothed nonparametric estimation. A simulation experiment is conducted to study the finite sample performance of the proposed test in comparison with some related tests.

Keywords: Asymptotic normality; Cramér–von Mises statistic; Empirical distribution function; Hypothesis testing; Serial independence; Weighting

1. Introduction

The detection of serial dependence is important for non-Gaussian and non-linear time series. Tests for serial independence are useful diagnostic tools in fitting non-linear time series and identifying appropriate lags, especially when the time series has zero autocorrelation (see Granger and Teräsvirta (1993)). For instance, to detect non-linearity we can test whether the residuals from a linear fit are independent. We can also test the random walk hypothesis for a time series by testing whether its first differences are serially independent. Tests for independence are useful in other contexts as well (see Robinson (1991a)). Independence is still a testable hypothesis in populations with infinite second-order moments.

In practice, correlation tests (e.g. Box and Pierce (1970) and Ljung and Box (1978)) are widely used to test serial independence, but they are not consistent against alternatives with zero autocorrelation. The lack of consistency is unsatisfying from both theoretical and practical viewpoints. Also, these tests require second- or higher order moments, excluding their applications to time series with infinite second-order moments.

There have been many nonparametric tests for serial independence, e.g. Chan and Tran (1992), Delgado (1996), Hjellvik and Tjøstheim (1996), Pinkse (1997), Robinson (1991a) and

†*Address for correspondence*: Department of Economics and Department of Statistical Sciences, College of Arts and Sciences, Uris Hall, Cornell University, Ithaca, NY 14853-7601, USA.
E-mail: yh20@cornell.edu

 1369–7412/98/60429

Skaug and Tjøstheim (1993a, b, 1996). These tests are consistent against serial dependences up to a finite order. Some of them require smoothed nonparametric density estimation. Both theory and simulation studies (e.g. Skaug and Tjøstheim (1993b, 1996)) suggest that the finite sample performances of smoothed density-based tests may depend heavily on the choice of the smoothing parameters. Resampling methods such as bootstrap and permutation methods have been used to obtain accurate sizes (e.g. Chan and Tran (1992), Delgado (1996), Hjellvik and Tjøstheim (1996) and Skaug and Tjøstheim (1993b, 1996)).

One important test that avoids smoothed nonparametric estimation is Skaug and Tjøstheim's (1993a) test based on the empirical distribution function, which extends the test of Blum *et al.* (1961) to a time series setting in a significant way. The test statistic is invariant to any order preserving transformation. The distribution generating the data can be continuous or discrete; when it is continuous, the test is distribution free. No moment is required; this is attractive for time series whose variances are infinite, as often arises in economics and high frequency financial time series (e.g. Fama and Roll (1968) and Mandelbrot (1967)). Finally, the statistic depends only on ranking the observations and is simple to compute. We note that Delgado (1996) also used the empirical distribution function via an alternative approach.

Built on Skaug and Tjøstheim's (1993a) approach, this paper proposes a new test for serial independence. It compares the pairwise empirical distribution functions of a time series with the products of its marginals for various lags. The test has some new attributes. First, it is consistent against all pairwise dependences for continuous random variables. Thus, the test may be useful when no prior information is available. It may also be expected to have good power against long memory processes because a long lag is used (see Robinson (1991b)). Second, different weights are given to different lags; typically larger weights are given to lower order lags. Non-uniform weighting is expected to give better power than uniform weighting against alternatives whose dependence decays to zero as the lag increases, as is often observed for seasonally adjusted stationary time series. Third, although our statistic is a weighted sum of Cramér–von Mises statistics, it has a null asymptotic one-sided $N(0, 1)$ distribution no matter whether the data are generated from a discrete or a continuous distribution. We consider some 'leave-one-out' test statistics, which improve size performances in small samples. We also present a slightly modified version of Skaug and Tjøstheim's (1993a) test that improves the size in small samples when relatively many lags are used. We emphasize that the new test should be viewed as not competing with but as a complement to Skaug and Tjøstheim's (1993a) test, because they work in different regimes and have their own merits. Skaug and Tjøstheim's (1993a) test is relevant when relatively few lags are tested, whereas the test proposed applies for relatively many lags. In addition, the present test is built on Skaug and Tjøstheim's (1993a) approach, and one may perform better in some situations while the other performs better in other situations. Simulation shows that the new test has good power against linear — both short and long memory — processes and some non-linear processes, but like other tests based on the empirical distribution functions it has relatively low power against Engle's (1982) autoregressive conditional heteroscedastic process. For this class of alternatives, smoothed density-based tests (e.g. Skaug and Tjøstheim (1996)) may be expected to perform better.

Section 2 introduces the test statistic and derives its asymptotic normality. Section 3 establishes consistency against all pairwise dependences and derives an optimal weighting scheme for various lags. In Section 4, we use simulation methods to compare the new test with Skaug and Tjøstheim's (1993a) test, and Skaug and Tjøstheim's (1996) density-based test. The proofs are briefly sketched in Appendix A; details are available from the author on request.

2. The approach and test statistics

Consider a real-valued strictly stationary time series $\{X_t\}_{t=-\infty}^{\infty}$ with marginal distribution $G(x) = P(X_t \leqslant x)$ and pairwise distribution function $F_j(x, y) = P(X_t \leqslant x, X_{t-j} \leqslant y), (x, y) \in \mathbb{R}^2$, $j = 0, \pm 1, \ldots$. Suppose that a random sample $\{X_t\}_{t=1}^{n}$ of size n is observed. We are interested in testing the null hypothesis that $\{X_t\}$ is serially independent.

Existing tests for serial independence are all based on some measure of dependence. Because X_t and X_{t-j} are independent if and only if $g(X_t)$ and $g(X_{t-j})$ are independent for any continuous monotonic function g, it is desirable to use a measure that is invariant to g (see Skaug and Tjøstheim (1996) for further discussion). Hoeffding (1948) proposed the measure

$$D^2(j) = \int_{\mathbb{R}^2} \{F_j(u, v) - G(u)\, G(v)\}^2 \, \mathrm{d}F_j(u, v). \tag{1}$$

This is invariant to any order preserving transformation. By Hoeffding's (1948) theorem 3.1, when $F_j(u, v)$ has continuous pairwise and marginal density functions, $D^2(j) = 0$ if and only if X_t and X_{t-j} are independent. In this case, tests based on measure (1) are consistent against all pairwise dependences. If $F_j(u, v)$ is discontinuous, however, it is possible that $D^2(j) = 0$ but X_t and X_{t-j} are not independent. See Hoeffding (1948), p. 548, for an example. In this case, tests based on measure (1) are not consistent against all pairwise dependences.

Measure (1) or its analogue has been used to test independence. Hoeffding (1948) used a U-statistic estimator for an analogue of measure (1) to test independence between two identically and independently distributed random variables. Blum *et al.* (1961) used an empirical distribution function-based estimator for an analogue of measure (1) to test independence among the components of an identically and independently distributed random vector. Also see Deheuvels (1981) and Carlstein (1988).

For $j = 0, 1, \ldots$ define the pairwise empirical distribution function of (X_t, X_{t-j})

$$\hat{F}_j(x, y) = (n - j)^{-1} \sum_{t=j+1}^{n} \mathbf{1}(X_t \leqslant x)\, \mathbf{1}(X_{t-j} \leqslant y),$$

where $\mathbf{1}(\cdot)$ is the indicator function. A consistent estimator for measure (1) can be given by

$$\hat{D}_n^2(j) = (n - j)^{-1} \sum_{t=j+1}^{n} \{\hat{F}_j(X_t, X_{t-j}) - \hat{F}_j(X_t, \infty)\, \hat{F}_j(\infty, X_{t-j})\}^2. \tag{2}$$

Skaug and Tjøstheim (1993a) were the first to use measure (2) to test for serial independence. They considered the test statistic

$$\mathrm{ST1_a} = (n - 1) \sum_{j=1}^{p} \hat{D}_n^2(j). \tag{3}$$

When $\{X_t\}_{t=1}^{n}$ are independently and identically distributed,

$$\mathrm{ST1_a} \rightarrow \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_{ij} \chi_{ij}^2(p)$$

in distribution as $n \rightarrow \infty$, where the $\chi_{ij}^2(p)$ are independent χ^2 random variables with p degrees of freedom and the λ_{ij} are weights. For continuous random variables, $\lambda_{ij} = (ij\pi^2)^{-2}$, and the test is distribution free. For discrete random variables, the λ_{ij} depend on the data-generating process and must be estimated; the test is not distribution free.

Statistic (3) can be viewed as an appropriate and natural extension of Box and Pierce's

(1970) correlation test. It can detect pairwise dependences up to lag p, including those with zero autocorrelation. Some evidence (Skaug and Tjøstheim (1993a), p. 600) suggests that this pairwise approach has better power than an alternative approach that tests joint independence of $(X_{t-1}, \ldots, X_{t-p})$. It also gives a much simpler limit distribution than the joint testing approach. Skaug and Tjøstheim (1993a) documented that the most significant asset of statistic (3) is that it is very close to correlation tests in power against linear processes and is considerably better than correlation tests in power against a variety of non-linear processes, although it has relatively low power against Engle's (1982) autoregressive conditional heteroscedastic process. In addition, the size of statistic (3) is quite accurate for moderate sample sizes compared with existing smoothed density-based tests for serial independence, especially when p is small. For large p, Skaug and Tjøstheim (1993a) pointed out that the asymptotic critical values of statistic (3) progressively lead to a test with a level that is too high as p increases for small n. The bootstrap and permutation methods, advocated by Skaug and Tjøstheim (1993b, 1996) to produce the correct levels for smoothed density-based tests, can be applied to statistic (3) as well.

In this paper, we develop a distribution theory that gives a simple and reasonable approximation for statistic (3) when p is large. We show that for large p it is possible to obtain an $N(0, 1)$ limit distribution for statistic (3), after proper standardization. However, for power, we may want to check more lags as n increases. This ensures consistency against all pairwise dependences for continuous random variables. In addition, different weights can be given to different lags. In particular, more weights can be given to more recent information, i.e. to lower order lags. This may improve power against stationary time series whose dependence decays to zero as the lag increases. These considerations suggest the statistic

$$V_n = \sum_{j=1}^{n-1} k^2(j/p)(n-j)\, \hat{D}_n^2(j), \tag{4}$$

where k is a kernel function satisfying the following assumption.

Assumption 1. $k:\mathbb{R} \to [-1, 1]$ is symmetric, continuous at 0 and all except a finite number of points, with $k(0) = 1$, $\int_0^\infty k^2(z)\,\mathrm{d}z < \infty$ and $|k(z)| \leqslant C|z|^{-b}$ as $z \to \infty$ for some $b > \frac{1}{2}$ and $0 < C < \infty$.

This assumption helps to ensure that statistic (4), after division by n, converges in probability to $\Sigma_{j=1}^\infty D^2(j)$, thus delivering a consistent test against all pairwise dependences for continuous random variables. Here, the continuity of k with $k(0) = 1$ ensures that the bias of statistic (4) vanishes. The condition $\int_0^\infty k^2(z)\,\mathrm{d}z < \infty$ implies that $k(z) \to 0$ as $z \to \infty$. This ensures that the asymptotic variance of statistic (4) vanishes. The truncated, Bartlett, Daniell, Parzen, quadratic spectral and Tukey kernels (e.g. Priestley (1981), pages 441–442) all satisfy assumption 1. Except for the truncated kernel, all have non-uniform weights. Note that all $n-1$ lags are used in statistic (4) if k has unbounded support.

Statistic (4) is a weighted sum of von Mises statistics. When the truncated kernel (i.e. $k(z) = \mathbf{1}(|z| \leqslant 1)$) is used, we obtain

$$\mathrm{ST2_a} = \sum_{j=1}^{p} (n-j)\, \hat{D}_n^2(j), \tag{5}$$

which is asymptotically equivalent to statistic (3). Thus, we can see that both statistic (3) and statistic (5) are based on the truncated kernel or uniform weighting. The use of $n-j$ in statistic (5) is similar in spirit to Ljung and Box's (1978) modified version of Box and Pierce's

(1970) test. Our simulation study later shows that statistic (5) has better size than statistic (3) for large p in small samples.

We also consider a leave-one-out empirical distribution function

$$\hat{F}_{j(t)}(X_t, X_{t-j}) = (n-j-1)^{-1} \sum_{s=j+1, s\neq t}^{n} \mathbf{1}(X_s \leqslant X_t)\, \mathbf{1}(X_{s-j} \leqslant X_{t-j}),$$

where the sum excludes the tth observation. This yields the alternative statistic

$$V_n^* = \sum_{j=1}^{n-2} k^2(j/p)(n-j-1)\, \hat{D}_n^{*2}(j), \tag{6}$$

where

$$\hat{D}_n^{*2}(j) = (n-j)^{-1} \sum_{t=j+1}^{n} \{\hat{F}_{j(t)}(X_t, X_{t-j}) - \hat{F}_{j(t)}(X_t, \infty)\, \hat{F}_{j(t)}(\infty, X_{t-j})\}^2.$$

Both statistic (4) and statistic (6) have the same limit distribution, but asymptotic analysis shows that statistic (6) has a smaller approximation error for the limit distribution. Thus, statistic (6) may give better sizes in small samples, as is confirmed in our simulation study. Similarly, we can consider

$$\mathrm{ST1_b} = (n-1) \sum_{j=1}^{p} \hat{D}_n^{*2}(j), \tag{7}$$

$$\mathrm{ST2_b} = \sum_{j=1}^{p} (n-j-1)\, \hat{D}_n^{*2}(j). \tag{8}$$

These two tests have the same limit distribution as statistics (3) and (5).

We first give our generic test statistics.

Theorem 1. Suppose that assumption (1) holds, and $p = cn^{\nu}$ for some $0 < \nu < 1$ and $0 < c < \infty$. Define

$$M_a = \sum_{j=1}^{n-1} k^2\left(\frac{j}{p}\right)\{(n-j)\,\hat{D}_n^2(j) - \hat{A}_0\} \Big/ \left\{2\hat{B}_0 \sum_{j=1}^{n-2} k^4\left(\frac{j}{p}\right)\right\}^{1/2},$$

$$M_b = \sum_{j=1}^{n-2} k^2\left(\frac{j}{p}\right)\{(n-j-1)\,\hat{D}_n^{*2}(j) - \hat{A}_0\} \Big/ \left\{2\hat{B}_0 \sum_{j=1}^{n-3} k^4\left(\frac{j}{p}\right)\right\}^{1/2},$$

where

$$\hat{A}_0 = \left[n^{-1} \sum_{t=1}^{n} \hat{G}(X_t)\{1 - \hat{G}(X_t)\}\right]^2,$$

$$\hat{B}_0 = \left(n^{-2} \sum_{t=1}^{n} \sum_{s=1}^{n} [\hat{G}\{\min(X_t, X_s)\} - \hat{G}(X_t)\,\hat{G}(X_s)]^2\right)^2,$$

with

$$\hat{G}(u) = n^{-1} \sum_{t=1}^{n} \mathbf{1}(X_t \leqslant u).$$

If $\{X_t\}_{t=1}^n$ are identically and independently distributed, then $M_a - M_b \to 0$ in probability, and $M_a \to N(0, 1)$ and $M_b \to N(0, 1)$ in distribution.

Theorem 1 applies to discrete, continuous or mixed distributions. If $G(x)$ is known to be continuous, we can use the following simplified statistics.

Theorem 2. Suppose that the conditions of theorem 1 hold. Define

$$M_a = 90 \sum_{j=1}^{n-1} k^2\left(\frac{j}{p}\right)\left\{(n-j)\,\hat{D}_n^2(j) - \frac{1}{36}\right\} \Big/ \left\{2 \sum_{j=1}^{n-2} k^4\left(\frac{j}{p}\right)\right\}^{1/2},$$

$$M_b = 90 \sum_{j=1}^{n-2} k^2\left(\frac{j}{p}\right)\left\{(n-j-1)\,\hat{D}_n^{*2}(j) - \frac{1}{36}\right\} \Big/ \left\{2 \sum_{j=1}^{n-3} k^4\left(\frac{j}{p}\right)\right\}^{1/2}.$$

If $\{X_t\}_{t=1}^n$ are identically and independently distributed continuous random variables, then $M_a - M_b \to 0$ in probability, and $M_a \to N(0, 1)$ and $M_b \to N(0, 1)$ in distribution.

We note that theorem 2 does not apply to discrete distributions, because the mean and asymptotic variance of $(n-j)\,\hat{D}_n^2(j)$ cannot be computed in discrete cases.

Although the von Mises statistic $(n-j)\,\hat{D}_n^2(j)$ does not follow an asymptotic χ^2-distribution, theorems 1 and 2 show that a weighted sum of $(n-j)\,\hat{D}_n^2(j)$ has an asymptotic $N(0, 1)$ distribution for large p after centring and scaling. To obtain an intuitive idea, consider for example the use of the truncated kernel. In this case, M_a of theorem 2 becomes

$$M_a = \sum_{j=1}^{p} \left\{(n-j)\,\hat{D}_n^2(j) - \frac{1}{36}\right\} \Big/ \left(\frac{2p}{90^2}\right)^{1/2}.$$

As shown in Skaug and Tjøstheim (1993a),

$$(n-j)\,\hat{D}_n^2(j) \to \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} (k\pi)^{-2}(l\pi)^{-2}\,\chi_{kl}^2(1)$$

in distribution as $n \to \infty$, thus having mean $1/36$ and asymptotic variance $2/90^2$. In addition, $\mathrm{cov}\{(n-i)\,\hat{D}_n^2(i), (n-j)\,\hat{D}_n^2(j)\} \to 0$ for $i \neq j$ as $n \to \infty$, suggesting that $(n-i)\,\hat{D}_n^2(i)$ and $(n-j)\,\hat{D}_n^2(j)$ are asymptotically uncorrelated or independent. Therefore, $\{(n-j)\,\hat{D}_n^2(j), j = 1, \ldots, p\}$ can be viewed as an asymptotically identically and independently distributed sequence with mean $1/36$ and variance $2/90^2$. The sum of this sequence, after differencing the mean and dividing by the standard deviation, will converge to $N(0, 1)$ in distribution as p becomes large. Our proof, of course, does not depend on this simplistic intuition. Instead, we develop a dependent degenerate V-statistic projection theory to approximate statistic (4) as a weighted sum of degenerate V-statistics over lags, which, after standardization, is then shown to be asymptotically $N(0, 1)$ by using an appropriate martingale limit theorem (e.g. Brown (1971)). See Appendix A for more details. For degenerate V-statistics (or related U-statistics) of dependent processes, see (for example) Carlstein (1988) and Sen (1963).

The $N(0, 1)$ approximation is convenient for inference. For small n or small p, however, it may not be accurate. As a practical alternative, the bootstrap or permutation, as advocated in Skaug and Tjøstheim (1993b, 1996), can be used as a remedy for obtaining the right level. Indeed, the current test situation is ideally suited to these resampling methods, which can be expected to yield as good a level as the best asymptotic approximation. In particular, permutation gives the exact level. However, it is impossible to compute the level exactly in practice for all except very small sample sizes. Hence, Monte Carlo methods must be used to obtain an approximation.

3. Consistency

To state our consistency theorem, we impose some additional regularity conditions.

Assumption 2.

$$\int_0^\infty z^{1/2}\, k^2(z)\, \mathrm{d}z < \infty.$$

Assumption 3. Put $U_t = G(X_t)$, where G is the continuous marginal distribution of the strictly stationary mixing process $\{X_t\}$ with strong mixing coefficient $\alpha(j) = O(j^{-1-\delta})$ as $j \to \infty$ for some $\delta > 0$. The joint distribution of (U_t, U_{t-j}) has a continuous joint density function on $[0, 1]^2$ that is bounded by some $\Delta < \infty$, where Δ does not depend on j.

Theorem 3. Suppose that assumptions 1–3 hold, and $p = cn^\nu$ for some $0 < \nu < \frac{1}{3}$ and $0 < c < \infty$. Let M_a and M_b be defined as in theorem 2. Then, with probability approaching 1 as $n \to \infty$,

$$(p^{1/2}/n)M_\mathrm{a} \to 90 \sum_{j=1}^{\infty} D^2(j) \Big/ \left\{ 2 \int_0^\infty k^4(z)\, \mathrm{d}z \right\}^{1/2},$$

$$(p^{1/2}/n)M_\mathrm{b} \to 90 \sum_{j=1}^{\infty} D^2(j) \Big/ \left\{ 2 \int_0^\infty k^4(z)\, \mathrm{d}z \right\}^{1/2}.$$

Theorem 3 implies that $\lim_{n\to\infty}\{P(M_\mathrm{a} > C_n)\} = 1$ for any non-stochastic sequence $\{C_n = o(n/p^{1/2})\}$, thus ensuring consistency of M_a against all pairwise dependences for continuous random variables. It suggests that when more data become available the tests M_a and M_b have power against an increasingly larger class of alternatives. Thus, M_a and M_b are useful when we have no prior information about possible alternatives. Because negative values of M_a and M_b can occur only under the null hypothesis asymptotically, M_a and M_b are one-sided tests; appropriate upper-tailed critical values of $N(0, 1)$ should be used.

Although the temporal condition on $\alpha(j)$ is mild, it rules out strong dependences. However, it is reasonable that any independence test should be expected to have strong power against strong dependences. In particular, M_a and M_b may be expected to have good power against strong dependences because a long lag is used. We shall investigate the power of M_a and M_b against long memory processes via simulation. It should also be noted that theorem 3 holds only for continuous random variables. For discrete random variables, M_a and M_b are not consistent against all pairwise dependences, because $D^2(j)$ can be 0 even if the time series is not pairwise independent. Furthermore, like other asymptotic notions, consistency against all pairwise dependences seems to be mostly of theoretical interest. In practice, p is always finite given any sample size n. When a kernel with bounded support (i.e. $k(z) = 0$ if $|z| > 1$) is used, p is the lag truncation number, and only the first p lags are tested. When a kernel with unbounded support is used, p is not a lag truncation number. In this case, all $n - 1$ lags are used, but the contributions from lags much larger than p are negligible.

An important issue is the choice of k. Intuitively, for stationary time series whose dependence as measured by statistic (1) decays to zero as the lag increases, it seems more efficient to give more weights to lower order lags. The asymptotic analysis below shows that this is indeed the case. To compare the asymptotic relative efficiency between two kernels, say k_1 and k_2, we use Bahadur's (1960) asymptotic slope criterion, which is suitable for large sample tests under fixed alternatives. Bahadur's asymptotic slope is the rate at which the asymptotic p-value of the test statistic goes to 0 as $n \to \infty$. Because M_a is asymptotically $N(0, 1)$ under the null hypothesis, its asymptotic p-value is $1 - \Phi(M_a)$, where Φ is the cumulative distribution function of $N(0, 1)$. Define

$$S_n(k) = -2 \ln\{1 - \Phi(M_a)\}.$$

Given $\ln\{1 - \Phi(z)\} = -\frac{1}{2} z^2 \{1 + o(1)\}$ as $z \to \infty$, we have by theorem 3 that

$$(p/n^2)\, S_n(k) \to 90^2 \left\{ \sum_{j=1}^{\infty} D^2(j) \right\}^2 \Big/ \left\{ 2 \int_0^{\infty} k^4(z)\,\mathrm{d}z \right\}$$

in probability. Following Bahadur (1960), we call $90^2\{\Sigma_{j=1}^{\infty} D^2(j)\}^2/2 \int_0^{\infty} k^4(z)\,\mathrm{d}z$ the asymptotic slope of M_a. Under the conditions of theorem 3, it is straightforward to show that Bahadur's asymptotic relative efficiency of k_2 to k_1 is

$$\mathrm{ARE}(k_2 : k_1) = \left\{ \int_0^{\infty} k_1^4(z)\,\mathrm{d}z \Big/ \int_0^{\infty} k_2^4(z)\,\mathrm{d}z \right\}^{1/(2-\nu)}.$$

Thus, k_2 is more efficient than k_1 if

$$\int_0^{\infty} k_2^4(z)\,\mathrm{d}z < \int_0^{\infty} k_1^4(z)\,\mathrm{d}z.$$

For example, Bahadur's asymptotic efficiency of the Bartlett kernel ($k_B(z) = (1 - |z|)\,\mathbf{1}(|z| \leqslant 1)$) to the truncated kernel ($k_T(z) = \mathbf{1}(|z| \leqslant 1)$) is $\mathrm{ARE}(k_B : k_T) = 5^{1/(2-\nu)} > 2.23$. Following reasoning that is analogous to Hong's (1996a, b) local power analysis for some correlation tests, we can obtain that the Daniell kernel

$$k(z) = \frac{\sin(\tau z \sqrt{3})}{\tau z \sqrt{3}}, \qquad z \in \mathbb{R}, \tag{9}$$

maximizes the Bahadur asymptotic slope over the class of kernels

$$\mathbb{K}(\tau) = \{k \text{ satisfies assumptions 1, 2, } k^{(2)} = \tau^2/2 > 0,\ K(\lambda) \geqslant 0 \text{ for } \lambda \in \mathbb{R}\}, \tag{10}$$

where

$$k^{(r)} = \lim_{z \to 0} [\{1 - k(z)\}/|z|^r]$$

is called the 'characteristic exponent' of function k and

$$K(\lambda) = (2\pi)^{-1} \int_{-\infty}^{\infty} k(z) \exp(-\mathrm{i}\lambda z)\,\mathrm{d}z$$

is the Fourier transform of k. This class includes the Daniell, Parzen and quadratic spectral kernels but rules out the truncated ($r = \infty$) and Bartlett ($r = 1$) kernels. Mainly because of assumption 2, $\mathbb{K}(\tau)$ is more restrictive than a class of kernels often considered to derive optimal kernels using appropriate criteria (see Andrews (1991) and Priestley (1962)) in the spectral analysis literature. Hong (1996a, b) showed that the Daniell kernel (9) maximizes the local power of some correlation tests over a class of kernels slightly more general than $\mathbb{K}(\tau)$. The result obtained here suggests that the optimality of the Daniell kernel (9) carries over to the non-linear Hoeffding measure (1). However, the optimality of the Daniell kernel (9) seems to be only of theoretical interest. Simulation studies later show that commonly used non-uniform kernels have similar powers for M_a and M_b, but they are all more powerful than the truncated kernel in most cases.

One advantage of Skaug and Tjøstheim's (1993a) test is that it does not require choosing any kernel function and p as a function of n. For M_a and M_b, p behaves like a smoothing parameter as we require $p \to \infty$ as $n \to \infty$ to ensure asymptotic null normality and consistency against all pairwise dependences. Of course, M_a and M_b are different from smoothed density-based tests because M_a and M_b do not require smoothed nonparametric estimation for each lag. To test all pairwise dependences, a smoothed density-based test would also need to let $p \to \infty$ as $n \to \infty$, in addition to the original smoothing parameter for density estimation. In the present context, the optimal p for M_a and M_b depends on the alternative dependence structure. Skaug and Tjøstheim (1993a) noted that choosing p too large will decrease the power of statistic (3). The same is true of M_a and M_b as well in general. Skaug and Tjøstheim (1993a) recommended choosing p larger than the smallest significant lag included in the model. It is reasonable to use a 'rule of thumb' consisting of taking p equal to the largest significant lag included in the model. Of course, this rule may not make sense for certain alternatives such as strongly dependent processes. We investigate the effect of the choice of p on both size and power via simulation. Our simulation finds that this rule of thumb applies well to the truncated kernel, but for non-uniform kernels maximal power is often achieved for p larger than the largest significant lag included in the model. In addition, the use of the non-uniform kernel alleviates the loss from choosing p too large because the kernel discounts the loss of degrees of freedom for higher order lags; this makes the power less sensitive to the choice of p. Nevertheless, the issue of choosing an optimal p is important. We defer this complicated issue to other work.

4. Monte Carlo evidence

We now study finite sample performances of the tests proposed. In addition to an identically and independently distributed process, we also consider following alternatives: AR(1),

$$X_t = 0.2X_{t-1} + \epsilon_t;$$

ARFIMA(0, d, 0),

$$X_t = (1-L)^{-0.2}\epsilon_t, \qquad LX_t = X_{t-1};$$

ARCH(1),

$$X_t = \epsilon_t(1 + 0.5X_{t-1}^2)^{1/2};$$

GARCH(1, 1),

$$X_t = \epsilon_t h_t^{1/2}, \qquad h_t = 1 + 0.1X_{t-1}^2 + 0.8h_{t-1};$$

TAR(1),

$$X_t = \begin{cases} -0.5X_{t-1} + \epsilon_t, & \text{if } X_{t-1} \geqslant 1, \\ 0.4X_{t-1} + \epsilon_t, & \text{if } X_{t-1} < 1; \end{cases}$$

NMA,

$$X_t = 0.8\epsilon_{t-1}\epsilon_{t-2} + \epsilon_t;$$

NAR(3)

$$X_t = 0.25X_{t-2} - 0.4\epsilon_{t-1}X_{t-1} + 0.2\epsilon_{t-2}X_{t-2} + 0.5\epsilon_{t-3}X_{t-3} + \epsilon_t;$$

NAR(5)

$$X_t = 0.3(\epsilon_{t-1}X_{t-1} - \epsilon_{t-2}X_{t-2} + \epsilon_{t-3}X_{t-3} - \epsilon_{t-4}X_{t-4} - \epsilon_{t-5}X_{t-5}) + \epsilon_t;$$

EXP(3)

$$X_t = 0.2 \sum_{j=1}^{3} X_{t-j} \exp(-\tfrac{1}{2}X_{t-j}^2) + \epsilon_t;$$

EXP(10),

$$X_t = 0.8 \sum_{j=1}^{10} X_{t-j} \exp(-\tfrac{1}{2}X_{t-j}^2) + \epsilon_t.$$

Here, AR(1) is a first-order autoregressive process, ARFIMA(0, d, 0) is a fractionally differenced integrated process, a popular long memory model (see Robinson (1991b, 1994)), ARCH(1) is a first-order autoregressive conditional heteroscedastic process, GARCH(1, 1) is a generalized ARCH process of order (1, 1), TAR(1) is a first-order threshold autoregressive process, NMA is a non-linear moving average process, NAR(3) and NAR(5) are non-linear autoregressive processes of orders 3 and 5 respectively, and EXP(3) and EXP(10) are non-linear exponential autoregressive processes of orders 3 and 10 respectively. These models are a fairly representative selection of both linear and non-linear time series. Except for the ARFIMA(0, d, 0) process, all have been used in a variety of existing simulation studies for independence tests. We consider two types of ϵ_t: normal and log-normal with zero mean and unit variance, for $n = 100$ and $n = 200$. The process ARFIMA(0, d, 0) is generated using Davies and Harte's (1987) fast Fourier transform algorithm. Except for the ARFIMA(0, d, 0) process, we generated $n + 100$ observations for each n and then discarded the first 100 to reduce the effect of initial values. For the ARFIMA(0, d, 0) process, we generated $n + 156 = 2^8$ observations and discarded the first 156 for $n = 100$, and generated $n + 312 = 2^9$ observations and discarded the first 312 for $n = 200$. The simulation experiments were conducted using the GAUSS 386 random number generator on a Cyrix personal computer.

To examine effects of various weights on the tests proposed, we considered five kernels: the truncated, Bartlett, Daniell, Parzen and quadratic spectral kernels. The last three kernels belong to the class $\mathbb{K}(\pi/\sqrt{3})$ in expression (10). To investigate the effect of choosing different values of p, we considered p from 1 to 20; this covers a sufficiently large range of p given $n = 100$ and $n = 200$. We also studied Skaug and Tjøstheim's (1993a) test $\mathrm{ST1_a}$ and its variants $\mathrm{ST1_b}$, $\mathrm{ST2_a}$ and $\mathrm{ST2_b}$, as well as Skaug and Tjøstheim's (1996) smoothed density-based test

$$J = \sum_{j=1}^{p} \hat{J}_n(j),$$

where

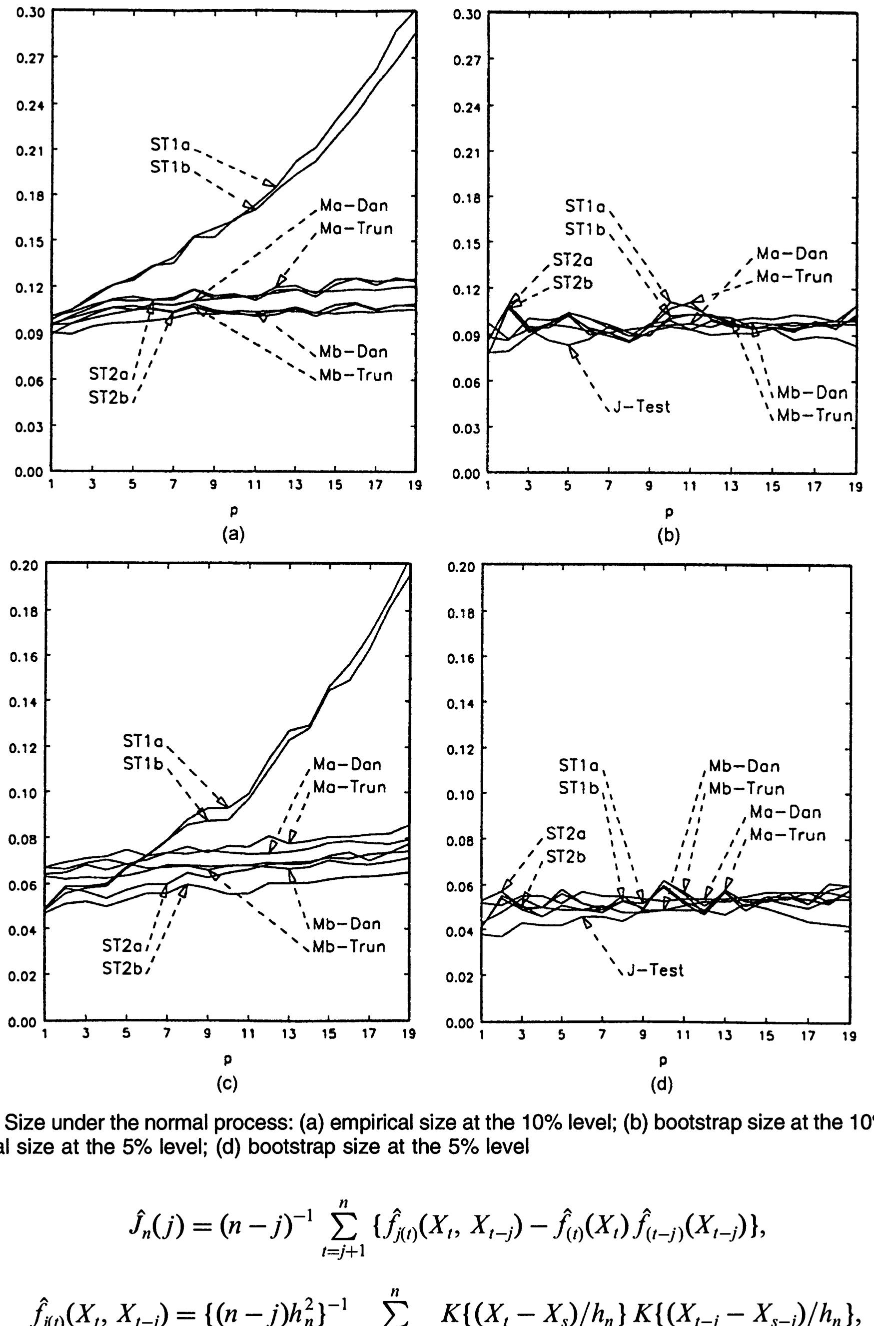

Fig. 1. Size under the normal process: (a) empirical size at the 10% level; (b) bootstrap size at the 10% level; (c) empirical size at the 5% level; (d) bootstrap size at the 5% level

$$\hat{J}_n(j) = (n-j)^{-1} \sum_{t=j+1}^{n} \{\hat{f}_{j(t)}(X_t, X_{t-j}) - \hat{f}_{(t)}(X_t)\hat{f}_{(t-j)}(X_{t-j})\},$$

$$\hat{f}_{j(t)}(X_t, X_{t-j}) = \{(n-j)h_n^2\}^{-1} \sum_{s=j+1, s\neq t}^{n} K\{(X_t - X_s)/h_n\}\, K\{(X_{t-j} - X_{s-j})/h_n\},$$

$$\hat{f}_{(t)}(X_t) = \{(n-1)h_n\}^{-1} \sum_{s=1, s\neq t}^{n} K\{(X_t - X_s)/h_n\},$$

with $K: \mathbb{R} \to \mathbb{R}^+$ a kernel function and h_n a bandwidth. Here, 'leave-one-out' bivariate and marginal density estimators are used. As in Skaug and Tjøstheim (1996), we first standardized

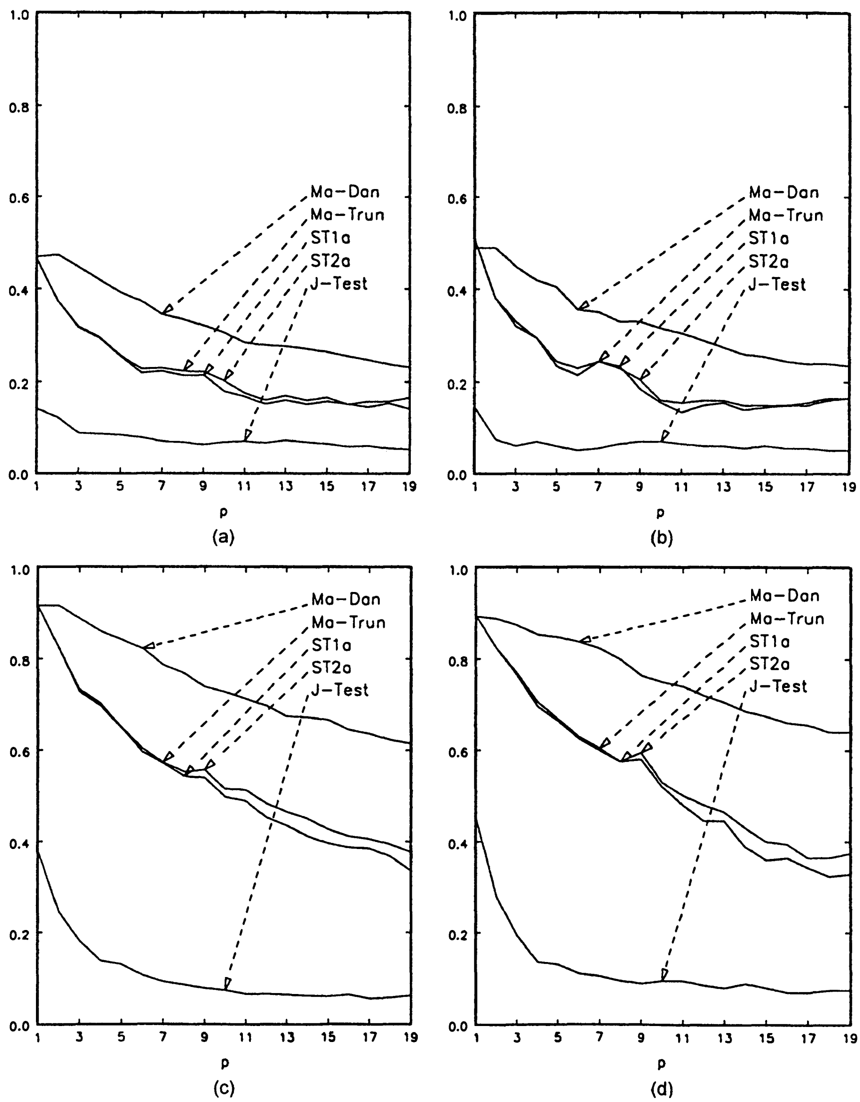

Fig. 2. Power at the 5% level under AR(1) and ARFIMA(0, *d*, 0) processes: (a) size-corrected power, AR(1), normal error; (b) bootstrap power, AR(1), normal error; (c) size-corrected power, AR(1), log-normal error; (d) bootstrap power, AR(1), log-normal error; (e) size-corrected power, ARFIMA(0, *d*, 0), normal error; (f) bootstrap power, ARFIMA(0, *d*, 0), normal error; (g) size-corrected power, ARFIMA(0, *d*, 0), log-normal error; (h) bootstrap power, ARFIMA(0, *d*, 0), log-normal error

the data by the sample deviation and used $h = n^{-1/6}$ with the Gaussian kernel. This test has been proven to be more powerful than or comparable with many existing smoothed density-based tests against a variety of non-linear processes (see Skaug and Tjøstheim (1993a, b, 1996)). Under independence, $(n/p)^{1/2}J \to N(0, \sigma_1^2)$ in distribution as $n \to \infty$ for some σ_1^2.

In addition to asymptotic inference, we also used a bootstrap procedure for all the tests to study bootstrap size and power. To describe the procedure, we first consider M_a. Given a sample $\{X_t\}_{t=1}^n$, we generate m bootstrap samples $\{X_{tl}^*\}_{t=1}^n$, $l = 1, \ldots, m$. For each bootstrap

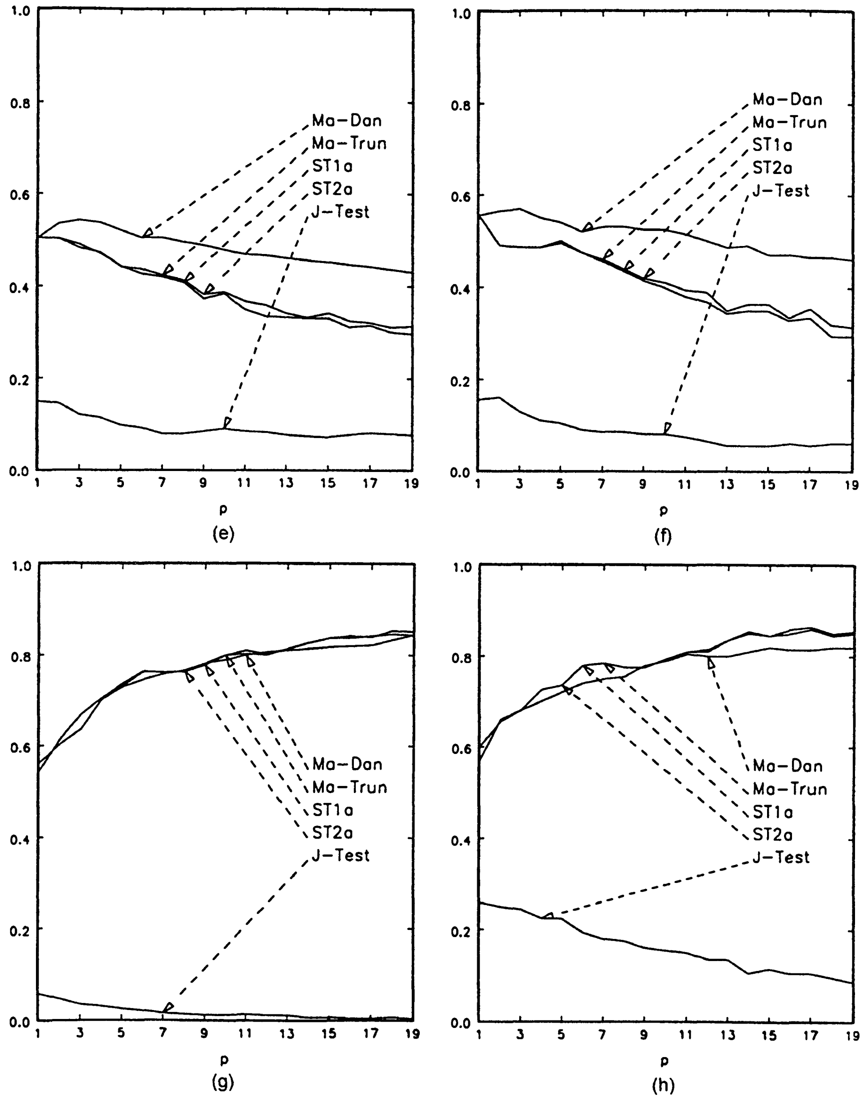

Fig. 2. (*continued*)

sample $\{X^*_{tl}\}^n_{t=1}$, we compute a bootstrap statistic M^*_{al}. To estimate the distribution function of the bootstrap statistic M^*_a, we could use the empirical distribution function of $\{M^*_{al}\}^m_{l=1}$ if m is sufficiently large. Because our simulation experiment is rather extensive, we choose $m = 50$. For such a small m, we follow Hjellvik and Tjøstheim (1996), section 3.3, to estimate the bootstrap distribution function $F^*(z) = P(M^*_a \leqslant z)$ by a smoothed nonparametric kernel estimator

$$\hat{F}^*_m(z) = \int_{-\infty}^{z} \left[m^{-1} \sum_{l=1}^{m} \{h_m \surd(2\pi)\}^{-1} \exp\{-(y - M^*_{al})^2/2h^2_m\} \right] \mathrm{d}y$$

$$= m^{-1} \sum_{l=1}^{m} \Phi\{(z - M^*_{al})/h_m\},$$

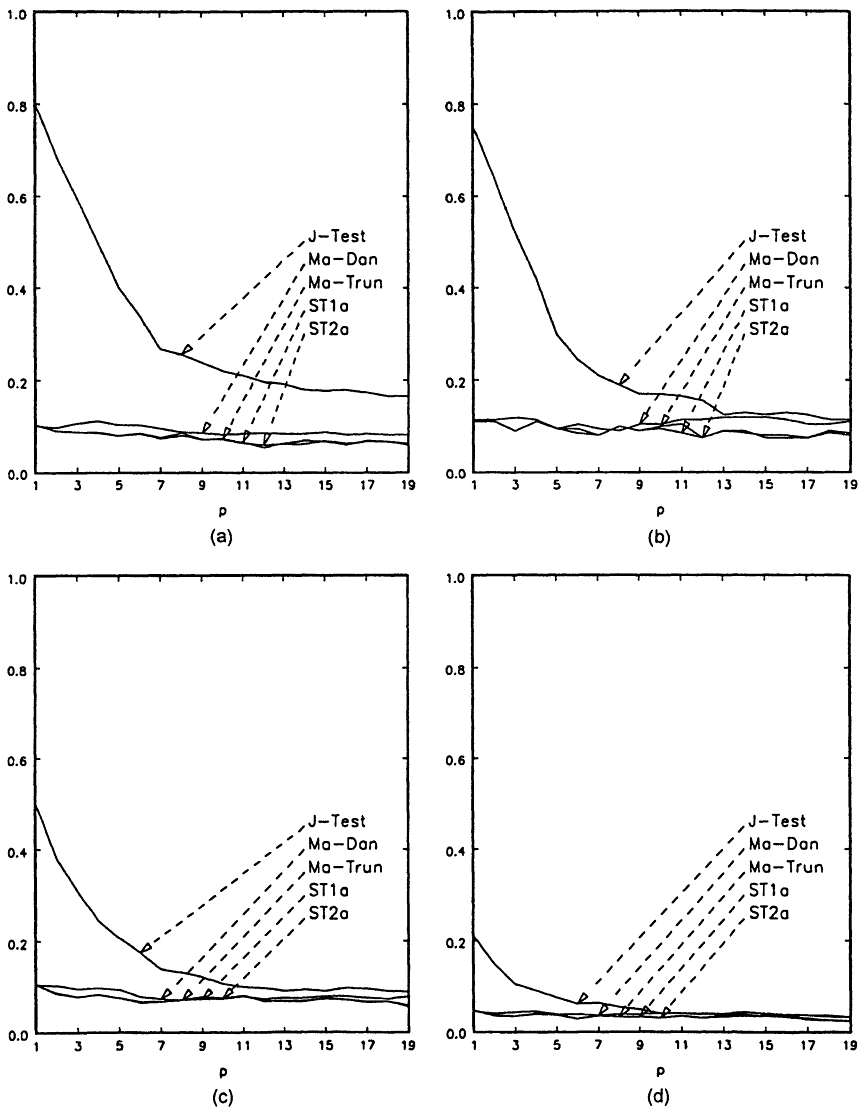

Fig. 3. Power at the 5% level under ARCH(1) and GARCH(1, 1) processes: (a) size-corrected power, ARCH(1), normal error; (b) bootstrap power, ARCH(1), normal error; (c) size-corrected power, ARCH(1), log-normal error; (d) bootstrap power, ARCH(1), log-normal error; (e) size-corrected power, GARCH(1, 1), normal error; (f) bootstrap power, GARCH(1, 1), normal error; (g) size-corrected power, GARCH(1, 1), log-normal error; (h) bootstrap power, GARCH(1, 1), log-normal error

where Φ is the cumulative distribution function of $N(0, 1)$, $h_m = \hat{S}_m m^{-1/5}$ is the bandwidth and $\hat{S}_m$ is the sample standard deviation of $\{M^*_{al}\}_{l=1}^m$. For more details, see Hjellvik and Tjøstheim (1996). Having obtained $\hat{F}^*_m$, we can reject the null hypothesis of independence if $\hat{F}^*_m(M_a) > 1 - \alpha$, where α is the significance level and M_a is the test statistic based on the original series $\{X_t\}_{t=1}^n$. The same procedure is applied to the other tests.

We first consider sizes by using both asymptotic and bootstrap critical values. For $ST1_a$, $ST1_b$, $ST2_a$ and $ST2_b$, we tabulate their asymptotic critical values using 10000 simulations of

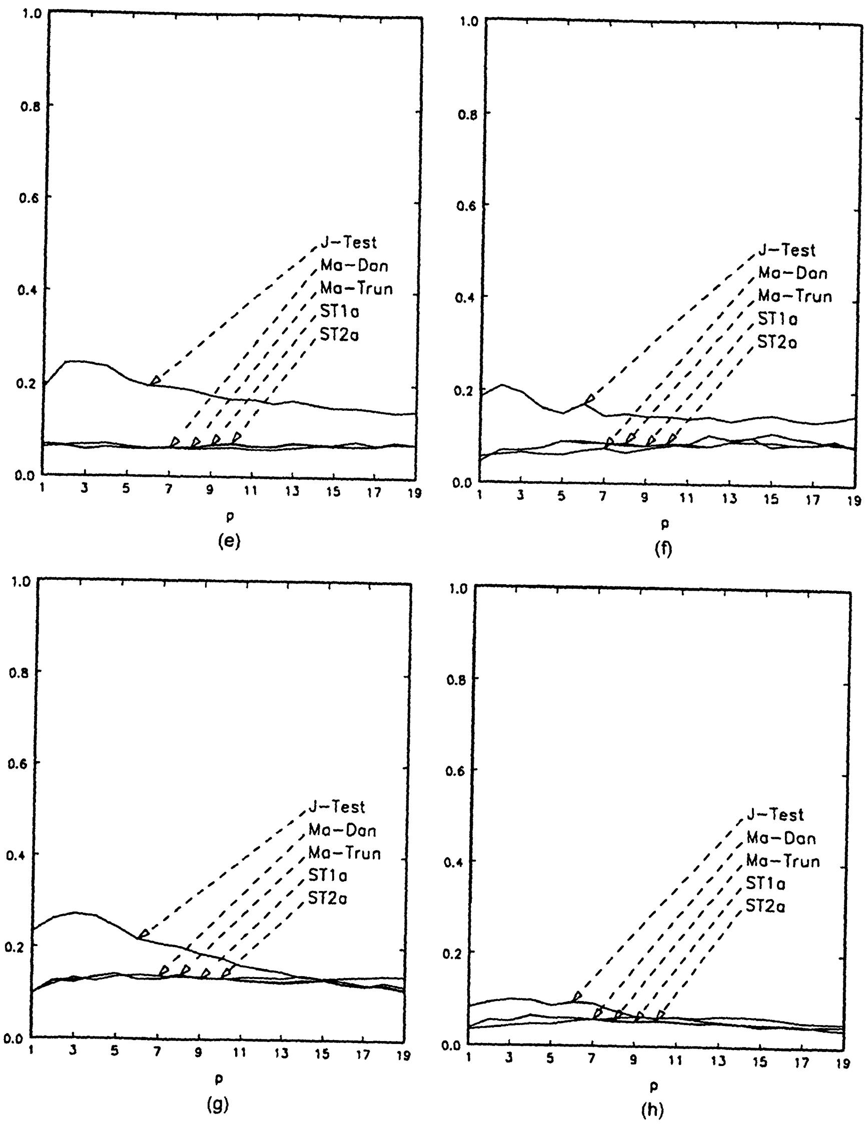

Fig. 3. (*continued*)

$$\sum_{i=1}^{200}\sum_{j=1}^{200}(ij\pi^2)^{-2}\,\chi^2_{ij}(p),$$

a truncated version of the limit distribution of $\mathrm{ST1_a}$. For the *J*-test, we only study its bootstrap size. Sizes for asymptotic and bootstrap critical values are based on 5000 and 1000 replications respectively. Fig. 1 reports sizes of M_a, M_b, $\mathrm{ST1_a}$, $\mathrm{ST1_b}$, $\mathrm{ST2_a}$, $\mathrm{ST2_b}$ and the *J*-test under the null hypothesis of independence at the 10% and 5% levels, for $n = 100$ and normal innovations. We only report the Daniell and truncated kernels for M_a and M_b, because the four non-uniform kernels perform similarly. We have the following observations.

(a) In terms of asymptotic critical values, M_a and M_b have reasonable sizes, although they

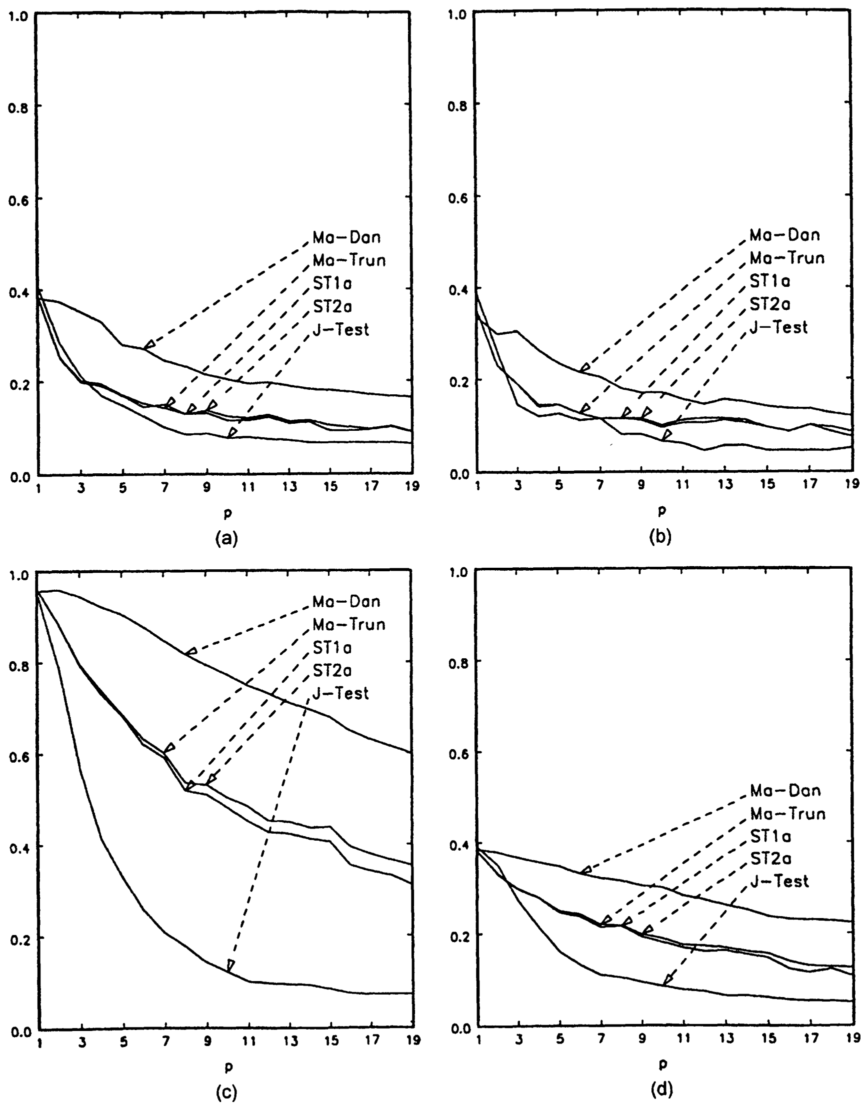

Fig. 4. Power at the 5% level under TAR(1) and NMA processes: (a) size-corrected power, TAR(1), normal error; (b) bootstrap power, TAR(1), normal error; (c) size-corrected power, TAR(1), log-normal error; (d) bootstrap power, TAR(1), log-normal error; (e) size-corrected power, NMA, normal error; (f) bootstrap power, NMA, normal error; (g) size-corrected power, NMA, log-normal error; (h) bootstrap power, NMA, log-normal error

show slight overrejections at the 5% level. Both the Daniell and the truncated kernels have similar sizes. Sizes are robust to the choice of p. The leave-one-out version M_b has slightly better sizes than M_a has, suggesting some gain from using the leave-one-out statistics. The tests $\mathrm{ST1}_a$, $\mathrm{ST2}_a$, $\mathrm{ST1}_b$ and $\mathrm{ST2}_b$ have precise sizes for small p. For large p, $\mathrm{ST1}_a$ and $\mathrm{ST1}_b$ tend to give some overrejection, but the modified statistics $\mathrm{ST2}_a$ and $\mathrm{ST2}_b$ continue to have good sizes. The leave-one-out version $\mathrm{ST2}_b$ has slightly better sizes than $\mathrm{ST2}_a$.

(b) All the tests have good bootstrap sizes. In particular, bootstrap sizes are better than

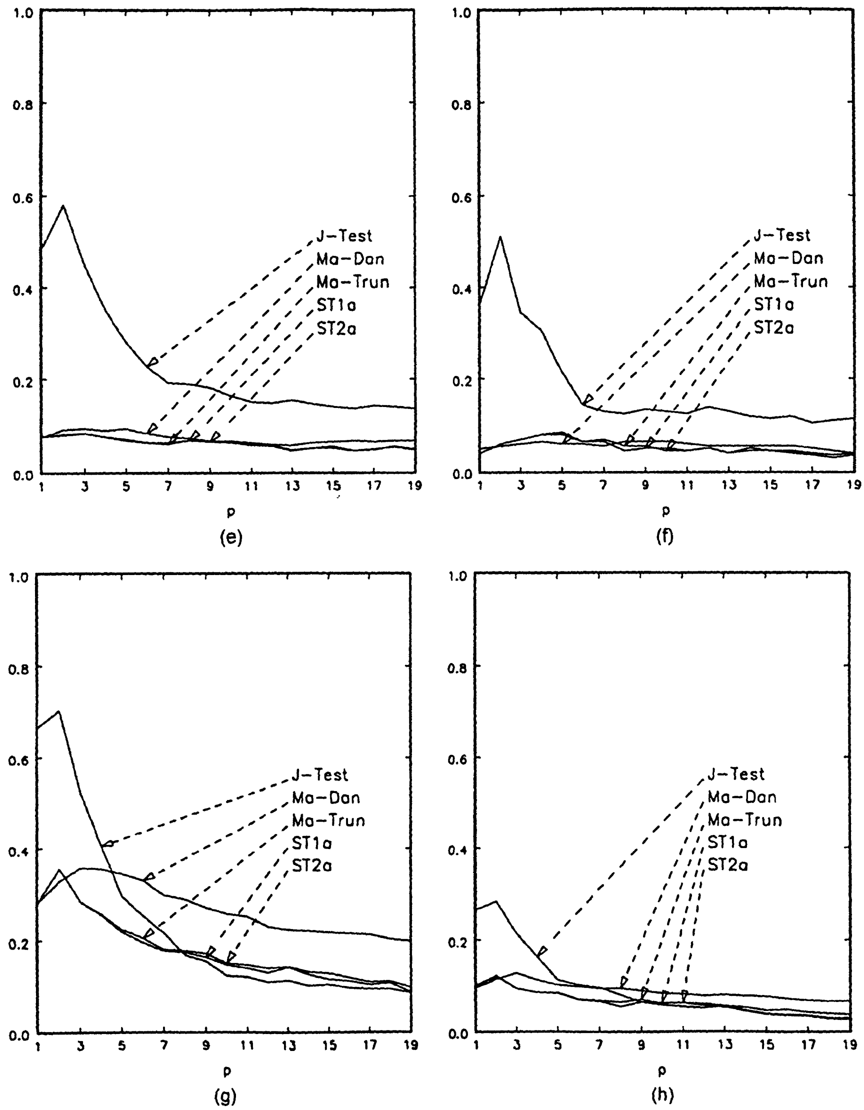

Fig. 4. (*continued*)

those given by the asymptotic approximation at the 5% level for all the tests.

We also consider sizes at the 1% level. At this level, the asymptotic approximation gives overrejections to various degrees for all the tests in most cases. The bootstrap sizes are significantly better and reasonable, with the *J*-test having the best size under normal innovations. The tests based on the empirical distribution function have invariant sizes at all the three levels under both normal and log-normal innovations. For $n = 200$, we studied sizes only using asymptotic critical values. The sizes are improved for all the tests, especially for $ST1_a$ and $ST2_a$.

Figs 2–6 report power at the 5% level for $n = 100$. To compare the tests on an equal basis, we use empirical critical values simulated from 5000 replications under the null hypothesis of

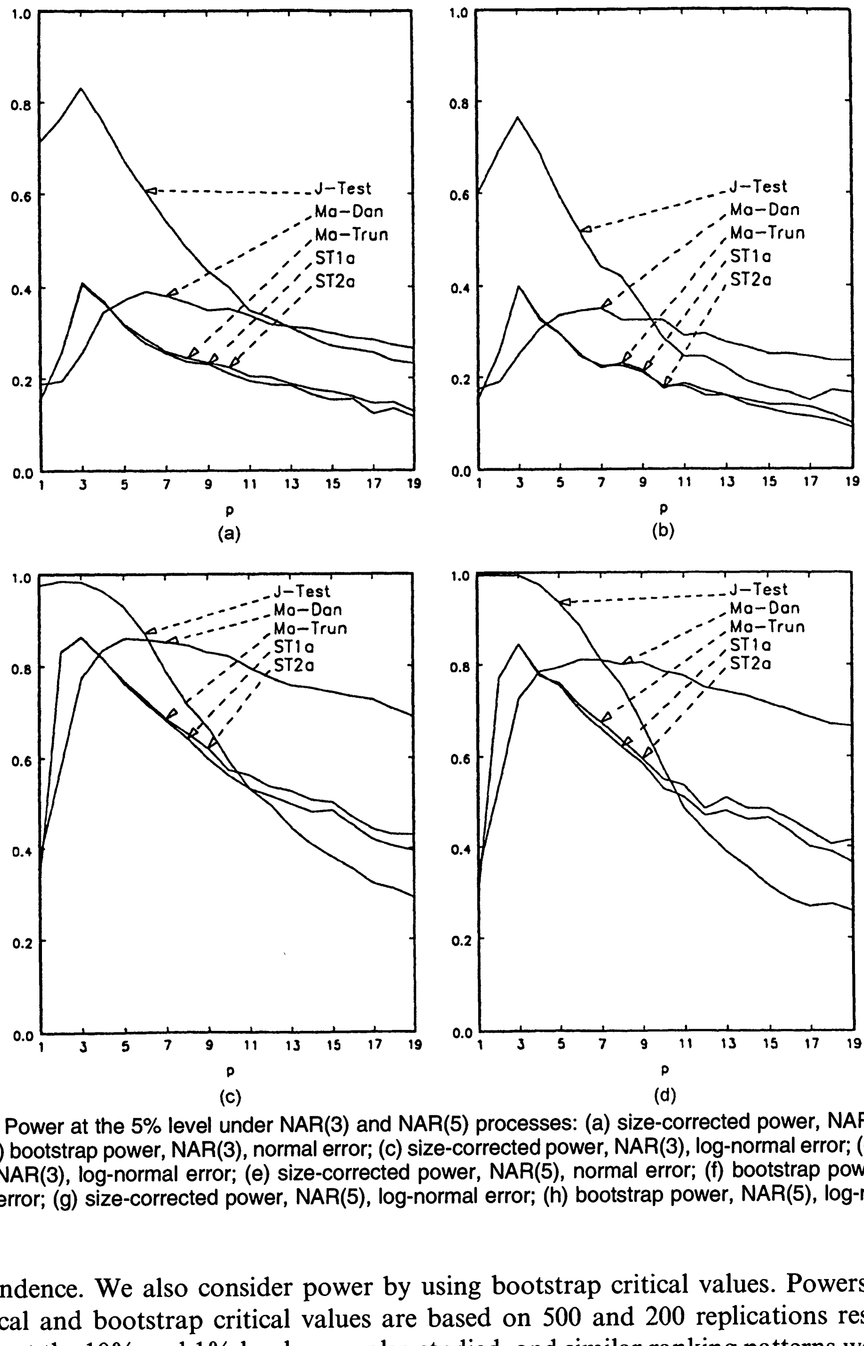

Fig. 5. Power at the 5% level under NAR(3) and NAR(5) processes: (a) size-corrected power, NAR(3), normal error; (b) bootstrap power, NAR(3), normal error; (c) size-corrected power, NAR(3), log-normal error; (d) bootstrap power, NAR(3), log-normal error; (e) size-corrected power, NAR(5), normal error; (f) bootstrap power, NAR(5), normal error; (g) size-corrected power, NAR(5), log-normal error; (h) bootstrap power, NAR(5), log-normal error

independence. We also consider power by using bootstrap critical values. Powers by using empirical and bootstrap critical values are based on 500 and 200 replications respectively. Powers at the 10% and 1% levels were also studied, and similar ranking patterns were found. Both M_a and M_b have similar powers. Also, $ST1_a$ and $ST2_a$ have power similar to $ST1_b$ and $ST2_b$ respectively. For clarity, we only report results for M_a, $ST1_a$, $ST2_a$ and the *J*-test. Again, only the Daniell and the truncated kernels for M_a are reported, because the four non-uniform kernels have similar powers. (In some cases, the Bartlett kernel is slightly more powerful than the other three non-uniform kernels, but this is not inconsistent with the

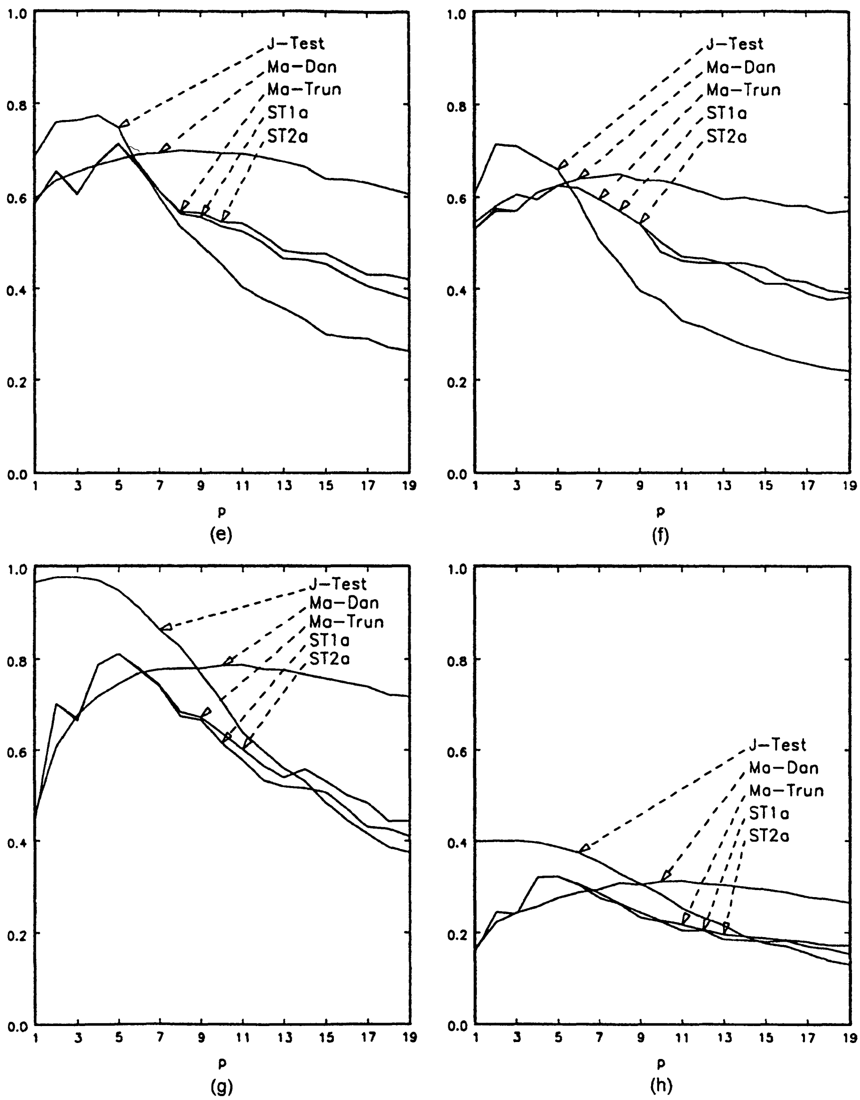

Fig. 5. (*continued*)

optimality of the Daniell kernel because the Bartlett kernel is outside the class of kernels over which the Daniell kernel is optimal.) We can make the following observations.

(a) For M_a, the Daniell kernel is more powerful than or comparable with the truncated kernel in most cases, suggesting the gains from using non-uniform weighting. For the alternatives NAR(3), NAR(5), EXP(3) and EXP(10), the truncated kernel is more powerful than or comparable with the Daniell kernel for small p, but as p becomes large the Daniell kernel becomes dominant. The tests $ST1_a$ and $ST2_a$ have roughly the same power as the truncated kernel-based M_a-test. (The truncated kernel-based M_a-test and $ST2_a$ have identical power because there is an exact relationship between them.)

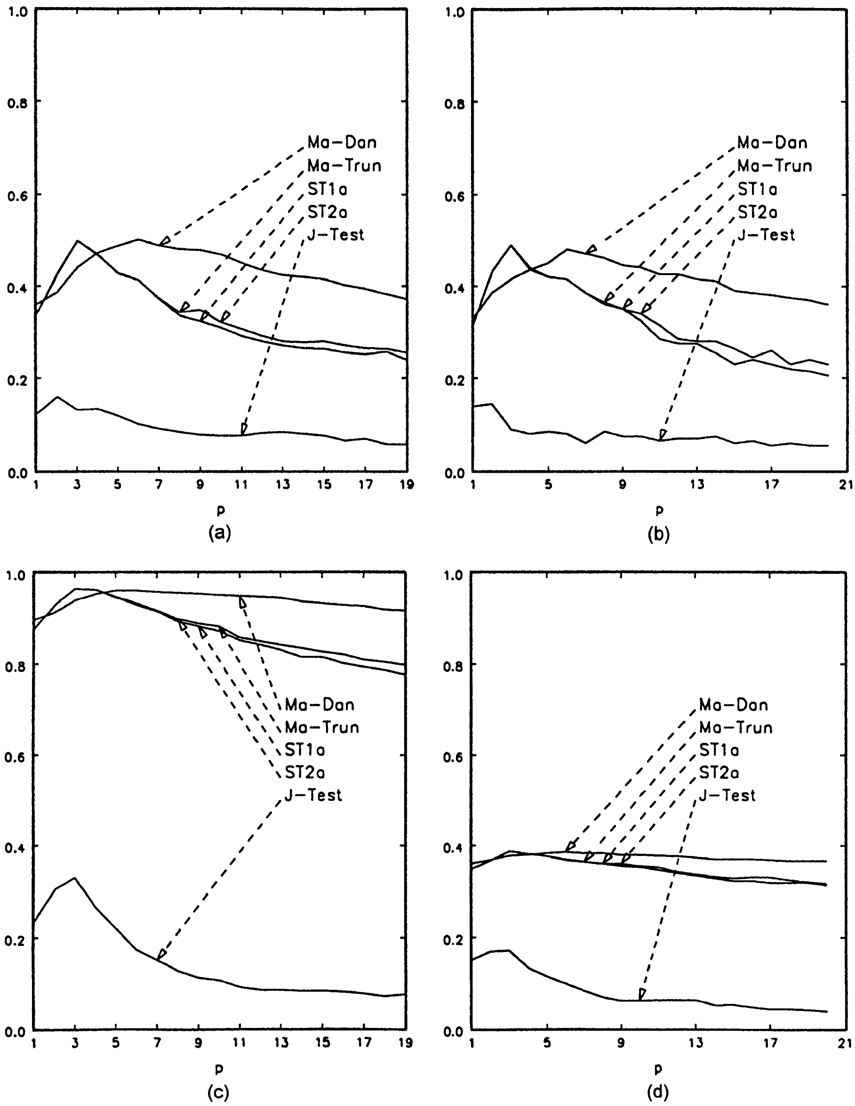

Fig. 6. Power at the 5% level under EXP(3) and EXP(10) processes: (a) size-corrected power, EXP(3), normal error; (b) bootstrap power, EXP(3), normal error; (c) size-corrected power, EXP(3), log-normal error; (d) bootstrap power, EXP(3), log-normal error; (e) size-corrected power, EXP(10), normal error; (f) bootstrap power, EXP(10), normal error; (g) size-corrected power, EXP(10), log-normal error; (h) bootstrap power, EXP(10), log-normal error

(b) For the alternatives AR(1), TAR(1), NAR(3), NAR(5), EXP(3) and EXP(10), $ST1_a$, $ST2_a$ and the truncated kernel-based M_a-test achieve maximal powers when p is equal to the largest significant lag included in the model. This is consistent with the findings of Skaug and Tjøstheim (1993a). For the ARFIMA(0, d, 0) process, $ST1_a$, $ST2_a$ and the truncated kernel-based M_a-test achieve maximal powers at $p = 1$ under normal innovations, but their powers grow with p under log-normal innovations. In contrast, the Daniell kernel generally achieves the maximal power when p is larger than the largest significant lag. This is true even for the AR(1) process. A possible reason is that

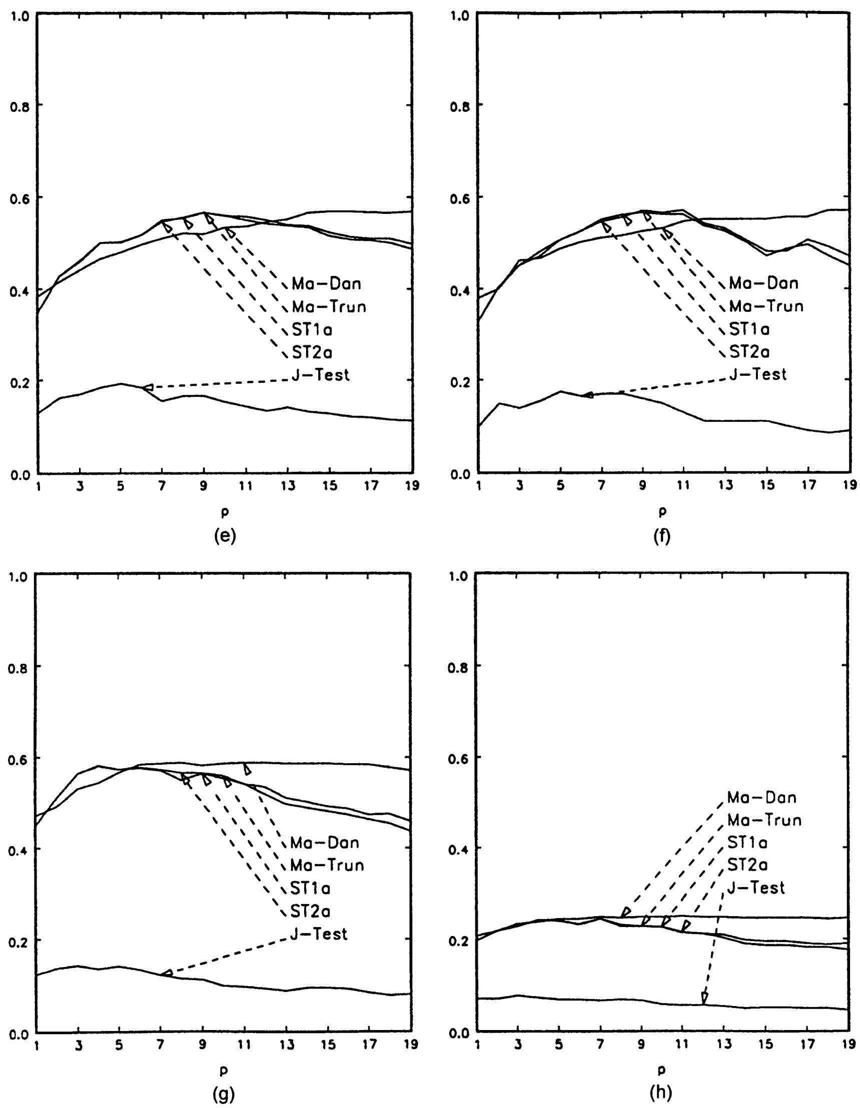

Fig. 6. (*continued*)

$D^2(j)$ may not vanish when j is larger than the largest significant lag, although its magnitude may be smaller. Thus, to include some extra terms beyond the largest significant lag may enhance the power if $D^2(j)$ is sufficiently large to offset the loss of additional degrees of freedom. For the ARFIMA(0, d, 0) process, the Daniell kernel has the maximal power when $p = 3$ under normal innovations, and its power also grows with p under log-normal innovations. Note that the Daniell kernel often makes power less sensitive to the choice of p, because it discounts higher order lags, which usually contribute less in power.

(c) None of the tests uniformly dominates the others against all the alternatives under study. The J-test is more powerful against ARCH(1), GARCH(1) and NMA processes with normal innovations, whereas the tests based on the empirical distribution function are more powerful than the J-test against the AR(1), ARFIMA(0, d, 0), TAR(1), EXP(3) and EXP(10) processes. For NAR(3), NAR(5) and NMA with log-normal innovations, the J-test is more powerful when p is small, whereas the tests based on the empirical distribution function are more powerful for large p. It seems that the tests based on the empirical distribution function are more powerful against the alternatives whose dependences are mostly captured by the conditional mean, whereas the J-test is more powerful against the alternatives whose dependences are mostly captured by the conditional second moments. Almost all the tests have better size-corrected powers under log-normal innovations than under normal innovations; only the J-test has better powers against the ARFIMA(0, d, 0) and ARCH(1) processes under normal innovations than under log-normal innovations.

(d) The bootstrap powers for all the tests are similar to those using empirical critical values under normal innovations. In this case, the loss of power due to bootstrapping is small. For log-normal innovations, however, the bootstrap powers are substantially smaller than the size-corrected powers in quite a few cases. An exception is the J-test against the ARFIMA(0, d, 0) process, whose bootstrap power is much better than its size-corrected power.

For $n = 200$, only size-corrected powers were considered. All the tests have better powers against all the alternatives to various degrees, but the relative rankings remain unchanged. The tests based on the empirical distribution function still have relatively low powers against the ARCH(1), GARCH(1, 1) and NMA processes. However, this does not mean that they should not be used to detect such alternatives. If interest indeed is in detecting these alternatives, it is sensible to apply the tests based on the empirical distribution function to the square of the original series. This can be expected to yield good power against ARCH-type alternatives.

Acknowledgements

I would like to thank two referees, the Joint Editor, N. Kiefer and T. Lyons for helpful comments and suggestions that led to a significant improvement of this paper.

Appendix A

We sketch the proof of theorems 1–3 here. Throughout, $0 < \Delta < \infty$ denotes a generic finite constant that may differ in different places.

A.1. Proof of theorem 1

We prove for M_a only; the proof for M_b is similar. The proof of theorem 1 is based on propositions 1–5 stated below. The proofs of these propositions are available from the author on request.

Proposition 1 shows that the weighted sum of von Mises statistics V_n in equation (4) can be approximated by a weighted sum of third-order V-statistics related to various lags.

Proposition 1. Put $h(x, y) = \mathbf{1}(x \leqslant y) - G(y)$. For $0 < j < n$, define

$$\hat{H}_n^2(j) = (n-j)^{-1} \sum_{t=j+1}^{n} \hat{H}_j^2(X_t, X_{t-j}),$$

where

$$\hat{H}_j(x, y) = (n-j)^{-1} \sum_{t=j+1}^{n} h(X_t, x)\, h(X_{t-j}, y).$$

Then

$$\sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\hat{D}_n^2(j) = \sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\hat{H}_n^2(j) + O_P(p/n^{1/2}).$$

The next result is a projection theory for dependent degenerate V-statistics.

Proposition 2. For $0 < j < n$, define

$$\tilde{H}_n^2(j) = \int \hat{H}_j^2(x, y)\,\mathrm{d}G(x)\,\mathrm{d}G(y),$$

where $\hat{H}_n^2(j)(x, y)$ is defined in proposition 1. Then

$$\sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\hat{H}_n^2(j) = \sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\tilde{H}_n^2(j) + O_P(p/n^{1/2}).$$

To state subsequent results, put

$$A_{ts}(j) = \int h(X_t, x)\,h(X_s, x)\,h(X_{t-j}, y)\,h(X_{s-j}, y)\,\mathrm{d}G(x)\,\mathrm{d}G(y).$$

Note that $A_{ts}(j) = A_{st}(j)$. In addition, $E\{A_{ts}(j)|\mathcal{F}_{t-1}\} = 0$ almost surely for all $t > s$ and all $j > 0$, where and hereafter $\{\mathcal{F}_t\}$ is the sequence of σ-fields consisting of $\{X_\tau, \tau \leqslant t\}$. We can now write

$$\sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\tilde{H}_n^2(j) = \sum_{j=1}^{n-1} k^2(j/p)(n-j)^{-1} \sum_{t=j+1}^{n} A_{tt}(j) + 2\sum_{j=1}^{n-2} k^2(j/p)(n-j)^{-1} \sum_{t=j+1}^{n} \sum_{s=j+1}^{t-1} A_{ts}(j)$$

$$= \hat{A}_n + \hat{B}_n, \text{ say.}$$

Proposition 3 shows that $\hat{A}_n$ can be approximated by a non-stochastic term.

Proposition 3. Let $\hat{A}_n$ be defined as above, and define

$$A_0 = \left[\int G(x)\,\{1 - G(x)\}\,\mathrm{d}G(x)\right]^2.$$

Then

$$\hat{A}_n - A_0 \sum_{j=1}^{n-1} k^2(j/p) = O_P(p/n^{1/2}).$$

The statistic $\hat{B}_n$ is a weighted sum of degenerate second-order U-statistics with mean 0. By rearranging the summation indices, we can write

$$\hat{B}_n = \sum_{t=3}^{n} \left\{ \sum_{s=2}^{t-1} \sum_{j=1}^{s-1} 2k^2(j/p)(n-j)^{-1}\,A_{ts}(j) \right\}.$$

Because $E\{A_{ts}(j)|\mathcal{F}_{t-1}\} = 0$ for all $t > s$ and all $j > 0$, $\hat{B}_n$ is a sum of a martingale difference sequence. By applying Brown's (1971) martingale limit theorem, we can show that a properly standardized version of $\hat{B}_n$ converges in distribution to $N(0, 1)$, as stated below.

Proposition 4. Define

$$B_0 = \left(\int [G\{\min(x, y)\} - G(x)\,G(y)]^2\,\mathrm{d}G(x)\,\mathrm{d}G(y)\right)^2.$$

Then

$$\left\{2B_0 \sum_{j=1}^{n-2} k^4(j/p)\right\}^{-1/2} \hat{B}_n \xrightarrow{\mathrm{d}} N(0, 1).$$

Both A_0 and B_0 must be estimated if we do not know whether $G(x)$ is continuous or discrete. The following result shows that $\hat{A}_0$ and $\hat{B}_0$ defined in theorem 1 are consistent for A_0 and B_0.

Proposition 5. Let $\hat{A}_0$ and $\hat{B}_0$ be defined as in theorem 1. Then

$$\hat{A}_0 - A_0 = O_P(n^{-1/2}),$$
$$\hat{B}_0 - B_0 = o_P(1).$$

Now, combining propositions 1–3 yields

$$\sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\hat{D}_n^2(j) = A_0 \sum_{j=1}^{n-1} k^2(j/p) + \hat{B}_n + O_P(p/n^{1/2})$$
$$= \hat{A}_0 \sum_{j=1}^{n-1} k^2(j/p) + \hat{B}_n + O_P(p/n^{1/2}),$$

where the second equality follows from proposition 5 and $p \to \infty$. It follows that

$$\left\{2B_0 \sum_{j=1}^{n-2} k^4(j/p)\right\}^{-1/2} \sum_{j=1}^{n-1} k^2(j/p)\{(n-j)\,\hat{D}_n^2(j) - \hat{A}_0\} = \left\{2B_0 \sum_{j=1}^{n-2} k^4(j/p)\right\}^{-1/2} \hat{B}_n + O_P(p^{1/2}/n^{1/2}) \xrightarrow{\mathrm{d}} N(0,\,1)$$

by proposition 4 and $p = cn^{\nu}$ for $0 < \nu < 1$, where we have also made use of the fact that

$$\sum_{j=1}^{n-2} k^4(j/p) = p \int_0^{\infty} k^4(z)\,\mathrm{d}z\,\{1 + o(1)\}.$$

By replacing B_0 by $\hat{B}_0$, we have

$$M_{\mathrm{a}} \xrightarrow{\mathrm{d}} N(0,\,1)$$

by Slutsky's theorem and $\hat{B}_0 - B_0 = o_P(1)$. This completes the proof for M_{a}. □

A.2. Proof of theorem 2

For a continuous distribution G, $G(X_t)$ is uniformly distributed on (0, 1). This fact suggests that we can compute A_0 and B_0 directly:

$$A_0 = \left\{\int_0^1 u(1-u)\,\mathrm{d}u\right\}^2 = \frac{1}{6^2}$$

and

$$B_0 = \left[\int_0^1 \int_0^1 \{\min(u_1,\,u_2) - u_1 u_2\}^2\,\mathrm{d}u_1\,\mathrm{d}u_2\right]^2 = \frac{1}{90^2}.$$

Hence, we need not use $\hat{A}_0$ and $\hat{B}_0$. The desired result follows immediately. □

A.3. Proof of theorem 3

The consistency of M_{a} follows from the fact that

$$\sum_{j=1}^{n-1} k^r(j/p) = p \int_0^{\infty} k^r(z)\,\mathrm{d}z\,\{1 + o(1)\}$$

for $r \geqslant 2$, $p = cn^{\nu}$ for $0 < \nu < \frac{1}{3}$, and two propositions stated below. The proofs of these propositions are available from the author on request.

Proposition 6. Put $D_j(x,\,y) = F_j(x,\,y) - G(x)\,G(y)$. Define

$$\tilde{D}_n^2(j) = (n-j)^{-1} \sum_{t=j+1}^{n} D_j^2(X_t,\,X_{t-j}).$$

Then

$$n^{-1} \sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\hat{D}_n^2(j) = n^{-1} \sum_{j=1}^{n-1} k^2(j/p)(n-j)\,\tilde{D}_n^2(j) + o_P(1).$$

Proposition 7. Let $D^2(j)$ be defined as in equation (1) and $\tilde{D}_n^2(j)$ be defined as in proposition 6. Then

$$n^{-1} \sum_{j=1}^{n-1} k^2(j/p)(n-j)\, \tilde{D}_n^2(j) = \sum_{j=1}^{\infty} D^2(j) + o_P(1).$$

References

Andrews, D. W. K. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.
Bahadur, R. R. (1960) Stochastic comparison of tests. *Ann. Math. Statist.*, **31**, 276–295.
Blum, J., Kiefer, J. and Rosenblatt, M. (1961) Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.*, **32**, 485–498.
Box, G. and Pierce, D. (1970) Distribution of residual autocorrelations in autoregressive integrated moving average time series models. *J. Am. Statist. Ass.*, **65**, 1509–1527.
Brown, B. M. (1971) Martingale central limit theorems. *Ann. Math. Statist.*, **42**, 59–66.
Carlstein, E. (1988) Degenerate *U*-statistics based on non-independent observations. *Calcutta Statist. Ass. Bull.*, **37**, 55–65.
Chan, N. H. and Tran, L. T. (1992) Nonparametric tests for serial dependence. *J. Time Ser. Anal.*, **13**, 102–113.
Davies, R. and Harte, D. (1987) Tests for Hurst effect. *Biometrika*, **74**, 95–101.
Deheuvels, P. (1981) An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multiv. Anal.*, **11**, 102–113.
Delgado, M. A. (1996) Testing serial independence using the sample distribution function. *J. Time Ser. Anal.*, **17**, 271–285.
Engle, R. (1982) Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.
Fama, E. F. and Roll, R. (1968) Some properties of symmetric stable distributions. *J. Am. Statist. Ass.*, **63**, 817–836.
Granger, C. and Teräsvirta, T. (1993) *Modelling Nonlinear Economic Relationships*. New York: Oxford University Press.
Hjellvik, V. and Tjøstheim, D. (1996) Nonparametric statistics for testing of linearity and serial independence. *J. Nonparam. Statist.*, **6**, 223–251.
Hoeffding, W. (1948) A nonparametric test of independence. *Ann. Math. Statist.*, **19**, 546–557.
Hong, Y. (1996a) Consistent testing for serial correlation of unknown form. *Econometrica*, **64**, 837–864.
———(1996b) Testing for independence between two covariance stationary time series. *Biometrika*, **83**, 615–625.
Ljung, G. M. and Box, G. E. P. (1978) A measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.
Mandelbrot, B. (1967) The variation of some other speculative prices. *J. Bus.*, **40**, 393–413.
Pinkse, J. (1997) Consistent nonparametric testing for serial independence. *J. Econometr.*, to be published.
Priestley, M. (1962) Basic considerations in the estimation of spectra. *Technometrics*, **4**, 551–564.
———(1981) *Spectral Analysis and Time Series*, vol. 1. London: Academic Press.
Robinson, P. M. (1991a) Consistent nonparametric entropy-based testing. *Rev. Econ. Stud.*, **58**, 437–453.
———(1991b) Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *J. Econometr.*, **47**, 67–84.
———(1994) Time series with strong dependence. In *Proc. 6th Wrld Congr. Advances in Econometrics* (ed. C. Sims), vol. 1, pp. 47–95. Cambridge: Cambridge University Press.
Sen, P. K. (1963) On the properties of U-statistics when the observations are not independent. *Calcutta Statist. Ass. Bull.*, **12**, 69–92.
Skaug, H. J. and Tjøstheim, D. (1993a) Nonparametric test of serial independence based on the empirical distribution function. *Biometrika*, **80**, 591–602.
———(1993b) Nonparametric tests of serial independence. In *Developments in Time Series Analysis: the Priestley Birthday Volume* (ed. T. Subba Rao), pp. 207–229. London: Chapman and Hall.
———(1996) Measures of distance between densities with application to testing for serial independence. In *Time Series Analysis in Memory of E. J. Hannan* (eds P. M. Robinson and M. Rosenblatt), pp. 363–377. New York: Springer.