

A Versatile and Robust Metric Entropy Test of Time-Reversibility, and other Hypotheses*

Jeff Racine
Department of Economics
McMaster University
Hamilton, ON L8S 4M4, Canada
racinej@mcmaster.ca

Esfandiar Maasoumi
Department of Economics
Southern Methodist University
Dallas, TX 75275-0496, USA
maasoumi@mail.smu.edu

May 6, 2005

Abstract

We examine the performance of a metric entropy statistic as a robust test for time-reversibility (TR), symmetry, and serial dependence. It also serves as a measure of goodness-of-fit. The statistic provides a consistent and unified basis in model search, and is a powerful diagnostic measure with surprising ability to pinpoint areas of model failure. We provide empirical evidence comparing the performance of the proposed procedure with some of the modern competitors in nonlinear time-series analysis, such as robust implementations of the BDS and characteristic function-based tests of TR, along with correlation-based competitors such as the Ljung-Box Q -statistic. Unlike our procedure, each of its competitors is motivated for a different, specific, context and hypothesis. Our evidence is based on Monte Carlo simulations along with an application to several stock indices for the US equity market.

*The authors would like to thank Dee Dechert, Thomas Cover, Yi-Ting Chen, Adrian Pagan, and participants at the IEE Conference in Honor of Arnold Zellner (September 2003, American University) and the Canadian Econometric Study group (September 2004) for useful comments and suggestions. The usual caveat applies.

1 Introduction

Since at least Tong (1990) there has been a greater appreciation that the characteristics of economic and financial time series typically go far beyond serial correlation and volatility clustering. For instance, the unemployment rate may behave asymmetrically in expansions and recessions, and the volatility of stock returns seems more sensitive to negative news than good news. Business cycle asymmetry and volatility asymmetries influence economic theories and the design of empirical models. As Tong (1990) pointed out, time-irreversibility is a broad concept of asymmetries in time-series, and we propose here a unified approach to testing time-irreversibility to complement conventional independence and other tests. We follow the same principles and statistics to test several hypotheses, as well as assess model fit and predictive performance. This is of value since testing is useful in a decision setting where models are adopted or rejected based on their performance.

A strictly stationary process is time-reversible (TR) if its finite dimensional distributions are invariant to the reversal of the time indices. Linear processes with non-Gaussian innovations (see Weiss (1975), Tong (1990), and Hallin, Lefevre & Puri (1988)) and nonlinear processes with regime-switching structures, such as the Self-Exciting Threshold Autoregressive (SETAR) processes, the exponential GARCH (EGARCH) processes, and the Smooth Transition Autoregressive (STAR) processes, are examples of generally time-irreversible processes. Note that, since the finite dimensional distributions of a sequence of independently and identically distributed (*iid*) random variables are products of marginal distributions, time-reversibility is a necessary condition for an *iid* random variable.

In motivating a test for serial independence against time-irreversibility, Chen (2003) writes “Testing serial independence against time-irreversibility would be important for motivating and checking these nonlinear models, just like testing serial independence against serial correlation and volatility clustering is important for the ARMA-GARCH models.” Chen (2003) provides a good discussion of competing tests for TR, such as the one proposed by Chen, Chou & Kuan (2000) which is based on the characteristic function. Like our proposed test in this paper, an advantage of the Chen et al. (2000) test is its robustness to the moment condition failure of heavily-tailed data. Chen (2003) uses the expectation of an odd-symmetric function of a random variable and its lag as a time-irreversibility measure. This is different from the basis used by Chen et al. (2000); see Section 2.2 for details.

Researchers depend on a battery of specification and diagnostic tests designed to guide the specification process. Most of these statistics trace their origins to linear time-series models, though they are often used in nonlinear settings with varying degrees of success. The best known examples are correlation-based statistics such as the autocorrelation function and the Ljung-Box Q statistic (Ljung & Box (1978)), though of course information criteria such as the Akaike Information Criterion (Akaike (1981)) have also proved to be very popular. A complementary set of statis-

tics trace their origins to nonlinear time-series models. This class would include, for example, the chaos-based BDS statistic (Brock, Dechert & Scheinkman (1987), Brock, Dechert, Scheinkman & LeBaron (1996)) and, more recently, the TR tests mentioned in the previous paragraph and in Chen & Kuan (2002) and Chen (2003).

But, specification and diagnostic tests are more meaningfully placed in the context of decision making whereby a model's adequacy is ranked and judged by goodness-of-fit and predictive performance. It is typically the case that principles and innovations that guide the choice of test statistics change from one test to another, and are also unrelated to principles and methods used to assess model choice, fit, and prediction. For instance, one would not think of using, say, a Q statistic to measure goodness-of-fit! As will become clear, an advantage of the entropy methods of this paper, and similar "distribution based" procedures, is that they avoid confusing and sometimes contradictory sets of principles and motivations.

Barnett, Gallant, Hinich, Jungeilges, Kaplan & Jensen (1997) have studied the performance of a range of popular diagnostic statistics, and they outline the generally unsatisfactory performance of several of these approaches to testing for nonlinearity. Not even those tests having their origins in nonlinear settings are immune to performance problems, however. For instance, Cromwell, Labys & Terraza (1994, pg 32–36) outline the use of the BDS statistic as a diagnostic tool for linear time-series modelling, the approach motivated mainly through expectation of power against linear, nonlinear, and chaotic (deterministic) alternatives. Unfortunately, the usual application of the BDS statistic in most studies is nonrobust due to the presence of unacceptably large size distortions, reflecting a failure of the asymptotic distribution theory in finite-sample settings. Somewhat surprisingly, the use of resampling methods to correct for such size distortions reveals an underlying lack of power relative to alternative statistics; see Belaire-Franch & Contreras (2002) for size and power performance of a permutation-based BDS test, and Chen & Kuan (2002) who consider a bootstrap-based BDS test. Time-reversibility can be shown to be a necessary condition for serially independent processes, thus the TR statistic can be directly applied as a diagnostic tool. However, there is a breakdown in the asymptotic distribution theory in finite-sample settings similar to that found for the BDS test, while a resampled version of the TR test may lack power in other settings.

Information theoretic tests are increasingly found to be superior in a variety of contexts; see Hong & White (forthcoming), Granger, Maasoumi & Racine (2004), and Skaug & Tjøstheim (1993), among others. In fact, BDS and other correlation integrals too may be viewed as special approximations of certain mutual information measures. Indeed, such relations may be used to obtain alternative nonparametric estimates for entropy measures, as proposed by Diks & Manzan (2002). It appears that how entropy measures are approximated, as well as how the actual statistics are implemented, have far reaching consequences for test properties and performance.

In this paper we pursue an information theoretic approach to the general problem of testing and model selection. We adopt Granger et al.'s (2004) metric entropy statistic and illustrate how it

can serve as a versatile diagnostic tool for guiding model specification in several new directions. In addition to being well-suited to testing for serial dependence as described in Granger et al. (2004) and measuring goodness-of-fit in nonlinear models as described in Maasoumi & Racine (2002), we propose using this statistic to test for time-reversibility. Our particular formulation of the null of time-reversibility necessitates testing for “symmetry.” The versatility of the entropy approach allows us, however, to use it to also test for symmetry in other contexts. We will examine the performance of the entropy based method relative to a “robust BDS” test, the TR test of Chen & Kuan (2002), and common correlation based statistics such as the Q statistic. Our approach to testing for time-reversibility appears to have good performance when used to identify suitable models and lags, being correctly sized yet having improved power relative to competing approaches. We also underscore the prescriptive nature of the statistic. That is, should the statistic indicate model failure, it also identifies a likely culprit for this failure thereby suggesting a direction in which an improved model may lie.

The rest of the paper proceeds as follows. Section 2 presents an overview of the proposed test of symmetry and time-reversibility, along with the comparison tests considered herein. Section 3 outlines a modest Monte Carlo experiment designed to examine finite-sample size and power of the proposed test of symmetry and time-reversibility, and provides the asymptotic distribution of the statistic. Section 4 present results extending the application performed by Chen & Kuan (2002) on financial models of six major US stock indices, while Section 5 presents some concluding remarks.

2 Overview of the BDS, TR, and Metric Entropy Test Statistics

We briefly describe the tests which are compared in the current paper, the BDS test (Brock et al. (1987), Brock et al. (1996)), the TR test of Chen & Kuan (2002), and the entropy-based test S_ρ (Granger et al. (2004), Maasoumi & Racine (2002)). We refer interested readers to the original papers for detailed descriptions of size and power performances of the respective tests.

2.1 The BDS Test

The BDS test statistic is based on the correlation integral of a time-series $\{Y_t\}_{t=1}^T$. The generalized K -order correlation integral is given by:

$$C_K(Y, \epsilon) = \left[\int \left(\int I(\|y - y'\| \leq \epsilon) f_Y(y') dy' \right)^{K-1} f_Y(y) dy \right]^{\frac{1}{K-1}},$$

where $I(\cdot)$ denotes the indicator function, $f_Y(\cdot)$ denotes the marginal density of Y , and $\|Y\| = \sup_{i=1, \dots, \dim Y} |y_i|$, the sup norm. The distance parameter ϵ is like a bandwidth and behaves accordingly. When the elements of Y are *iid*, the correlation integral factorizes. The BDS test

statistic is based on C_K , $K = 2$. This gives the expected probability of ϵ -neighbourhoods.

For small ϵ and dimensionality parameter m , the inner integral (probability) in $C_K(\cdot)$ behaves as $\epsilon^m f_Y(y)$ over the ϵ -neighbourhood. This allows us to see an approximate relationship between the correlation integral and various entropies.

The test's finite-sample distribution has been found to be poorly approximated by its limiting $N(0, 1)$ distribution. In particular, the asymptotic-based test has been found to suffer from substantial size distortions, often rejecting the null 100% of the time *when the null is in fact true*. Recently, tables providing quantiles of the finite-sample distribution have been constructed in certain cases which attempt to correct for finite-sample size distortions arising from the use of the asymptotic distribution (see Kanzler (1999) who assumed true Gaussian error distributions), though the asymptotic version of the test is that found in virtually all applied settings. A number of authors have noted that its finite-sample distribution is sensitive to the embedding dimension, dimension distance, and sample size, thus tabulated values are not likely to be useful in applied settings.¹ However, a simple permutation-based resampled version of the BDS statistic does yield a correctly sized test (Belaire-Franch & Contreras (2002), Diks & Manzan (2002)), hence we elect to use this ‘‘robust BDS’’ approach implemented by Chen & Kuan (2002) for what follows.

2.2 The TR Test

Recently, Chen & Kuan (2002) have suggested using a modified version of the TR test of Chen et al. (2000) as a diagnostic test for time-series models. This is a characteristic function-based test, and its authors recommend it in part as it requires no moment conditions hence is of wider applicability than existing time-reversibility tests. A stationary process is said to be ‘time-reversible’ if its distributions are invariant to the reversal of time indices; independent processes are time-reversible. If time-reversibility does not hold (i.e., the series is ‘time-irreversible’), then there is asymmetric dependence among members of the series in the sense that the effect of, say, Y_s on Y_t is different from that of Y_t on Y_s ; the threshold autoregressive (TAR) model is one example of a time-irreversible series.

When a series is time-reversible then the distribution of $Y_t - Y_{t-k}$ is symmetric ($k = 1, 2, \dots$), while failure of this symmetry condition indicates asymmetric dependence. A distribution is symmetric if and only if the imaginary part of its characteristic function is zero (i.e., $h(\omega) = E[\sin(\omega(Y_t - Y_{t-k}))] = 0 \quad \forall \quad \omega \in \mathbb{R}^+$). A necessary condition is

$$E[\psi_g(Y_t - Y_{t-k})] = E \left[\int_{\mathbb{R}^+} \sin(\omega(Y_t - Y_{t-k}))g(\omega) d\omega \right] = 0, \quad (1)$$

¹We note that in applied settings the user is required to set the embedding dimension (m) and the size of the dimensional distance (ϵ). One often encounters advice to avoid using the test on samples of size 500 or smaller, while one also encounters advice on setting ϵ in the range $0.5\sigma_y$ to $2.0\sigma_y$ of a time-series $\{Y_t\}_{t=1}^T$ along with advice on setting m in the range 2 – 8.

where $g(\cdot)$ is a weighting function, and Chen et al. (2000) therefore propose a test based on the sample analog of (1) given by

$$C_{g,k} = \sqrt{T_k} \left(\frac{\bar{\psi}_{g,k}}{\bar{\sigma}_{g,k}} \right),$$

where $T_k = T - k$, $\bar{\psi}_{g,k} = \sum_{t=k+1}^T \psi_g(Y_t - Y_{t-k})/T_k$, and $\bar{\sigma}_{g,k}^2$ is a consistent estimator of the asymptotic variance of $\sqrt{T_k} \bar{\psi}_{g,k}$. Chen et al. (2000) choose $g(\cdot)$ to be the exponential distribution function with parameter $\beta > 0$ yielding a test that is straightforward to compute, and under H_0 , it can be shown that $C_{g,k}$ has a limiting $N(0, 1)$ distribution. However, this version of the test suffers from substantial size distortions. A modified version of this test appropriate for testing asymmetry of residuals arising from a time-series model which is called the TR test (Chen & Kuan (2002)) is given by

$$\hat{C}_{g,k} = \sqrt{T_k} \left(\frac{\hat{\psi}_{g,k}}{\hat{\nu}_{g,k}} \right),$$

where $\hat{\psi}_{g,k} = \sum_{t=k+1}^T \psi_g(\hat{\epsilon}_t - \hat{\epsilon}_{t-k})/T_k$ (for our purposes $\hat{\epsilon}_t$ represents standardized residuals from a time-series model) and where $\hat{\nu}_{g,k}^2$ is a consistent estimator of the asymptotic variance of $\sqrt{T_k} \hat{\psi}_{g,k}$ which is obtained by bootstrapping (Chen & Kuan (2002, pg 568)).

2.3 Metric Entropy Tests

Entropy-based measures of divergence have proved to be powerful tools for a variety of tasks. They may form the basis for measures of nonlinear dependence, as in Granger et al. (2004), but as we shall demonstrate can also be used to assess goodness-of-fit in nonlinear models or form the basis of tests for symmetry or time-reversibility.

Granger et al. (2004) consider the case of the K -Class entropy, as in:

$$H_K(Y) = \frac{1}{K-1} \left[1 - \int f_Y^{K-1} f_Y(y) dy \right], \quad K \neq 1$$

(equal to Shannon's entropy for $K = 1$),

$$\simeq \frac{1}{K-1} \left\{ 1 - [C_K(Y; \epsilon)/\epsilon^m]^{K-1} \right\},$$

or, for Renyi's entropy:

$$H_q(Y) = \frac{1}{q-1} \log \int [f_Y(y)]^{q-1} f_Y(y) dy, \quad q \neq 1,$$

$$\simeq -\log C_q(Y; \epsilon) + m \log \epsilon.$$

A nonparametric estimate of $C_q(\cdot)$ may then be used to obtain a similar estimate of $H_q(\cdot)$, as

cogently argued by Diks & Manzan (2002).

Mutual Information measures test the significance of different metrics of divergence between the joint distribution and the product of the marginals. For instance:

$$I(X, Y) = \int \int \ln(f_{X,Y}(x, y)/f_X(x)f_Y(y)) \times f_{X,Y}(x, y) dx dy$$

is the simple Shannon mutual information. And, generally

$$I_q(X, Y) = H_q(X) + H_q(Y) - H_q(X, Y),$$

for any q , relates mutual information to the respective entropies. The conditional mutual information, given a third variable Z , is:

$$I(X, Y|Z) = \int \int \int \ln(f_{X|y,z}(x|y, z)/f_{X|z}(x|z))f_{X,Y,Z}(x, y, z) dx dy dz.$$

More generally,

$$\hat{I}_q(X, Y|Z) = \ln \hat{C}_q(X, Y, Z; \epsilon) - \ln \hat{C}_q(X, Z; \epsilon) - \ln \hat{C}_q(Y, Z; \epsilon) + \ln \hat{C}_q(Z; \epsilon)$$

reveals the relation between conditional mutual information and the correlation integral as the basis for its nonparametric estimation. The unconditional relation is obtained by removing Z . Extensive results on the connection between correlation integrals and information theory are given in Prichard & Theiler (1995).²

Following the arguments in Granger et al. (2004) and Maasoumi & Racine (2002) in favour of metric entropies, we choose $K = 1/2$ in the K -class entropy defined above. This is a normalization of the Bhattacharya-Matusita-Hellinger measure of dependence given by

$$\begin{aligned} S_\rho &= \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(f_1^{\frac{1}{2}} - f_2^{\frac{1}{2}} \right)^2 dx dy \\ &= \frac{1}{4} I_{\frac{1}{2}}, \end{aligned} \tag{2}$$

where $f_1 = f(x, y)$ is the joint density and $f_2 = f(x) \cdot f(y)$ is the product of the marginal densities

²The choice $q = 2$ is by far the most popular in chaos analysis as it allows for efficient estimation algorithms. Note that the conditional mutual information $I_q(X, Y|Z)$ is not positive definite for $q \neq 1$. It is thus possible to have variables X and Y which are conditionally dependent given Z , but for which $I_2(X, Y|Z)$ is zero or negative. Also, as noted by Diks & Manzan (2002), if $I_2(X, Y|Z)$ is zero, the test based on it does not have unit power asymptotically against conditional dependence. This situation, while exceptional, is quite undesirable. Since $I_2(X, Y|Z)$ is usually either positive or negative, a one-sided test rejecting for $I_2(X, Y|Z)$ large, is not optimal. Diks & Manzan (2002) argue that, in practise I_2 behaves much like I_1 in that we usually observe larger power for one-sided tests (rejecting for large I_2) than for two-sided tests. This led them to propose $q = 2$, together with a one-sided implementation of the test.

of the random variables X and Y . $S_\rho = 0$ if and only if X and Y are independent, and is otherwise positive and less than one.

This produces a test statistic that satisfies the following properties: (i) it is well defined for both continuous and discrete variables, (ii) it is normalized to zero if X and Y are independent, and lies between 0 and 1, (iii) the modulus of the measure is equal to unity if there is a *measurable* exact (nonlinear) relationship, $Y = g(X)$ say, between the random variables, (iv) it is equal to or has a simple relationship with the (linear) correlation coefficient in the case of a bivariate normal distribution, (v) it is a metric, that is, it is a true measure of “distance” and not just of divergence, and (vi) the measure is invariant under continuous and strictly increasing transformations $h(\cdot)$. This is useful since X and Y are independent if and only if $h(X)$ and $h(Y)$ are independent. This invariance is also useful in deriving asymptotic distributions and conducting resampling experiments on pivotal statistics.

Our TR test relies on testing for symmetry. To test symmetry, either in a series (raw or residuals), or in their lag differences (as for time-reversibility case), the densities f_1 and f_2 will be defined accordingly. This demonstrates the unification ability and general versatility of the information/entropy measures. The connection is worth summarizing. For any two density functions f_1 and f_2 , the asymmetric (with respect to f_1) K -class entropy *divergence measure* is:

$$I_k(f_2, f_1) = \frac{1}{k-1} \left[\int 1 - (f_2^k/f_1^k) dF_1 \right], \quad k \neq 1,$$

such that $\lim_{k \rightarrow 1} I_k(\cdot) = I_1(\cdot)$, the Shannon cross entropy (divergence) measure. When one distribution is the joint, and the other is the product of the marginals, this latter measure is called the “mutual information” outlined earlier. Once the divergence in both directions (of f_1 and f_2) are added, a symmetric measure is obtained which, for $K = 1$, is well known as the Kullback-Leibler measure. The symmetric K -class measure at $K = 1/2$ is as follows: $I_{1/2} = I_{1/2}(f_2, f_1) + I_{1/2}(f_1, f_2) = 2M(f_1, f_2) = 4B(f_1, f_2)$, where $M(\cdot) = \int (f_1^{1/2} - f_2^{1/2})^2 dx$ is the Matusita *distance*, and, $B(\cdot) = 1 - \rho^*$ is the Bhattacharya *distance* with $0 \leq \rho^* = \int (f_1 f_2)^{1/2} \leq 1$ being a measure of “affinity” between the two densities. $B(\cdot)$ and $M(\cdot)$ are rather rare measures of divergence since they satisfy the triangular inequality and are, therefore, proper measures of *distance*. Other *divergence* measures are capable of characterizing desired null hypotheses (such as independence) but may not be appropriate when these distances are compared across models, sample periods, or agents. These comparisons are often implicit in inferences.

The S_ρ measure is “robust” since it is capable of producing familiar outcomes when we deal with truly linear/Gaussian processes but is robust to departures from such settings. For testing the null of serial independence, we employ kernel estimators of $f(y, y_{-j})$, $f(y)$, and $f(y_{-j})$, $j = 1, 2, \dots, K$ originally proposed by Parzen (1962), with likelihood cross-validation used for bandwidth selection leading to density estimators that are “optimal” according to the Kullback-Leibler criterion; see

Silverman (1986, page 52) for details. For a test of *symmetry*, we proceed in a similar manner. The null distribution of the kernel-based implementation of S_ρ is obtained via a resampling approach identical to that used for the “robust BDS” test described above (see Granger et al. (2004) for details). R Code for computing the entropy metric and for computing the “robust BDS” test (Ihaka & Gentleman (1996)³) is available from the authors upon request.

In Section 4 we compare the relative performance of the TR, “robust BDS,” and the S_ρ entropy-based tests on the basis of their diagnostic ability in an empirical setting. All tests are correctly-sized due to their reliance on resampling procedures (see Section 3 for finite-sample performance of the entropy tests), therefore relative performance boils down to a comparison of power. We note that it is known that BDS is a test of the hypothesis $E[f(X, Y)] - E[f(X)] \times E[f(Y)] = 0$, whereas mutual information tests are concerned with the expected divergence between $f(x, y)$ and $f(x) \times f(y)$ (relative to $f(x, y)$). The latter are one-to-one representations of independence and imply the null of concern to BDS, but not vice versa.

2.3.1 Testing for Unconditional and Conditional Symmetry

Consider a stationary series $\{Y_t\}_{t=1}^T$. Let $\mu_y = E[Y_t]$, let $f_1(y)$ denote the density function of the random variable Y_t , let $\tilde{Y}_t = -Y_t + 2\mu_y$ denote a rotation of Y_t about its mean, and let $f_2(\tilde{y})$ denote the density function of the random variable \tilde{Y}_t . Note that, if $\mu_y = 0$, then $\tilde{Y}_t = -Y_t$, though in general this will not be so.

We say a series is *symmetric about the mean* (median, mode) if $f_1(y) \equiv f_2(\tilde{y})$ almost surely. Testing for asymmetry about the mean therefore naturally involve testing the null:

$$H_0 : f_1(y) = f_2(\tilde{y}) \text{ for all } y.$$

Note that symmetry about the mode or median may be a more natural characterization. One could of course clearly rotate a distribution around these measures of central tendency, and in what follows one would replace the mean with the appropriate statistic.

Tests for the presence of “conditional asymmetry” can be based upon standardized residuals from a regression model (see Belaire-Franch & Peiro (2003)). Let

$$Y_t = h(\Omega_t, \beta) + \sigma(\Omega, \lambda)e_t,$$

denote a general model for this process, where Ω_t is a conditioning information set, $\sigma(\Omega, \lambda)$ the conditional standard deviation of Y_t , and e_t is a zero mean unit variance error process independent of the elements of Ω_t . If $\mu_e = 0$, then tests for conditional asymmetry involve the following null:

$$H_0 : f_1(e) = f_2(-e) \text{ for all } e.$$

³See <http://www.r-project.org>

Bai & Ng (2001) construct tests based on the empirical distribution of e_t and that of $-e_t$. Belaire-Franch & Peiro (2003) apply this and other tests to the Nelson & Plosser (1982) data updated to include 1988.

We propose testing for symmetry using the metric entropy given in (2) with f_1 and f_2 as defined above. The unknown densities are estimated using kernel methods (see Granger et al. (2004) for details), while the null distribution is obtained via resampling from the pooled sample $\{Y_t, \tilde{Y}_t\}$. Note that H_0 holds if and only if $S_\rho = 0$. This is the main source of superior power for our entropy tests. Section 3 presents Monte Carlo evidence on the test's finite-sample performance, and reveals that the proposed approach is correctly sized and has power under the alternative.

2.3.2 Testing for Time-Reversibility

There are a number of potential approaches to testing for time-reversibility. For instance, one approach would be to consider the equality of the joint distribution of any *finite set* of time-series with the joint distribution of the same set in reverse order. In a nonparametric implementation, however, this can face the curse of dimensionality when the length of the finite set is large. Another approach is to note that reversibility implies that, $\forall k, f(Y_t, Y_{t-k}) = f(Y_{t-k}, Y_t)$, and Darolles, Florens & Gouriéroux (forthcoming) consider an elegant statistic based upon kernel estimation of such bivariate distributions at any lag. Yet another approach is possible, however, by recalling that when a series is time-reversible then $\forall k, f(Y_t - Y_{t-k})$ is symmetric. That is, one could instead look at the symmetry of *pairwise* distributions of $Y_t - Y_{t-k}$, for several values of k . In a nonparametric setting this circumvents the curse of dimensionality which has obvious appeal, and it is this avenue that we choose to pursue here. In fact, looking at the symmetry of pairwise distributions of $Y_t - Y_{t-k}$ for several values of k was also the route taken by Chen et al. (2000), who utilized the characteristic function as the basis for their test. Chen (2003) also proposed a portmanteau version of the latter method based on the sum of a finite number of pairwise differences with good performance. By considering testing for symmetry of the marginal distribution of k th differences, we highlight the versatility of the metric entropy by leveraging the symmetry test outlined in Section 2.3.1 in order to test time-reversibility, i.e., both tests are based on the same statistic, the former for levels, the latter for k th differences. Section 3 presents some Monte Carlo evidence on the test's finite-sample performance for some popular linear and nonlinear time-series processes. In this section we also offer some remarks regarding the asymptotic distribution of the entropy statistics and their poor performance. This has led us to focus on resampling techniques.

3 Finite-Sample Behaviour

3.1 Testing for Symmetry

We consider the finite-sample performance of the kernel-based test for symmetry, the null hypothesis being that the distribution is symmetric. We set the number of bootstrap replications underlying the test to 99, and consider a range of sample sizes. For each DGP, we conduct 1,000 Monte Carlo replications. The bandwidth is selected via likelihood cross-validation. We consider a range of DGP's ranging from symmetric to highly asymmetric. Table 1 summarizes the finite-sample performance of the proposed test in the form of empirical rejection frequencies, i.e., the proportion of rejections out of 1,000 Monte Carlo replications. We consider tests with nominal size $\alpha = 0.10$, 0.05, and 0.01.

Table 1: Empirical rejection frequencies at levels $\alpha = 0.10, 0.05, 0.01$. The degree of asymmetry increases as we go from left to right. The $N(\mu, \sigma^2)$ column corresponds to empirical size, while the remaining columns correspond to empirical power. n denotes the sample size.

n	$N(\mu, \sigma^2)$	$\chi^2(120)$	$\chi^2(80)$	$\chi^2(40)$	$\chi^2(20)$	$\chi^2(10)$	$\chi^2(5)$	$\chi^2(1)$
$\alpha = 0.10$								
50	0.101	0.155	0.195	0.285	0.414	0.611	0.845	0.957
100	0.116	0.247	0.309	0.446	0.697	0.899	0.991	1.000
200	0.097	0.340	0.461	0.737	0.939	0.997	1.000	1.000
$\alpha = 0.05$								
50	0.062	0.093	0.124	0.166	0.252	0.456	0.690	0.881
100	0.071	0.152	0.185	0.316	0.551	0.806	0.966	0.993
200	0.046	0.230	0.325	0.611	0.885	0.994	1.000	1.000
$\alpha = 0.01$								
50	0.018	0.027	0.038	0.059	0.108	0.229	0.444	0.666
100	0.025	0.058	0.067	0.144	0.323	0.600	0.863	0.956
200	0.011	0.112	0.171	0.398	0.724	0.956	0.999	0.998

We observe from Table 1 that the test is correctly sized. Empirical size does not differ from nominal for any of the sample sizes considered. As the degree of asymmetry increases, we observe that power increases as it also does when the sample size increases.

3.2 Testing for Time-Reversibility

Next, we consider two simulation experiments, one designed to examine finite-sample power and the other to examine finite-sample size of the proposed metric entropy test of time-reversibility.

In order to examine the test's power, we consider a time-irreversible two-regime self-exciting

threshold autoregression model (SETAR) of the form

$$Y_t = (1 - I(Y_{t-d} > r))(\alpha_0 + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + \sigma_1 \epsilon_t) + I(Y_{t-d} > r)(\beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \sigma_2 \epsilon_t) \quad (3)$$

where $\epsilon \sim N(0, 1)$, and where $I(\cdot)$ is the familiar indicator function. This is commonly referred to as a SETAR(2; p, p), and serves as a common illustration of a time-irreversible process. Following Clements, Franses, Smith & Van Dijk (2003, pg 366) we set $p = 1$, $d = 1$, and $r = 0.15$, and let $(\alpha_0, \alpha_1, \sigma_1) = (-1.25, -0.7, 2)$ and $(\beta_0, \beta_1, \sigma_2) = (0, 0.3, 1)$.

We consider 1,000 Monte Carlo replications drawn from this DGP and apply the proposed test for a range of lags for $Y_t - Y_{t-k}$. Results are presented in Table 2. The null is that the series is time-reversible, hence empirical rejection frequencies exceeding 0.10, 0.05, and 0.01 respectively reflect power at the respective sizes.

Table 2: Empirical rejection frequencies for the SETAR model for $k = 1, 2, 3$.

k	n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	50	0.728	0.532	0.217
1	100	0.926	0.854	0.549
1	200	0.999	0.997	0.934
2	50	0.129	0.062	0.007
2	100	0.154	0.081	0.016
2	200	0.212	0.130	0.035
3	50	0.080	0.033	0.007
3	100	0.073	0.033	0.004
3	200	0.071	0.031	0.003

In order to examine the test's size, Table 3 presents results for a time-reversible AR(1) model of the form $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \epsilon_t$, with $(\alpha_0, \alpha_1) = (1, 0.5)$ and $\epsilon_t \sim N(0, 1)$.

It is evident from tables 2 and 3 that the test is correctly sized (entries in Table 3 do not differ significantly from nominal size) and has power approaching 1 as n increases (see, e.g., entries in Table 2 for $k = 1$ approach 1 as n increases). We have considered a range of null and alternative models not reported here for space considerations, and results appear to be robust across a range of specifications. We are confident that the test will perform as expected in applied settings.

3.3 Asymptotic Approximations to the Null Distribution

Skaug & Tjøstheim (1996) obtained the asymptotic distributions of several similar statistics using

Table 3: Empirical rejection frequencies for the AR(1) model for $k = 1, 2, 3$.

k	n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
1	50	0.081	0.037	0.007
1	100	0.105	0.053	0.005
1	200	0.107	0.048	0.007
2	50	0.100	0.039	0.007
2	100	0.109	0.053	0.014
2	200	0.105	0.052	0.010
3	50	0.092	0.044	0.005
3	100	0.083	0.043	0.013
3	200	0.097	0.046	0.006

the same sample splitting techniques used by Robinson (1991).

Considering the moment form of our measure, and replacing the joint CDF with the empirical CDF, we obtain the following approximation:

$$\tilde{S}_\rho = \frac{1}{n-j} \sum_{t=j+1}^n \left[1 - \sqrt{\frac{\hat{g}(y_t)\hat{h}(y_{t-j})}{\hat{f}_1(y_t, y_{t-j})}} \right]^2 w(y_t, y_{t-j}). \quad (4)$$

The weight function $w(x, y) = 1\{(x, y) \in S_2\} = S \times S$, with $S = [a, b]$ for $a < b$. It has the effect of trimming out the extreme observations (though we don't actually use it in applications), and avoids the difficult tail areas for logarithmic entropy functions. Skaug & Tjøstheim (1993) and Skaug & Tjøstheim (1996) prove the consistency of a portmanteau version of \tilde{S}_ρ (for a joint test of serial independence at several lags) for a weighted and non-normalized version of S_ρ . Given that we are testing ‘‘symmetry’’ of the first differences which are *iid* under the null hypothesis, we are able to adopt the same techniques if we further assume the following:

- A1.** The marginal densities are bounded and uniformly continuous on \mathbb{R}^1 .
- A2.** The kernel function $K(\cdot)$ is bounded and satisfies: $\int K(u)du = 0$, $\int u^2K(u)du < \infty$, and has the representation $K(x) = \int \tilde{K}(\eta)e^{i\eta x}d\eta$ where $i = \sqrt{-1}$ and \tilde{K} is an absolutely integrable function of a real variable.
- A3.** The bandwidth $h_n = cn^{-1/\beta}$ for some $c > 0$ and $4 < \beta < 8$.

It is generally assumed that the processes are strong mixing with an exponentially decaying mixing coefficient. Under these assumptions proof of consistency of classes of measures, including \tilde{S}_ρ , follows from Skaug & Tjøstheim (1993) and also Robinson (1991).

Asymptotic normality of the measure \tilde{S}_ρ is given by the following theorem which is adapted from Skaug & Tjøstheim (1996).

Theorem 1. *Let $\{Y_t, t \geq 1\}$ be iid random variables. Under assumptions A1-A3 we have:*

$$n^{1/2}\tilde{S}_\rho \rightarrow N\left(0, \frac{1}{4}\sigma^2\right) \quad (5)$$

where

$$\sigma^2 = \left[\int f(y)w(y)dy \right]^2 \left[\int 1 - f(y)w(y)dy \right]^2. \quad (6)$$

Proof. Slight adaptation of Theorem 1 of Skaug & Tjøstheim (1996). □

As has been noted by Robinson (1991) and others, when $w(\cdot) = 1$ we obtain $\sigma^2 = 0$, a degenerate distribution. Therefore, the weight function is a theoretically important device for obtaining the asymptotic distribution of these type of statistics.

The large-sample Gaussian distribution is known to be a very poor approximation to the finite-sample one. Skaug & Tjøstheim (1993) and Skaug & Tjøstheim (1996) have also studied this issue in the context of testing for serial independence. They note that the asymptotic variance is very poorly estimated in the case of the Hellinger and cross entropy measures, which renders asymptotic inferences quite unreliable. These same reasons suggest that bootstrapping “asymptotically pivotal” statistics may not perform well in this context.

A serious problem with the asymptotic approach in a kernel context is that the asymptotic-based null distribution would not depend on the bandwidth, while the value of the test statistic does so directly. This is partly because the bandwidth vanishes asymptotically. This is a serious drawback in practise, since the outcome of such asymptotic-based tests tends to be quite sensitive to the choice of bandwidth. This has been reported by a number of authors including Robinson (1991) who, in a kernel context, noted “substantial variability in the [test statistic] across bandwidths was recorded,” which would be quite disturbing in applied situations.

The asymptotic theory for residual-based tests will require similar techniques to those used by Chen (2003) who exploits the asymptotic properties of the Gaussian quasi-maximum-likelihood estimators (QMLEs), to extend the original-series-based tests as model diagnostic checks for a general model. Since these model diagnostic checks are based on a formal asymptotic method, they can be implemented without bootstrapping. We have not developed a similar theory for our tests based on the standardized residuals in this paper.

We now turn to several applications of a bootstrap implementation of the entropy metric, including as a measure of goodness-of-fit, a measure of nonlinear “serial” dependence in both the raw return series and the residuals of certain models, and as a test of time-reversibility in pairwise comparisons for a small set of k lags.

4 Application: Dynamic Model Specification for Stock Returns

Section 3 outlines Monte Carlo experiments which reveal that the proposed tests are correctly sized under the null and have power approaching one under the alternative as the sample size increases. We now turn to an illustrative application, closely following the work of Chen & Kuan (2002) to facilitate comparison. Since all tests considered herein employ resampling methods yielding tests that are correctly sized, this application will serve to underscore power differences, albeit in an applied setting.

4.1 Data Sources

We use the data series found in Chen & Kuan (2002), who apply their TR characteristic function-based test to residuals from a variety of models of daily returns of six major US stock indices. The indices are the Dow Jones Industrial Averages (DJIA), New York Stock Exchange Composite (NYSE), Standard and Poor's 500 (S&P500), National Association of Securities Dealers Automated Quotations Composite (NASDAQ), Russell 2000 (RS2000), and Pacific Exchange Technology (PE-TECH). Each series contains $T = 2,527$ observations running from January 1 1991 through December 31 2000, and we let $Y_t = 100 \times (\log P_t - \log P_{t-1})$ denote the daily return of the index P_t .

4.2 Assessing Dependence in the Series

We begin by considering whether or not there exists potential nonlinear dependence in the raw series themselves. We therefore compute our metric entropy \hat{S}_ρ using $k = 1, 2, \dots, 5$ lags for each raw series, and use this as the basis for a test of independence following Granger et al. (2004). Bandwidths were selected via likelihood cross-validation, and the Gaussian kernel was used throughout. We construct P -values for the hypothesis that each series is a serially independent white-noise process and, for comparison, we also compute P -values for the Q test. Results are summarized in Table 4.

As can be seen from Table 4, there is extremely strong and robust evidence in favour of dependence being present in all of these series based on \hat{S}_ρ . However, the correlation-based Q statistic fails to capture this dependence for the DJIA, NASDAQ, and S&P500 across a range of lags. This suggests, as has been often observed, that correlation-based tests often lack power in nonlinear settings.

Next, we follow Chen & Kuan (2002) who assess the suitability of two classes of popular time-series models which have been used to model such processes.

Table 4: P -values for the entropy-based and Q tests for serial independence at various lags on the raw data series.

Series	\hat{S}_ρ					Q				
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
DJIA	0.00	0.00	0.00	0.00	0.00	0.36	0.13	0.04	0.09	0.14
NASDAQ	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.07	0.14	0.18
NYSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PETECH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RS2000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S&P500	0.01	0.00	0.00	0.00	0.00	0.00	0.48	0.02	0.05	0.02

4.3 Assessing Goodness-of-Fit

As in Chen & Kuan (2002), we consider two models, the GARCH(p, q) (Bollerslev (1986)) and EGARCH(p, q) (Nelson (1991)) specifications for a time-series $Y_t | \Psi_{t-1} = \epsilon_t \sim N(0, h_t)$ which we now briefly outline. The GARCH(p, q) model may be expressed as

$$\text{GARCH}(p, q) : \quad h_t = \omega + \sum_{i=1}^q \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^p \gamma_j h_{t-j}, \quad (7)$$

where $p \geq 0$, $q > 0$, $\omega > 0$, $\alpha_i \geq 0$, and $\gamma_j \geq 0$, while the EGARCH(p, q) model may be written as

$$\text{EGARCH}(p, q) : \quad \ln(h_t) = \omega + \sum_{i=1}^q \alpha_i g(z_{t-i}) + \sum_{j=1}^p \gamma_j \ln(h_{t-j}), \quad (8)$$

$$\text{where } g(z_t) = \theta z_t + \gamma [|z_t| - E|z_t|] \text{ and } z_t = \epsilon_t / \sqrt{h_t}.$$

These models will serve multiple illustrative purposes for what follows. In order to facilitate direct comparison with Chen & Kuan (2002), for all GARCH($1, k$) and EGARCH($1, k$) models listed below, the value of k is chosen to be the maximum k ($k \leq 5$) such that their TR statistic is significant at the 5% level.

We first consider the question of whether these nonlinear models differ in terms of their goodness-of-fit for a given series as these models are often considered equal on the basis of ‘‘goodness-of-fit’’ criteria. However, goodness-of-fit criteria such as R^2 are correlation-based. Out of concern that the ‘equality’ of models might be an artifact of using correlation-based measures of fit, we therefore compute the entropy measure with $f = f(\hat{Y}_t, Y_t)$ as the joint density of the predicted and actual excess returns, and $f_1 = f(\hat{Y}_t)$ and $f_2 = f(Y_t)$ as the respective marginal densities. If predicted and actual returns are independent, this metric will yield the value zero, and will increase as the model’s predictive ability improves.

In order to determine whether or not two model's measures of goodness-of-fit differ statistically, we require the sampling distribution of the goodness-of-fit measure itself. To obtain the percentiles for our goodness-of-fit statistic we employed the stationary bootstrap of Politis & Romano (1994) and 399 bootstrap replications. Results are summarized in Table 5.

Table 5: \hat{S}_ρ measures of goodness-of-fit along with their resampled 90% interval estimates.

Series	\hat{S}_ρ	$[pct_5, pct_{95}]$
DJIA [EGARCH(1, 1)]	0.07	[0.06, 0.15]
DJIA [EGARCH(1, k)]	0.06	[0.05, 0.16]
DJIA [GARCH(1, 1)]	0.07	[0.06, 0.13]
DJIA [GARCH(1, k)]	0.07	[0.06, 0.14]
NASDAQ [EGARCH(1, 1)]	0.12	[0.09, 0.25]
NASDAQ [EGARCH(1, k)]	0.12	[0.09, 0.29]
NASDAQ [GARCH(1, 1)]	0.12	[0.10, 0.23]
NASDAQ [GARCH(1, k)]	0.12	[0.09, 0.21]
NYSE [EGARCH(1, 1)]	0.06	[0.05, 0.09]
NYSE [EGARCH(1, k)]	0.05	[0.05, 0.09]
NYSE [GARCH(1, 1)]	0.06	[0.05, 0.09]
NYSE [GARCH(1, k)]	0.06	[0.05, 0.10]
PETECH [EGARCH(1, 1)]	0.16	[0.14, 0.19]
PETECH [EGARCH(1, k)]	0.16	[0.13, 0.19]
PETECH [GARCH(1, 1)]	0.16	[0.14, 0.20]
PETECH [GARCH(1, k)]	0.16	[0.14, 0.21]
RS2000 [EGARCH(1, 1)]	0.08	[0.05, 0.14]
RS2000 [EGARCH(1, k)]	0.07	[0.05, 0.14]
RS2000 [GARCH(1, 1)]	0.08	[0.06, 0.14]
RS2000 [GARCH(1, k)]	0.08	[0.06, 0.15]
S&P500 [EGARCH(1, 1)]	0.08	[0.06, 0.10]
S&P500 [EGARCH(1, k)]	0.07	[0.06, 0.11]
S&P500 [GARCH(1, 1)]	0.08	[0.06, 0.10]
S&P500 [GARCH(1, k)]	0.07	[0.06, 0.10]

Given results summarized in Table 5, it is evident that there does not appear to be any significant difference between models in terms of their fidelity to the data for a given series. This common finding leads naturally to residual-based specification testing to which we now proceed.

4.4 Assessing Model Specification

Next, we focus on using S_ρ as a diagnostic tool. Under the null of correct model specification, the model residuals would be indistinguishable from white noise at all lags. Table 6 reports the associated P -values for the entropy-based test for *serial independence* at various lags on time-series

models' residuals using 99 permutation replications, the Ljung-Box Q test, the modified BDS test, and those for Chen & Kuan's (2002) TR test. It should be noted that time-reversibility is only necessary for *iid*-ness. Due to consistency for the raw series, our test will have power against both symmetric and asymmetric alternatives, as well as such time-irreversible processes as STAR, SETAR, EGARCH and many Regime Switching models.

Table 6 reveals that the correlation-based Q statistic almost uniformly fails to have power in the direction of misspecification for both the GARCH and EGARCH models. The "robust BDS" test performs better, though relative to the TR test, the BDS also fails to have power in a number of instances. The competing TR test performs quite well, though we note that it too lacks power for several cases of EGARCH(1, k). In particular, Chen & Kuan (2002), on the basis of their reversibility tests, conclude that expanded EGARCH specifications are correct, further noting that "the [proposed] test detects volatility asymmetry that cannot be detected by the BDS test [...] providing more information regarding how a model should be refined" (Chen & Kuan (2002, pg 577)). Table 6 reveals that the correlation-based Q test, the chaos-based BDS test, and characteristic function-based TR test fail to reject the EGARCH(1, k) specification across series. In contrast, the entropy-based TR test detects misspecification across EGARCH(1, k) models for every series at lags 1 and/or 2 at the 5% level except the NASDAQ (though the appropriateness of this model is rejected at lag 2 at the 10% level.). As demonstrated in our Monte Carlo evidence in Section 3, this is not due to any size distortion.

4.5 Assessing Time-Reversibility

As has been noted in the literature, time-irreversibility is the norm rather than exception for nonlinear (financial) time-series (see Tong (1990)). The time-reversibility hypothesis for these same raw return series is rejected by Chen & Kuan (2002). Table 7 contains our findings for the time-reversibility properties of the raw series as well as their standardized residuals. For the raw series (first panel) we find similar results by looking at $k = 1, 2, \dots, 5$. Every return series except the NASDAQ and RS2000 fails the TR test at some lag. Interestingly, S&P500 fails at small lags and at lag 5, whereas DJIA fails marginally only at the first lag. Several panels focus on the standardized residuals of the models described above. We see that the GARCH models perform badly since they appear to induce irreversibility in all the residual series, and at many more lags! A portmanteau test will not be necessary in these cases. EGARCH(1, 1) does considerably better, but still failing the NYSE and PETECH residuals. EGARCH(1, k) does much better only coming close to borders of rejection at some lags for NASDAQ ($k = 3$) and RS2000 ($k = 1$). The P -values indicate that the residuals of the EGARCH(1, k) generally enjoy the time-reversibility property with greater confidence than the original series.

The performance of the entropy tests point to good power as, indicated earlier, and as shown

in Skaug & Tjøstheim (1993) for the raw series.⁴ We conclude that the relative failure to detect this misspecification by competing tests would merely reflect their lack of power in some directions. Relative to its peers, the entropy-based test has two features to recommend its use as a diagnostic tool:

1. It has generally higher power than competing correlation-based and even characteristic-function-based tests.
2. It provides surprisingly strong indications of where models fail (e.g., at lags 1 and 2) hence provides prescriptive advice for model refinement in that limited sense. One constructive way to proceed would be to use our statistic as a goodness-of-fit measure, as was done above, to choose between competing models that accommodate time-irreversibility, such as the STAR, SETAR, EGARCH and switching regression models. This may be an informed selection process where rejections at specific lags, as above, guide the choice of competing models. Additionally, we may employ our entropy statistic as a measure of out of sample predictive performance, as was done in Maasoumi & Racine (2002) in a similar context. These substantive specification searches are beyond the scope of the current paper.

5 Conclusion

We consider the application of a versatile metric entropy for detecting departures from serial independence, and time-reversibility of series and residuals of some popular models to aid in the construction of parametric time-series models. Applications indicate this “diagnostic” approach may offer unusually *constructive* prescriptions for model specification, both when applied to the raw series and the residuals of models, and in conjunction with its use as a measure of fit and predictive performance.

References

- Akaike, H. (1981), ‘Likelihood of a model and information criteria’, *Journal of Econometrics* **16**, 3–14.
- Bai, J. & Ng, S. (2001), ‘A consistent test for conditional symmetry’, *Journal of Econometrics* **103**, 225–258.
- Barnett, B., Gallant, A. R., Hinich, M. J., Jungeilges, J. A., Kaplan, D. T. & Jensen, M. J. (1997), ‘A single-blind controlled competition among tests for nonlinearity and chaos’, *Journal of Econometrics* **82**, 157–192.

⁴The asymptotic distribution of the proposed entropy tests for the standardized residuals are yet to be developed. Experience has shown that such asymptotic approximations are generally poor and the bootstrap and other approximations to the permutation tests remain the procedure of choice. See, for example, Skaug & Tjøstheim (1996) on this issue, and Section 3.

- Belaire-Franch, J. & Contreras, D. (2002), ‘How to compute the BDS test: a software comparison’, *Journal of Applied Econometrics* **17**, 691–699.
- Belaire-Franch, J. & Peiro, A. (2003), ‘Conditional and unconditional asymmetry in U.S. macroeconomic time series’, *Studies in Nonlinear Dynamics and Econometrics* **7**, Article 4.
- Bollerslev, T. (1986), ‘Generalized autoregressive conditional heteroskedasticity’, *Journal of Econometrics* **31**, 307–27.
- Brock, W. A., Dechert, W. D. & Scheinkman, J. A. (1987), A test for independence based on the correlation dimension, University of Wisconsin-Madison Social Systems Research Institute Working Paper 8702, University of Wisconsin-Madison Social Systems Research Institute.
- Brock, W. A., Dechert, W. D., Scheinkman, J. A. & LeBaron, B. (1996), ‘A test for independence based on the correlation dimension’, *Econometric Reviews* **15**(3), 197–235.
- Chen, Y. (2003), ‘Testing serial independence against time irreversibility’, *Studies in Nonlinear Dynamics & Econometrics* **7**, 1–28.
- Chen, Y., Chou, R. & Kuan, C. (2000), ‘Testing time reversibility without moment restrictions’, *Journal of Econometrics* **95**, 199–218.
- Chen, Y. & Kuan, C. (2002), ‘Time irreversibility and EGARCH effects in US stock index returns’, *Journal of Applied Econometrics* **17**, 565–578.
- Clements, M. P., Franses, P. H., Smith, J. & Van Dijk, D. (2003), ‘On SETAR non-linearity and forecasting’, *Journal of Forecasting* **22**, 359–375.
- Cromwell, J. B., Labys, W. C. & Terraza, M. (1994), *Univariate Tests for Time Series Models*, Sage.
- Darolles, S., Florens, J.-P. & Gouriéroux, C. (forthcoming), ‘Kernel-based nonlinear canonical analysis and time reversibility’, *Journal of Econometrics* .
- Diks, C. & Manzan, S. (2002), ‘Tests for serial independence and linearity based on correlation integrals’, *Studies in Nonlinear Dynamics and Econometrics* **6**.
- Granger, C., Maasoumi, E. & Racine, J. S. (2004), ‘A dependence metric for possibly nonlinear time series’, *Journal of Time Series Analysis* **25**(5), 649–669.
- Hallin, M., Lefevre, C. & Puri, M. (1988), ‘On time-reversibility and the uniqueness of moving average representations for non-gaussian stationary time series’, *Biometrika* **75**, 170–171.
- Hong, Y. & White, H. (forthcoming), ‘Asymptotic distribution theory for nonparametric entropy measures of serial dependence’, *Econometrica* .
- Ihaka, R. & Gentleman, R. (1996), ‘R: A language for data analysis and graphics’, *Journal of Computational and Graphical Statistics* **5**(3), 299–314.
- Kanzler, L. (1999), ‘Very fast and correctly sized estimation of the BDS statistic’, *Christ Church and Department of Economics, University of Oxford* .

- Ljung, G. & Box, G. (1978), ‘On a measure of lack of fit in time series models’, *Biometrika* **65**, 297–303.
- Maasoumi, E. & Racine, J. S. (2002), ‘Entropy and predictability of stock market returns’, *Journal of Econometrics* **107**(2), 291–312.
- Nelson, C. R. & Plosser, C. (1982), ‘Trends and random walks in macro-economic time series: some evidence and implications’, *Journal of Monetary Economics* **10**, 139–162.
- Nelson, D. B. (1991), ‘Conditional heteroskedasticity in asset returns: A new approach’, *Econometrica* **59**(2), 347–370.
- Parzen, E. (1962), ‘On estimation of a probability density function and mode’, *The Annals of Mathematical Statistics* **33**, 1065–1076.
- Politis, D. N. & Romano, J. P. (1994), ‘The stationary bootstrap’, *Journal of the American Statistical Association* **89**, 1303–1313.
- Prichard, D. & Theiler, J. (1995), ‘Generalized redundancies for time series analysis’, *Physica D* **84**, 476–493.
- Robinson, P. M. (1991), ‘Consistent nonparametric entropy-based testing’, *Review of Economic Studies* **58**, 437–453.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.
- Skaug, H. & Tjøstheim, D. (1993), Nonparametric tests of serial independence, in S. Rao, ed., ‘Developments in Time Series Analysis’, Chapman and Hall, pp. 207–229.
- Skaug, H. & Tjøstheim, D. (1996), Testing for serial independence using measures of distance between densities, in P. Robinson & M. Rosenblatt, eds, ‘Athens Conference on Applied Probability and Time Series’, Springer Lecture Notes in Statistics, Springer.
- Tong, H. (1990), *Nonlinear Time Series - A Dynamic System Approach*, Oxford University Press.
- Weiss, G. (1975), ‘Time-reversibility of linear stochastic processes’, *Journal of Applied Probability* **12**, 831–836.

Table 6: Test results from various tests for serial independence at various lags on time-series models' residuals. For the \hat{S}_ρ and Q -tests we present P -values for the null of correct specification, while for the BDS and TR tests we present the actual statistics and flag those values which are significant with an asterisk. Therefore, for comparison purposes, all entries which are significant at the $\alpha = 0.05$ level are marked with an asterisk (i.e., P -values and actual statistics). Throughout we use k to denote lag, and n to denote embedding dimension.

Series	S_ρ Entropy Test					Q Test					BDS Test					TR Test				
	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 10$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 3$	$n = 4$	$n = 1$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$		
GARCH(1,1)																				
DJIA	0.00*	0.00*	0.00*	0.41	0.59	0.17	-0.55	-0.07	-0.21	-0.12	-0.12	-0.12	-0.12	-3.19*	-3.28*	-0.99	-1.16	-1.26		
NASDAQ	0.00*	0.00*	0.49	0.82	0.30	0.98	0.68	1.90	1.88	2.05*	2.05*	2.05*	2.05*	-5.52*	-4.15*	-2.11*	0.36	-2.02*		
NYSE	0.00*	0.00*	0.00*	0.52	0.01*	0.06	-1.29	-0.49	-0.69	-0.42	-0.42	-0.42	-0.42	-2.61*	-4.65*	-0.72	-1.77	-1.81		
PETECH	0.00*	0.00*	0.05*	0.88	0.46	0.44	-0.62	0.19	0.25	0.42	0.42	0.42	0.42	-4.45*	-3.59*	-2.40*	-0.76	-1.25		
RS2000	0.00*	0.01*	0.13	0.51	0.58	0.66	2.55*	3.68*	3.69*	3.89*	3.89*	3.89*	3.89*	-3.23*	-3.26*	-2.56*	0.32	-0.83		
S&P500	0.00*	0.00*	0.07	0.34	0.01*	0.28	-1.69	-0.81	-0.79	-0.30	-0.30	-0.30	-0.30	-3.38*	-4.65*	-0.89	-1.91	-2.59*		
GARCH(1,k)																				
DJIA	0.00*	0.00*	0.00*	0.23	0.47	0.14	-1.20	-1.49	-1.61	-1.32	-1.32	-1.32	-1.32	-2.93*	-2.80*	-0.72	-1.03	-1.29		
NASDAQ	0.00*	0.00*	0.19	0.72	0.54	0.96	-0.09	0.25	0.20	0.18	0.18	0.18	0.18	-5.38*	-3.81*	-2.23*	0.45	-1.86*		
NYSE	0.00*	0.00*	0.01*	0.31	0.03*	0.05*	-1.41	-1.55	-1.72	-1.32	-1.32	-1.32	-1.32	-2.75*	-4.43*	-0.46	-1.38	-1.65		
PETECH	0.00*	0.00*	0.05*	0.94	0.31	0.41	-0.85	-0.65	-0.45	-0.04	-0.04	-0.04	-0.04	-4.41*	-3.19*	-2.22*	-0.73	-1.30		
RS2000	0.01*	0.01*	0.34	0.10	0.20	0.51	1.77	2.03*	2.09*	2.01*	2.01*	2.01*	2.01*	-3.27*	-3.04*	-2.62*	0.13	-0.37		
S&P500	0.00*	0.00*	0.20	0.33	0.00*	0.25	-1.67	-1.75	-1.77	-1.14	-1.14	-1.14	-1.14	-3.63*	-4.06*	-0.72	-1.74	-2.47*		
EGARCH(1,1)																				
DJIA	0.00*	0.25	0.60	0.43	0.85	0.18	-1.04	-0.74	-0.87	-0.79	-0.79	-0.79	-0.79	-1.96*	-1.98*	-0.14	-0.41	-0.54		
NASDAQ	0.00*	0.00*	0.65	0.60	0.53	0.99	0.42	1.56	1.55	1.73	1.73	1.73	1.73	-3.70*	-2.72*	-1.03	1.29	-1.27		
NYSE	0.00*	0.06	0.12	0.75	0.08	0.12	-1.83	-1.21	-1.41	-1.08	-1.08	-1.08	-1.08	-1.11	-3.27*	-0.31	-0.77	-0.81		
PETECH	0.00*	0.00*	0.33	0.32	0.68	0.59	-1.33	-0.66	-0.50	-0.31	-0.31	-0.31	-0.31	-3.09*	-2.06*	-1.28	0.21	-0.26		
RS2000	0.01*	0.02*	0.48	0.26	0.84	0.69	2.79*	3.88*	3.89*	4.26*	4.26*	4.26*	4.26*	-1.56	-1.89	-1.96*	1.10	-0.14		
S&P500	0.00*	0.00*	0.67	0.61	0.06	0.21	-2.57*	-1.83	-1.90	-1.39	-1.39	-1.39	-1.39	-1.98*	-3.01*	0.27	-0.92	-1.65		
EGARCH(1,k)																				
DJIA	0.00*	0.29	0.10	0.22	0.73	0.13	-0.83	-0.66	-0.72	-0.56	-0.56	-0.56	-0.56	-0.10	0.40	-0.52	-0.92	-0.83		
NASDAQ	0.16	0.06	0.44	0.79	0.58	0.94	-1.00	-0.50	-0.67	-0.57	-0.57	-0.57	-0.57	-0.69	0.06	-1.11	-0.29	-1.07		
NYSE	0.01*	0.10	0.04*	0.42	0.05*	0.07	-0.95	-0.74	-0.96	-0.61	-0.61	-0.61	-0.61	0.75	-0.13	-0.22	-1.31	-1.02		
PETECH	0.13	0.00*	0.34	0.48	0.19	0.59	-1.18	-0.65	-0.64	-0.48	-0.48	-0.48	-0.48	-1.11	0.42	-0.63	-0.62	-1.09		
RS2000	0.02*	0.11	0.57	0.03*	0.66	0.48	1.48	1.47	1.32	1.73	1.73	1.73	1.73	0.79	-0.13	-1.90	-0.95	-1.31		
S&P500	0.00*	0.10	0.50	0.57	0.13	0.36	-1.17	-0.96	-1.14	-0.84	-0.84	-0.84	-0.84	-0.51	-0.25	0.65	-2.01*	-0.98		

Table 7: P -values for the entropy-TR tests.

Series	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Series					
DJIA	0.08	0.17	0.82	0.68	0.24
NASDAQ	0.41	0.32	0.52	0.79	0.72
NYSE	0.17	0.06	0.74	0.76	0.38
PETECH	0.02	0.05	0.12	0.52	0.58
RS2000	0.34	0.12	0.11	0.35	0.49
S&P500	0.04	0.03	0.52	0.44	0.08
GARCH(1, 1) Residuals					
DJIA	0.15	0.07	0.76	0.59	0.95
NASDAQ	0.12	0.00	0.66	0.48	0.52
NYSE	0.00	0.00	0.35	0.67	0.46
PETECH	0.00	0.00	0.09	0.82	0.43
RS2000	0.04	0.04	0.40	0.45	0.55
S&P500	0.13	0.00	0.76	0.42	0.44
GARCH(1, k) Residuals					
DJIA	0.08	0.05	0.80	0.73	0.76
NASDAQ	0.09	0.02	0.52	0.41	0.53
NYSE	0.00	0.01	0.10	0.79	0.48
PETECH	0.00	0.00	0.09	0.55	0.45
RS2000	0.04	0.11	0.34	0.45	0.65
S&P500	0.03	0.01	0.87	0.45	0.59
EGARCH(1, 1) Residuals					
DJIA	0.30	0.39	0.83	0.69	0.94
NASDAQ	0.37	0.24	0.68	0.70	0.98
NYSE	0.06	0.11	0.90	0.77	0.46
PETECH	0.09	0.01	0.49	0.85	0.99
RS2000	0.26	0.23	0.66	0.58	0.94
S&P500	0.34	0.30	0.99	0.67	0.81
EGARCH(1, k) Residuals					
DJIA	0.77	0.79	0.43	0.51	0.79
NASDAQ	0.49	0.35	0.12	0.45	0.65
NYSE	0.54	0.90	0.70	0.79	0.56
PETECH	0.69	0.41	0.77	0.63	0.80
RS2000	0.10	0.80	0.43	0.39	0.55
S&P500	0.55	0.82	0.49	0.45	0.97