A new statistic and practical guidelines for nonparametric Granger causality testing

Cees Diks & Valentyn Panchenko

Center for Nonlinear Dynamics in Economics and Finance

Department of Economics, University of Amsterdam, Roetersstraat 11

1018 WB Amsterdam, The Netherlands

May 4, 2005

Abstract

In this paper we introduce a new nonparametric test for Granger non-causality which avoids the over-rejection observed in the frequently used test proposed by Hiemstra and Jones (1994). After illustrating the problem by showing that rejection probabilities under the null hypothesis may tend to one as the sample size increases, we study the reason behind this phenomenon analytically. It turns out that the Hiemstra-Jones test for the null of Granger non-causality, which can be rephrased in terms of conditional independence of two vectors X and Z given a third vector Y, is sensitive to variations in the conditional distributions of X and Z that may be present under the null. To overcome this problem we replace the global test statistic by an average of local conditional dependence measures. By letting the bandwidth tend to zero at appropriate rates, the variations in the conditional distributions are accounted for automatically. Based on asymptotic theory we formulate practical guidelines for choosing the bandwidth depending on the sample size. We conclude with an application to historical returns and trading volumes of the Standard and Poor's index which indicates that the evidence for volume Granger-causing returns is weaker than suggested by the Hiemstra-Jones test.

Keywords: Financial time series; Granger causality; Nonparametric; Hypothesis testing; Size distortion; U-statistics;

JEL classification: C12, C51, E3

1 Introduction

Granger (1969) causality has turned out to be a useful notion for characterizing dependence relations between time series in economics and econometrics. Intuitively, for a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if past and current values of X contain additional information on future values of Y that is not contained in past and current Y-values alone. If we denote the information contained in past observations X_s and Y_s, $s \leq t$, by $\mathcal{F}_{X,t}$ and $\mathcal{F}_{Y,t}$, respectively, and let '$\sim$' denote equivalence in distribution, the formal definition is:

Definition 1 *For a strictly stationary bivariate time series process* $\{(X_t, Y_t)\}$, $t \in \mathbb{Z}$, $\{X_t\}$ *is a Granger cause of* $\{Y_t\}$ *if, for some* $k \geq 1$,

$$(Y_{t+1}, \ldots, Y_{t+k})|(\mathcal{F}_{X,t}, \mathcal{F}_{Y,t}) \not\sim (Y_{t+1}, \ldots, Y_{t+k})|\mathcal{F}_{Y,t}.$$

Since this definition is general and does not involve any modelling assumptions, such as a linear autoregressive model, it is often referred to as general or, by a slight abuse of language, nonlinear Granger causality.

Traditional parametric tests for Granger non-causality within linear autoregressive model classes have reached a mature status, and have become part of the standard toolbox of economists. The recent literature, due to the availability of ever cheaper computational power, has shown an increasing interest in nonparametric versions of the Granger non-causality hypothesis against general (linear as well as nonlinear) Granger causality. Among the various nonparametric tests for the Granger non-causality hypothesis, the Hiemstra and Jones (1994) test (hereafter HJ test) is the most frequently used among practitioners in economics and finance. Although alternative tests, such as that proposed by Bell *et al.* (1996), and by Su and White (2003), may also be applied in economics and finance, we limit ourselves to a discussion of the HJ test and our proposed modification of it.

The reason for considering the HJ test here in detail is our earlier finding (Diks and Panchenko, 2005) that this commonly used test can severely over-reject if the null hypothesis is true. The aim of the present paper is two-fold. First, we derive the exact conditions under which the HJ test over-rejects, and secondly we propose a new test statistic which does not suffer from this serious limitation. We will show that the reason for over-rejection of the HJ test is that the test statistic, due to its global nature, ignores the possible variation in conditional distributions that may be present under the null hypothesis. Our new test statistic, provided that the bandwidth tends to zero at an appropriate rate, automatically takes into account such variation under the null hypothesis while obtaining an asymptotically correct size.

The practical implication of our findings is far-reaching: all cases for which evidence for Granger causality was reported based on the HJ test may be caused by the tendency of the HJ test to over-reject. Reports of such evidence are numerous in the economics and finance literature. For instance, Brooks (1998) finds evidence for Granger causality between volume and volatility on the New York Stock Exchange, Abhyankar (1998) and Silvapulla and Moosa (1999) in futures markets, and Ma and Kanas (2000) in exchange rates. Further evidence for causality is reported in stock markets (Ciner, 2001), among real estate prices and stock markets (Okunev *et al.*, 2000, 2002) and between London Metal Exchange cash prices and some of its possible predictors (Chen and Lin, 2004). Although we do not claim that the reported Granger causality is absent in all these cases, we do state that the statistical justification is not warranted.

This paper is organized as follows. In section 2 we show that the HJ test statistic can give rise to rejection probabilities that tend to one with increasing sample size under the null hypothesis. In section 3 the reason behind this phenomenon is studied analytically and found to be related to a bias in the test statistic due to variations in conditional distributions. The analytic results suggest an alternative test statistic, described in Section 4, which automatically takes these variations into account, and can be shown to give asymptotic rejection rates equal to the nominal size for bandwidths tending to zero at appropriate rates. The theory is confirmed by the simulation results presented at the end of the section. In Section 5 we consider an application to S&P500 volumes and returns for which the HJ test indicates volume Granger-causing returns, while our test indicates that the evidence for volume causing returns is considerably weaker. Section 6 summarizes and concludes.

2 The Hiemstra-Jones Test

In testing for Granger non-causality, the aim is to detect evidence against the null hypothesis

$$H_0: \qquad \{X_t\} \text{ is not Granger causing } \{Y_t\},$$

with Granger causality defined according to Definition 1. We limit ourselves to tests for detecting Granger causality for $k = 1$, which is the case considered most often in practice. Under the null hypothesis Y_{t+1} is conditionally independent of $X_t, X_{t-1}, \ldots$, given $Y_t, Y_{t-1}, \ldots$. In a nonparametric setting, conditioning on the infinite past is impossible without a model restriction, such as an assumption that the order of the process is finite. Therefore, in practice conditional independence is tested using finite lags l_X and l_Y:

$$Y_{t+1}|(X_t^{l_X}; Y_t^{l_Y}) \sim Y_{t+1}|Y_t^{l_Y},$$

where $X_t^{l_X} = (X_{t-l_X+1}, \ldots, X_t)$ and $Y_t^{l_Y} = (Y_{t-l_Y+1}, \ldots, Y_t)$. For a strictly stationary bivariate time series $\{(X_t, Y_t)\}$ this is a statement about the invariant distribution of the l_X+l_Y+1-dimensional vector $W_t = (X_t^{l_X}, Y_t^{l_Y}, Z_t)$, where $Z_t = Y_{t+1}$. To keep the notation compact, and to bring about the fact that the null hypothesis is a statement about the invariant distribution of W_t, we often drop the time index and just write $W = (X, Y, Z)$, where the latter is a random vector with the invariant distribution of $(X_t^{l_X}, Y_t^{l_Y}, Y_{t+1})$. In this paper we only consider the choice $l_X = l_Y = 1$, in which case $W = (X, Y, Z)$ denotes a three-variate random variable, distributed as $W_t = (X_t, Y_t, Y_{t+1})$. Throughout we will assume that W is a continuous random variable.

The HJ test is a modified version of the Baek and Brock (1992) test for conditional independence, with critical values based on asymptotic theory. To motivate the test statistic it is convenient to restate the null hypothesis in terms of ratios of joint distributions. Under the null the conditional distribution of Z given $(X, Y) = (x, y)$ is the same as that of Z given $Y = y$ only, so that the joint probability density function $f_{X,Y,Z}(x, y, z)$ and its marginals must satisfy

$$\frac{f_{X,Y,Z}(x,y,z)}{f_{X,Y}(x,y)} = \frac{f_{Y,Z}(y,z)}{f_Y(y)}, \tag{1a}$$

or equivalently

$$\frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \frac{f_{Y,Z}(y,z)}{f_Y(y)} \tag{1b}$$

for each vector (x, y, z) in the support of (X, Y, Z). The last equation is identical to $f_{X,Z|Y}(x, z|y) = f_{X|Y}(x|y) f_{Z|Y}(z|y)$, which explicitly states that X and Z are independent conditionally on $Y = y$, for each fixed value of y.

The Hiemstra-Jones test employs ratios of correlation integrals to measure the discrepancy between the left- and right-hand-sides of (1a). For a multivariate random vector V taking values in $\mathbb{R}^{d_V}$ the associated correlation integral $C_V(\varepsilon)$ is the probability of finding two independent realisations of the vector at a distance smaller than or equal to ε:

$$\begin{aligned} C_V(\varepsilon) &= P[\|V_1 - V_2\| \le \varepsilon], \qquad V_1, V_2 \text{ indep.} \sim V \\ &= \int\int I(\|s_1 - s_2\| \le \varepsilon) f_V(s_1) f_V(s_2)\, \mathrm{d}s_2\, \mathrm{d}s_1 \end{aligned}$$

where $I(\|s_1-s_2\| \le \varepsilon)$ is the indicator function, which is one if $\|s_1-s_2\| \le \varepsilon$ and zero otherwise, and $\|x\| = \sup_{i=1,\ldots,d_V} |x_i|$ denotes the supremum norm. Hiemstra and Jones (1994) argue that Equation

(1a) implies for any $\varepsilon > 0$:

$$\frac{C_{X,Y,Z}(\varepsilon)}{C_{X,Y}(\varepsilon)} = \frac{C_{Y,Z}(\varepsilon)}{C_Y(\varepsilon)} \tag{2a}$$

or equivalently

$$\frac{C_{X,Y,Z}(\varepsilon)}{C_Y(\varepsilon)} = \frac{C_{X,Y}(\varepsilon)}{C_Y(\varepsilon)} \frac{C_{Y,Z}(\varepsilon)}{C_Y(\varepsilon)}. \tag{2b}$$

The HJ test consists of calculating sample versions of the correlation integrals in (2a), and then testing whether the left-hand- and right-hand-side ratios differ significantly or not. The estimators for each of the correlation integrals take the form

$$C_{W,n}(\varepsilon) = \frac{2}{n(n-1)} \sum\sum_{i<j} I_{ij}^W,$$

where $I_{ij}^W = I(\|W_i - W_j\| \leq \varepsilon)$. For the asymptotic theory we refer to Hiemstra and Jones (1994).

As stated in the introduction, the main motivation for the present paper is that in certain situations the HJ test rejects too often under the null, and we wish to formulate an alternative procedure to avoid this. Before investigating the reasons for over-rejection analytically, we use a simple example to illustrate the over-rejection numerically, and to show that simple remedies such as transforming the data to uniform marginals and filtering out GARCH structure do not work. Diks and Panchenko (2005) demonstrated that for a process with instantaneous dependence in conditional variance the actual size of the HJ test was severely distorted. Here we illustrate the same point for the similar process, but without instantaneous dependence:

$$\begin{aligned} X_t &\sim N\left(0, c + aY_{t-1}^2)\right) \\ Y_t &\sim N\left(0, c + aY_{t-1}^2)\right). \end{aligned} \tag{3}$$

This process satisfies the null hypothesis; $\{X_t\}$ is not Granger causing $\{Y_t\}$. The values for the coefficients a and c are chosen in such a way that the process remains stationary and ergodic ($c > 0$, $0 < a < 1$).

We performed some Monte Carlo simulations to obtain the empirical size of the HJ test for the ARCH process (3) with coefficients $c = 1$, $a = 0.4$. For various sample sizes, we generated 1 000 independent realisations of the bivariate process and determined the observed fraction of rejections of the null at a nominal size of 0.05. The solid line in Figure 1 shows the rejection rates found as a function of the time series length n. The simulated data were normalized to unit variance before the test was applied, and the bandwidth was set to $\varepsilon = 1$, which is within the common range $(0.5, 1.5)$ used in practice. For time series length $n < 500$ the test based on the original series under-rejects.

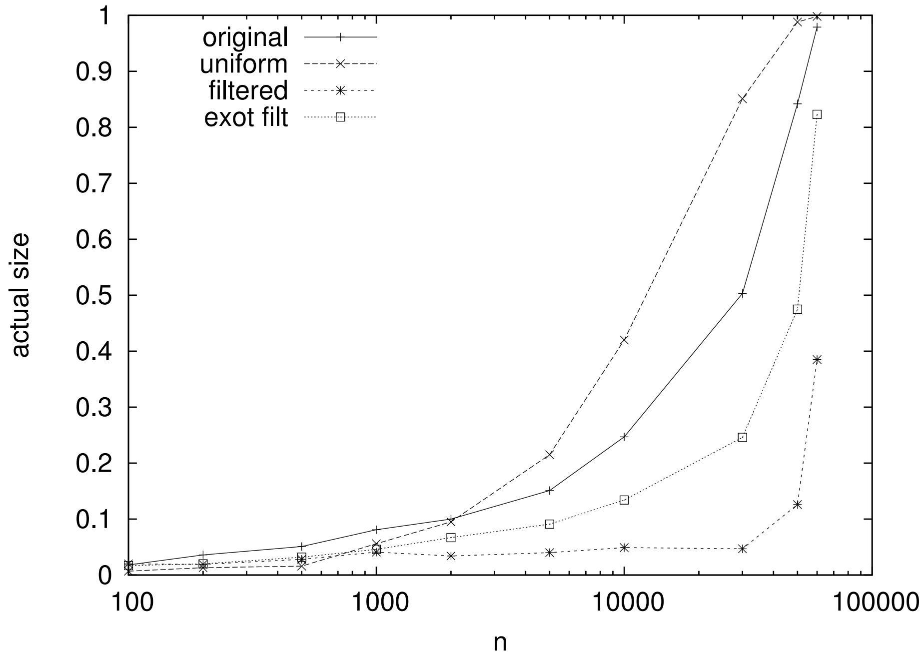

Figure 1: *Observed rejection rates (empirical size, number of realisations: 1 000) of the HJ test ($\varepsilon = 1$) for the bivariate ARCH process (3) as a function of the time series length n (nominal size 0.05) for: original data (solid line), uniformly transformed data (long-dashed line), ARCH filtered data (dashed line) and for data generated with model (4) and filtered with a misspecified ARCH(1) model (dotted line).*

Its size is close to nominal for series length $n = 500$. For longer series the actual size increases and becomes close to one when $n = 60\ 000$. The reason that the observed size increases with the series length n is that, as detailed in the next section, the test statistic is biased in that it does not converge in probability to zero under the null as the sample size increases. As the sample size increases the biases converges to a nonzero limit while the variance decreases to zero, giving rise to apparently significant values of the test statistic. In comparison with the process with instantaneous dependence considered in Diks and Panchenko (2005) the current process indicates less size distortion. This is due to the weaker covariance between the concentration measures H_X and H_Z for the current process, which is the main cause of the bias.

As suggested by Pompe (1993) in the context of testing for serial independence, transforming the time series to a uniform marginal distribution by using ranks, may improve the performance of the test. Here we investigate if it reduces the bias of the HJ test. The long-dashed line in Fig. 1 shows that the uniform transform improves the size for time series of length $n = 1\ 000$, but magnifies the size distortion for time series length $n > 2\ 000$.

As another solution one might argue that it is possible to filter out the conditional heteroskedasticity using a univariate (G)ARCH specification. This would remove the bias caused by the conditional heteroskedasticity in the HJ test. However such a filtering procedure has several drawbacks. First, it may affect the dependence structure and consequently the power of the test. Second, a (G)ARCH filter may not fully remove the conditional heteroskedasticity in the residuals. To illustrate the latter point we filtered the original series considered before by univariate ARCH(1) model. The parameters of the model were estimated for every realisation using the asymptotically efficient two stage procedure of Engle (1982). Figure 1 (dashed line) shows that the filtering removes the bias for time series length $n < 30\,000$, however the actual size remains distorted for longer series.

It is important to mention that in the previous case the correct model for the conditional variance of series Y_t was used and, as the next section clarifies, most of the source of the bias was removed. In practice the correct model is not known and the model used to filter out the heteroskedasticity is likely to be misspecified. To show the effect of model misspecification we generated data according to the following "exotic" ARCH model:

$$\begin{aligned} X_t &\sim N\left(0, c + aY_{t-1}^2 \exp(-bY_{t-1}^2)\right) \\ Y_t &\sim N\left(0, c + aY_{t-1}^2 \exp(-bY_{t-1}^2)\right). \end{aligned} \tag{4}$$

With parameters $c = 1$, $a = 2$ and $b = 0.4$ the process (4) is stationary and the fluctuations in the conditional variance are similar in magnitude as for the ARCH process (3) with the coefficients considered before. Instead of using a correctly specified filter we proceeded as before, calculating the size using a conventional ARCH(1) filter prior to application of the HJ test. The results represented by the dotted line in Fig. 1 indicate that the misspecified ARCH(1) filter is not able to remove large part of the source of bias and the sensitivity of the HJ test to dependence in the conditional variance leads to over-rejection, even for shorter time series.

3 Bias from correlations in conditional concentrations

In this section we show that the reason that the HJ test is inconsistent is that the assumption made by HJ that (1a) implies (2a) does not hold in general. In fact (2a) follows from (1a) only in specific cases, e.g. when the conditional distributions of Z and X given $Y = y$ do not depend on y. To see this, note that under the null hypothesis

$$\begin{aligned} & P[\|X_1 - X_2\| < \varepsilon, \|Z_1 - Z_2\| < \varepsilon | Y_1 = Y_2 = y] \\ & \quad = P[\|X_1 - X_2\| < \varepsilon | Y_1 = Y_2 = y] P[\|Z_1 - Z_2\| < \varepsilon | Y_1 = Y_2 = y], \end{aligned} \tag{5}$$

whereas Equation (2b) states

$$\begin{aligned} & P[\|X_1 - X_2\| < \varepsilon, \|Z_1 - Z_2\| < \varepsilon | \|Y_1 - Y_2\| < \varepsilon] \\ & \quad = P[\|X_1 - X_2\| < \varepsilon | \|Y_1 - Y_2\| < \varepsilon] P[\|Z_1 - Z_2\| < \varepsilon | \|Y_1 - Y_2\| < \varepsilon]. \end{aligned} \tag{6}$$

In general these conditions are not equivalent. In both equations a statement regarding the factorization of probabilities is made, but the events on which the conditioning takes place differ. In general, under the null the conditional distributions of X and Z are allowed to depend on Y. Therefore, the distributions of $X_1 - X_2$ and $Z_1 - Z_2$ will generally depend, under the null, on Y_1 and Y_2. Even for small ε the condition in Equation (6) holds for many close but very different Y_1, Y_2 pairs. Therefore, for small ε the left-hand-side of Equation 6 behaves as an average of that of Equation (5) over all possible values of y. Because factorization of densities is not preserved under averaging — $af_1(x)g_1(z) + (1-a)f_2(x)f_2(z)$ typically cannot be written as the product of a function of x and of z — the average probability on the left-hand-side of Equation (6) will typically not factorize in the form on the right-hand-side.

Although this argument shows that the relationship tested in the HJ test is generally inconsistent with the null hypothesis, one might argue that the test could still be asymptotically valid if appropriate measures are taken to eliminate the 'bias' in Eq. (2a) asymptotically, for example by allowing for the bandwidth ε to tend to zero at an appropriate rate with increasing sample size.

To see whether such an approach might work we examine the behavior of the fractions in (2a) for small values of the bandwidth ε. For continuous distributions the following small ε approximation is useful:

$$\begin{aligned} C_V(\varepsilon) &= \int\int I(\|s_1 - s_2\| \le \varepsilon) f_V(s_1) f_V(s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2 \\ &= \int\int_{B_\varepsilon(s_1)} f_V(s_2)\,\mathrm{d}s_2\, f_V(s_1)\,\mathrm{d}s_1 + \mathrm{o}(\varepsilon^{d_V}) \\ &= (2\varepsilon)^{d_V} \int f_V^2(s)\,\mathrm{d}s + \mathrm{o}(\varepsilon^{d_V}) \\ &= (2\varepsilon)^{d_V} H_V + \mathrm{o}(\varepsilon^{d_V}), \end{aligned} \tag{7}$$

where $B_\varepsilon(s_1)$ denotes a ball (or, since we use the supremum norm, a hypercube) with radius ε centered at s_1. The constant $H_V \equiv \int f_V^2(s)\,\mathrm{d}s = E\left[f_V(V)\right]$ can be considered as a *concentration measure* of V. To illustrate this, consider a family of univariate pdfs with scale parameter θ, that is, $f_V(v;\theta) =$

$\theta^{-1}g(\theta^{-1}v)$ for some pdf $g(\cdot)$. One readily finds $\int f_V^2(s;\theta)\,\mathrm{d}s = \frac{1}{\theta}\int g^2(s)\,\mathrm{d}s = \frac{\text{cnst.}}{\theta}$, which shows that, in the univariate case, the concentration measure is inversely proportional to the scale parameter θ. For later convenience, for a pair of vector-valued random variables (V, Y) of possibly different dimensions, we also introduce the *conditional concentration* of the random variable V given $Y = y$, as $H_V(y) = \int f_{V|Y}^2(v|y)\,\mathrm{d}v = (\int f_{V,Y}^2(v,y)\,\mathrm{d}v)/f_Y^2(y)$.

By comparing the leading terms of the expansion in powers of ε in equations (2b) and (7), we find that

$$\frac{E[f_{X,Y,Z}(X,Y,Z)]}{E[f_Y(Y)]} = \frac{E[f_{X,Y}(X,Y)]}{E[f_Y(Y)]}\frac{E[f_{Y,Z}(Y,Z)]}{E[f_Y(Y)]}. \tag{8}$$

That is, for ε small, testing the equivalence of the ratios in (2a) amounts to testing (8) instead of the null hypothesis. Unless some additional conditions hold, this will typically not be equivalent to testing the null hypothesis. To see what these additional conditions are it is useful to rewrite (8) as follows. For the left-hand-side one can write

$$\begin{aligned}\frac{E\left[f_{X,Y,Z}(X,Y,Z)\right]}{E[f_Y(Y)]} &= \frac{E_Y\left[E_{X,Z|Y}[f_{X,Z|Y}(X,Z|Y)f(Y)]\right]}{E[f_Y(Y)]}\\ &= \int E_{X,Z|Y=y}[f_{X,Z|Y}(X,Z|y)]w(y)\,\mathrm{d}y\\ &= \int H_{X,Z}(y)w(y)\,\mathrm{d}y,\end{aligned}$$

where $w(y)$ is a weight function given by $w(y) = f_Y^2(y)/\int f_Y^2(s)\,\mathrm{d}s$. This brings about the fact that the ratio on the left-hand-side of (8) for small ε is proportional to a weighted average of the conditional concentration $H_{X,Z}(y)$, with weight function $w(y)$. In a similar fashion, for the terms on the right-hand-side one derives

$$\frac{E\left[f_{X,Y}(X,Y)\right]}{E[f_Y(Y)]} = \int H_X(y)w(y)\,\mathrm{d}y, \qquad \text{and} \qquad \frac{E\left[f_{Y,Z}(Y,Z)\right]}{E[f_Y(Y)]} = \int H_Z(y)w(y)\,\mathrm{d}y.$$

Under the null hypothesis, Z is conditionally independent of X given $Y = y$, so that $H_{X,Z}(y)$ is equal to $H_X(y)H_Z(y)$, for all y. It follows that the left- and right-hand-sides of (8) coincide under the null if and only if $\int H_X(y)H_Z(y)w(y)\,\mathrm{d}y - \int H_X(y)w(y)\,\mathrm{d}y \int H_Z(y)w(y)\,\mathrm{d}y = 0$, or

$$\mathrm{Cov}(H_X(S), H_Z(S)) = 0, \tag{9}$$

where S is a random variable with pdf $w(y)$. Only under specific conditions, such as either $H_X(y)$ or $H_Z(y)$ being independent on y, (9) holds under the null, and hence (2a) as ε tends to zero. Also if $H_X(y)$ and $H_Z(y)$ depend on y, (9) may hold, but this is an exception rather than the rule. Typically the covariance between the conditional concentrations of X and Z given Y will not vanish, inducing a bias in the HJ test for small ε.

Therefore, letting the bandwidth tend to zero with increasing sample size in the HJ test would not provide a theoretical solution to the problem of over- or under-rejection caused by positive or negative covariance of the concentration measures respectively. In simulations for a particular process and small to moderate sample sizes one can often identify a seemingly adequate rate for bandwidths vanishing according to $\varepsilon_n = Cn^{-\beta}$, for which the size of the HJ test remains close to nominal. However, this does not imply that using the HJ test with such a sample size dependent bandwidth is advisable in practice. The optimal choices for C and β may depend strongly on the data generating process, and our results show that asymptotically the HJ test for typical processes (those with non-vanishing covariance of concentrations of X and Y) is inconsistent.

The fact that the conditional concentration measures of $X_t^{l_X}$ and Y_{t+1} given $Y_t^{l_Y}$ affect the leading bias term poses severe restrictions on applicability to economic and financial time series in which conditional heteroskedasticity is usually present. Consequently there is a risk of over-rejection by the HJ test which can not be easily eliminated either by using (G)ARCH filtering, or by using a bandwidth that decreases with the sample size. To avoid this problem, in the next section we suggest a new test statistic for which a consistent test is obtained as ε tends to zero at the appropriate rate. The idea is to measure the dependence between X and Z given $Y = y_i$ locally for each y_i. By allowing for the bandwidth to decrease with the sample size, variations in the local (fixed Y) distributions of X and Z given Y are automatically taken into account by the test statistic.

4 A modified test statistic

In comparing equations (1b) and (8) it can be noticed that although (1b) holds point-wise for any triple (x, y, z) in the support of $f_{X,Y,Z}(x, y, z)$, (8) contains separate averages for the nominator and the denominator of (1b), which do not respect the fact that the y-values on the rhs of (1b) should be identical. Because (1b) holds point-wise, rather than (8), the null hypothesis implies

$$q_g \equiv E\left[\left(\frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \frac{f_{X,Y}(X,Y)}{f_Y(Y)}\frac{f_{Y,Z}(Y,Z)}{f_Y(Y)}\right) g(X,Y,Z)\right] = 0$$

where $g(x, y, z)$ is a positive weight function Under the null hypothesis the term within the round brackets vanishes, so that the expectation is zero. Although q_g is not positive definite, a one-sided test, rejecting when its estimated value is too large, in practice is often found to have larger power than a two-sided test. In tests for serial dependence Skaug and Tjøstheim (1993) report good performance of a closely related unconditional test statistic (their dependence measure I_4 is an unconditional version of our term in round brackets).

We have considered several possible choices of the weight function g, being (i) $g_1(x,y,z) = f_Y(y)$, (ii) $g_2(x,y,z) = f_Y^2(y)$ and (iii) $g_3(x,y,z) = f_Y(y)/f_{X,Y}(x,y)$. Monte Carlo simulations using the stationary bootstrap (Politis and Romano, 1994) indicated that g_1 and g_2 behave similarly and are more stable than g_3. We will focus on g_2 in this paper, as its main advantage of over g_1 is that the corresponding estimator has a representation as a U-statistic, allowing the asymptotic distribution to be derived analytically for weakly dependent data, thus eliminating the need of the computationally more requiring bootstrap procedure. For the choice $g(x,y,z) = f_Y^2(y)$, we refer to the corresponding functional simply as q:

$$q = E\left[f_{X,Y,Z}(X,Y,Z)f_Y(Y) - f_{X,Y}(X,Y)f_{Y,Z}(Y,Z)\right].$$

A natural estimator of q based on indicator functions is:

$$T_n(\varepsilon) = \frac{(2\varepsilon)^{-d_X-2d_Y-d_Z}}{n(n-1)(n-2)}\sum_i \left[\sum_{k,k\neq i}\sum_{j,j\neq i}\left(I_{ik}^{XYZ}I_{ij}^{Y} - I_{ik}^{XY}I_{ij}^{YZ}\right)\right],$$

where $I_{ij}^W = I(\|W_i - W_j\| < \varepsilon)$. Note that the terms with $k = j$ need not be excluded explicitly as these each contribute zero to the test statistic. The test statistic can be interpreted as an average over local BDS test statistics (see Brock *et al.*, 1996), for the conditional distribution of X and Z, given $Y = y_i$.

If we denote local density estimators of a d_W-variate random vector W at W_i by

$$\widehat{f}_W(W_i) = \frac{(2\varepsilon)^{-d_W}}{n-1}\sum_{j,j\neq i} I_{ij}^W,$$

the test statistic simplifies to

$$T_n(\varepsilon) = \frac{(n-1)}{n(n-2)}\sum_i \left(\widehat{f}_{X,Y,Z}(X_i,Y_i,Z_i)\widehat{f}_Y(Y_i) - \widehat{f}_{X,Y}(X_i,Y_i)\widehat{f}_{Y,Z}(Y_i,Z_i)\right).$$

For an appropriate sequence ε_n of bandwidth values these estimators are consistent and the test statistic consist of a weighted average of local contributions $\widehat{f}_{X,Y,Z}(x,y,z)\widehat{f}_Y(y) - \widehat{f}_{X,Y}(x,y)\widehat{f}_{Y,Z}(y,z)$ which tend to zero in probability under the null hypothesis.

In Appendix A.1, using the approach proposed by Powell and Stoker (1996), we show that for $d_X = d_Y = d_Z = 1$ the test is consistent if we let the bandwidth depend on the sample size as

$$\varepsilon_n = Cn^{-\beta} \tag{10}$$

for any positive constant C and $\beta \in (\frac{1}{4}, \frac{1}{3})$. In that case the test statistic is asymptotically normally distributed in the absence of dependence between the vectors W_i. Under suitable mixing conditions (Denker and Keller, 1983) this can be extended to a time series context provided that covariances between the local density estimators are taken into account, giving:

Theorem 1 *For a sequence of bandwidths ε_n given by (10) with $C > 0$ and $\beta \in (\frac{1}{4}, \frac{1}{3})$ the test statistic T_n satisfies:*

$$\sqrt{n}\frac{(T_n(\varepsilon_n) - q)}{S_n} \xrightarrow{d} N(0, 1).$$

In appendix A.1 the asymptotic normality of T_n is shown under a decreasing bandwidth, while appendix A.3 considers the autocorrelation robust estimation of the asymptotic variance σ^2 by S_n^2.

4.1 Bandwidth choice

In the typical case where the local bias tends to zero at the rate ε^2 as in Condition 1 in Appendix A.1, the bandwidth choice which is optimal in that it asymptotically gives the estimator T_n with the smallest mean squared error (MSE) is given by

$$\varepsilon_n^* = C^* n^{-\frac{2}{7}}$$

with

$$C^* = \left(\frac{18 \cdot 3q_2}{4(E[s(W)])^2}\right)^{\frac{1}{7}} \tag{11}$$

as derived in appendix A.2.

To gain some insights into the order of magnitude of C^* it is helpful to calculate its value for some processes. Here we consider the ARCH process given in (3). The optimal C-value derived in the appendix is analytically hard to track since it involves the marginal distribution of the process. However, we can derive an approximate optimal value of C analytically by ignoring the deviation from normality of Y (an assumption which is reasonable for small a). Taking $Y \sim N(0, 1)$ and X, Z independent and $N(0, 1 + aY^2)$ conditional on Y, we find

$$q_2 = \frac{\mathrm{e}^{2/a}\mathrm{erfc}(\sqrt{2/a})}{1152\pi^2\sqrt{a}}, \tag{12}$$

where $\mathrm{erfc}(s) = 1 - \mathrm{erf}(s)$ and

$$E[s(W)] = \frac{\sqrt{6a/\pi}(3 + a) + (a(a - 6) - 9)\mathrm{e}^{3/(2a)}\mathrm{erfc}(\sqrt{3/(2a)})}{768\sqrt{2}a^{3/2}\pi^{3/2}}. \tag{13}$$

To investigate the behaviour of the bandwidth for small a, one may use the fact that

$$q_2 = \frac{1}{1152\sqrt{2}\pi^{3/2}} + \mathrm{o}(a) \qquad \text{and} \qquad E[s(W)] = a^2 \left(\frac{1}{288\sqrt{3}\pi^2} + \mathrm{o}(a) \right).$$

This suggests that as a tends to zero the (asymptotically) optimal bandwidth diverges at the rate $a^{-4/7}$. This is consistent with the fact that larger bandwidths are optimal for a smaller correlation between the conditional concentrations of X and Z given Y.

The optimal bandwidth for (G)ARCH filtered data depends on the correlation of the conditional concentrations after filtering, which may depend strongly on the underlying data generating process. However, the consistency of the test does not require filtering prior to testing, and it is possible to obtain a rough indication of the optimal bandwidth for raw returns. Since the covariance between conditional concentrations for bivariate financial time series are mainly due to ARCH/GARCH effects, eqs (12) and (13) can be used together with an estimate of the ARCH coefficient a to obtain a rough indication of the optimal constant C^* for applications to unfiltered financial returns data. To provide a feel for the order of magnitude: for $a = 0.4$ one finds $C^* \simeq 8$. Note that this value is asymptotically optimal and may lead to unrealistically large bandwidths for small n. In applications we therefore truncate the bandwidth by taking

$$\varepsilon_n = \max(Cn^{-2/7}, 1.5). \tag{14}$$

4.2 Simulations

We use numerical simulations to investigate the behavior of the proposed T_n test with the shrinking bandwidth given by (10). As the underlying process for the simulations we choose the process (3) considered before, a bivariate conditional heteroskedastic process with lag one dependence. The interest in this process is stipulated by its relevance to econometrics and financial time series. The null hypothesis $\{X_t\}$ is not Granger causing $\{Y_t\}$ is satisfied.

Table 1 reports the T_n test rejection rates (both size and power) for increasing series length n with n-dependent bandwidths ε_n given by (10), for a nominal size of 0.05. The size computations were based on the ARCH process (3) with coefficients $c = 1$, $a = 0.4$. For β we used the theoretically optimal rate of $\frac{2}{7}$, and we chose $C = 8.62$ which empirically turned out to give fast convergence of the size to the nominal value 0.05. This C-value is close to the approximate optimal asymptotic value $C^* \simeq 8$ for $a = 0.4$ reported above.

n	100	200	500	1 000	2 000	5 000	10 000	20 000	60 000
ε	1.50	1.50	1.50	1.20	1.00	0.76	0.62	0.51	0.37
size	0.022	0.033	0.052	0.052	0.051	0.050	0.050	0.052	0.053
power	0.073	0.155	0.411	0.661	0.900	0.998	1.000	1.000	1.000

Table 1: *Observed rejection rates (size and power) of the T_n test for bivariate ARCH process (3) as a function of the time series length n and decreasing bandwidth ε according to (14) (nominal size 0.05). Number of realisations:* 10 000 *for* $n < 60\,000$, *and* 3 000 *for* $n = 60\,000$.

To compute the power we took the same process and reversed the roles of $\{X_t\}$ and $\{Y_t\}$, so that the relation tested became: $\{Y_t\}$ is not Granger causing $\{X_t\}$. For the power calculations the coefficient a was reduced to 0.1 to make the simulations more informative (for higher a the power was one in nearly all cases). The power of the test increases with n, in accordance with the consistency of the test under the decreasing bandwidth procedure.

To provide some guidance for choosing critical p-values in practice for small sample sizes, Figure 2 shows some size-size plots for small n ranging over nominal sizes between 0 and 0.15.

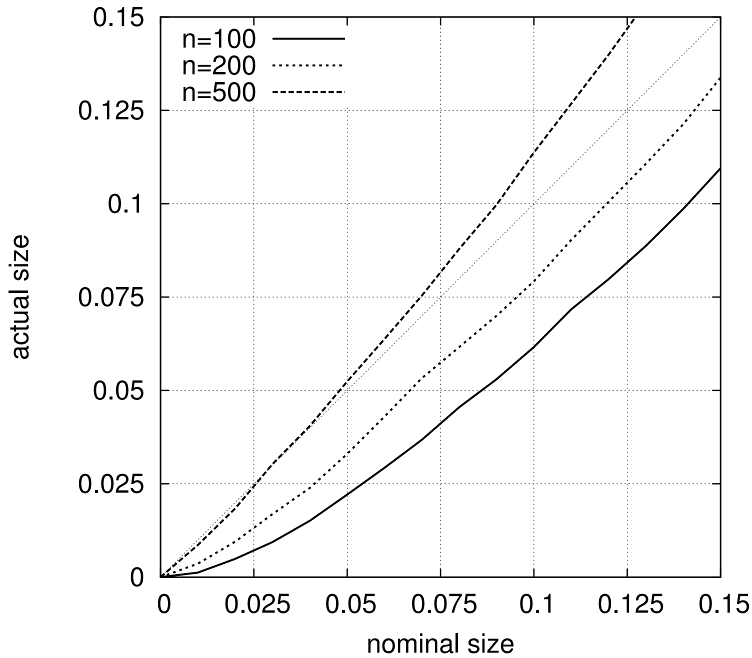

Figure 2: *Size-size plot of T_n test for process (3) with shrinking bandwidth for time series lengths $n = 100$ (solid line), 200 (dashed line), 500 (long-dashed line). The number of realisations is 10 000. The dotted line along the diagonal represents the ideal situation where the actual size and the nominal size coincide.*

Finally, we present some simulations for lags $l_X = l_Y$ larger than one, since these are used often for the HJ test. In the applications presented in the next section we compare both tests for larger values of l_X and l_Y as well, and to motivate this we should check if the empirical size of our new test does not exceed the nominal size for larger lags. Table 2 gives the empirical rejection rates for the bivariate ARCH process (3), again with $c = 1$ and $a = 0.4$, under the null hypothesis (that is, testing $\{X_t\}$ Granger causes $\{Y_t\}$) for lag lengths $l_X = l_Y$ ranging from 1 to 5. The results indicate that the rejection rate decreases with $l_X = l_Y$, and hence that the T_n test is progressively conservative for increasing lag lengths, so that the risk of rejecting under the null becomes small.

$l_X = l_Y$	$n = 1\ 000$	$n = 10\ 000$
1	0.0517	0.0502
2	0.0391	0.0316
3	0.0318	0.0197
4	0.0243	0.0112
5	0.0187	0.0099

Table 2: *Observed rejection rates (empirical size) of T_n test for bivariate ARCH process (3) as a function of number of lags $l_X = l_Y$ for time series length $n = 1\ 000$ and $n = 10\ 000$ with optimal bandwidth $\varepsilon = 1.2$ and $\varepsilon = 0.62$ respectively (nominal size 0.05, number of realisations $10\ 000$)*

5 Applications

We consider an application to daily volume and returns data for the Standard and Poor's 500 index in the period between January 1950 and December 1990. We deliberately have chosen this period to roughly correspond to the period for which Hiemstra and Jones (1994) found strong evidence for volume Granger-causing returns (1947 – 1990) for the Dow Jones index. To keep our results comparable with those of Hiemstra and Jones, we closely followed their procedure. That is, we adjusted for day-of-the-week and month-of-the-year effects on returns and percentage volume changes, using a two-step procedure in which we first adjust for effects in the mean, and subsequently in the variance. The calendar adjusted, standardized, returns and percentage volume change data were used to estimate a linear bivariate VAR model, the residuals of which are considered in the application below.

We applied the HJ and T_n test to the VAR residuals, before as well as after EGARCH(1,1) filtering the VAR residuals of the returns data. Table 3 shows the resulting T-values for the HJ and T_n test in both directions, for $l_X = l_Y = 1, \ldots, 8$ and for two different values of ε: 1.5, the value used by Hiemstra and Jones (1994) for the Dow Jones data, and 0.6, which is roughly the optimal value ($C^* \simeq 0.57$) we found from eqs (11)–(13) for the ARCH coefficient a, estimated from the data as 0.27.

	returns ⇒ volume				volume ⇒ returns			
	$\varepsilon = 1.5$		$\varepsilon = 0.6$		$\varepsilon = 1.5$		$\varepsilon = 0.6$	
$l_X = l_Y$	HJ	T_2	HJ	T_2	HJ	T_2	HJ	T_2
	before filtering							
1	9.476^{**}	9.415^{**}	10.298^{**}	8.850^{**}	5.351^{**}	5.106^{**}	5.736^{**}	4.893^{**}
2	10.989^{**}	11.076^{**}	10.616^{**}	8.182^{**}	6.671^{**}	6.447^{**}	6.818^{**}	5.396^{**}
3	10.909^{**}	10.662^{**}	9.112^{**}	6.425^{**}	6.026^{**}	5.683^{**}	5.717^{**}	3.948^{**}
4	10.758^{**}	9.823^{**}	7.934^{**}	5.121^{**}	6.029^{**}	5.552^{**}	4.692^{**}	2.887^{**}
5	10.118^{**}	8.856^{**}	5.821^{**}	3.540^{**}	5.695^{**}	5.191^{**}	2.837^{**}	1.234
6	9.428^{**}	7.903^{**}	4.391^{**}	2.603^{**}	5.935^{**}	5.338^{**}	3.314^{**}	1.604
7	8.959^{**}	7.4215^{**}	3.102^{**}	2.085^{*}	5.194^{**}	4.706^{**}	1.327	0.248
8	8.494^{**}	6.577^{**}	1.649^{*}	0.701	4.484^{**}	4.085^{**}	0.418	0.567
	after EGARCH filtering							
1	7.461^{**}	7.429^{**}	7.946^{**}	6.781^{**}	1.532	1.481	1.628	1.529
2	8.444^{**}	8.600^{**}	8.012^{**}	6.493^{**}	3.022^{**}	3.091^{**}	3.251^{**}	2.825^{**}
3	7.537^{**}	7.788^{**}	6.381^{**}	5.109^{**}	1.894^{*}	1.982^{*}	2.534^{**}	2.023^{*}
4	7.257^{**}	7.198^{**}	5.169^{**}	3.900^{**}	2.141^{*}	2.225^{*}	1.964^{*}	0.989
5	6.125^{**}	6.107^{**}	2.686^{**}	2.023^{*}	2.095^{*}	2.142^{*}	1.160	0.853
6	5.582^{**}	5.445^{**}	2.136^{*}	1.477	2.969^{**}	2.965^{**}	1.411	1.129
7	5.028^{**}	4.873^{**}	1.192	0.532	2.278^{*}	2.285^{*}	1.414	0.943
8	4.495^{**}	4.249^{**}	0.779	0.253	1.754^{*}	1.725^{*}	0.398	0.860

Table 3: *T-ratios for the S&P500 returns and volume data. Results are shown for the HJ test and T_n for bandwidth values of 1.5, the value used by Hiemstra and Jones (1994) and 0.6, corresponding to the optimal bandwidth for T_n (based on an estimated ARCH parameter 0.27). T-ratios before and after EGARCH filtering the returns are given, for $l_X = l_Y = 1, \ldots, 8$. The asterisks indicate significance at the 5% (*) and 1% (**) levels.*

The results obtained with both tests strongly indicate evidence for returns affecting future volume changes, for nearly all lags and both bandwidths. Only for large values of the lags $l_X = l_Y$ the evidence is somewhat weaker. Although both tests point in the same direction, when comparing the overall results for equal bandwidths and lags $l_X = l_Y$ the T-values are somewhat smaller for the T_n test than for the HJ test. As argued in the previous sections, the HJ test may be inconsistent due to a bias which cannot be removed simply by choosing a smaller bandwidth. To investigate the possible effects of this bias one should contrast the HJ test with our new test with an appropriately scaled bandwidth, which we have shown to be consistent asymptotically. That is, at least for the unfiltered data, one should actually compare the HJ test for $\varepsilon = 1.5$ with the T_n test for the adaptive bandwidth 0.6. In that case the table shows even larger differences between the T-values of the HJ test and the T_n test.

For the other causal direction — volume changes affecting future returns — the different results obtained for the HJ test with $\varepsilon = 1.5$ and the T_n test with $\varepsilon = 0.6$, for the filtered data is large enough to make a difference for obtaining significance at the 5% and 1% nominal level for several lags. Overall, the evidence for volume changes affecting future returns, although still present after filtering for lag $l_X = l_Y = 2$ and arguably 3, is much weaker for T_n with $\varepsilon = 0.6$ than for the HJ test with $\varepsilon = 1.5$.

In summary, our findings on the basis of the Standard and Poor's data indicate that the strong evidence for volume Granger causing returns obtained with the HJ test may be partly due to the bias we identified in the HJ test statistic. If the test is performed with the consistent T_n statistic with a near-optimal bandwidth, for which theory and simulations indicate that the actual size is close to nominal, the evidence for volume Granger causing returns tends to become weaker. Finally, since the T-values can be seen to decrease for smaller ε in most cases, the results also suggest that, when in doubt, it is better to use a smaller bandwidth. Intuitively this is related to the fact that it reduces the bias and increases the variance of the test statistic relative to the bias, so that the risk of over-rejection becomes smaller.

6 Concluding Remarks

Motivated by the fact that the HJ test can over-reject, as demonstrated in simulations, our aim was to construct a new test for Granger non-causality. By analyzing the HJ test analytically we found it to be biased even if the bandwidth tends to zero. Based on the analytic results, which indicated that the bias is caused by covariances in conditional concentrations, we proposed a new test statistic T_n that automatically takes the variation in concentrations into account.

By symmetrizing the new test statistic, we expressed it as a U-statistic for which we developed asymptotic theory under bandwidth values that tend to zero with the sample size at appropriate rates. The theory allowed us to derive the optimal rate as well as the asymptotically optimal multiplicative factor for the bandwidth. For ARCH type processes the optimal bandwidth can be expressed in terms of the ARCH coefficient, which is useful for getting an indication of the order of bandwidth magnitude to be used in practice for financial returns data. Simulations for the new test confirmed that the size converges to the nominal size fast as the sample size increases. Additional simulations indicated that the test becomes conservative for larger lags taken into account by the test.

In an application to relative volume changes and returns for historic Standard and Poor's index data we found that some of the strong evidence for relative volume changes Granger causing returns obtained with the HJ test may be related to its bias, since use of the new test, which is shown to be consistent, strongly weakens the evidence against the null hypothesis. This result suggests that some of the rejections of the Granger non-causality hypothesis reported in the literature may be spurious.

A Appendix

A.1 Asymptotic distribution of T_n

The test statistic T_n can be written in terms of a U-statistic by symmetrization with respect to the three different indices. This gives

$$T_n(\varepsilon) = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k \neq i} K(W_i, W_j, W_k)$$

with $W_i = (X_i^{l_X}, Y_i^{l_Y}, Z_i)$, $i = 1, \ldots, n$ and

$$K(W_i, W_j, W_k) = \frac{(2\varepsilon)^{-d_X - 2d_Y - d_Z}}{6} \begin{pmatrix} \left(I_{ik}^{XYZ} I_{ij}^{Y} - I_{ik}^{XY} I_{ij}^{YZ}\right) + \left(I_{ij}^{XYZ} I_{ik}^{Y} - I_{ij}^{XY} I_{ik}^{YZ}\right) + \\ \left(I_{jk}^{XYZ} I_{ji}^{Y} - I_{jk}^{XY} I_{ji}^{YZ}\right) + \left(I_{ji}^{XYZ} I_{jk}^{Y} - I_{ji}^{XY} I_{jk}^{YZ}\right) + \\ \left(I_{ki}^{XYZ} I_{kj}^{Y} - I_{ki}^{XY} I_{kj}^{YZ}\right) + \left(I_{kj}^{XYZ} I_{ki}^{Y} - I_{kj}^{XY} I_{ki}^{YZ}\right) \end{pmatrix} \tag{15}$$

For a given bandwidth ε the test statistic T_n is a third order U-statistic. To develop asymptotic distribution theory under a shrinking bandwidth ε_n we closely follow the methodology proposed by Powell and Stoker (1996). Although their main goal was to derive MSE (mean squared error) optimal bandwidths for point estimators, it turns out that similar considerations can be used to derive rates for the bandwidth that provide consistency and asymptotic normality of T_n. We first treat the analytically simplest case of a random sample $\{W_i\}_{i=1}^n$, and deal with dependence later.

Because T_n is a U-statistic, its finite sample variance is given by (see e.g. Serfling, 1980):

$$\mathrm{Var}\,(T_n) = \frac{9}{n}\zeta_1 + \frac{18}{n^2}\zeta_2 + \frac{6}{n^3}\zeta_3 + \mathrm{o}\left(\frac{\zeta_1}{n} + \frac{\zeta_2}{n^2} + \frac{\zeta_3}{n^3}\right),$$

where

$$\begin{aligned} \zeta_1 &= \mathrm{Cov}(K(W_1, W_2, W_3), K(W_1, W_2', W_3')) = \mathrm{Var}(K_1(W_1)) \\ \zeta_2 &= \mathrm{Cov}(K(W_1, W_2, W_3), K(W_1, W_2, W_3') = \mathrm{Var}(K_2(W_1, W_2)) \\ \zeta_3 &= \mathrm{Var}(K(W_1, W_2, W_3)), \end{aligned}$$

with W_1, W_2, W_3, W_2' and W_3' all independent and identically distributed according to W. The functions $K_1(w_1)$ and $K_2(w_1, w_2)$ are given by $K_1(w_1) = E[K(w_1, W_2, W_3)]$ and $K_2(w_1, w_2) = E[K(w_1, w_2, W_3)]$.

Following Powell and Stoker (1996), define $r(w, \varepsilon) = K_1(w, \varepsilon)$ and $r_0(w) = \lim_{\varepsilon \to 0} r(w, \varepsilon)$. It can be verified that

$$\begin{aligned} r_0(w) &= \tfrac{2}{3}f_{X,Y,Z}(x,y,z)f_Y(y) + \tfrac{1}{3}f_Y^2(y)H_{X,Z}(y) - \tfrac{1}{3}f_{X,Y}(x,y)f_{Y,Z}(y,z) \\ &\quad -\tfrac{1}{3}f_{Y,Z}(y,z)f_Y(y)\int f_{X,Y}(x',y)f_{X,Y,Z}(x',y,z)dx' \\ &\quad -\tfrac{1}{3}f_{X,Y}(x,y)f_Y(y)\int f_{Y,Z}(y,z')f_{X,Y,Z}(x,y,z')dz'. \end{aligned}$$

For example, the fourth term on the right-hand-side follows from:

$$\begin{aligned} (2\varepsilon)^{-d_X-2d_Y-d_Z}E_{W_k}[I_{jk}^{XY}I_{ji}^{YZ}] &= \int f_{X,Y}(x_k,y_k)\delta_{x_j,y_j}(x_k,y_k)I_{ij}^{YZ}\,\mathrm{d}x_k\,\mathrm{d}y_k(2\varepsilon)^{-d_Y-d_Z} + \mathrm{o}(1) \\ &= f_{X,Y}(x_j,y_j)I_{ij}^{YZ}(2\varepsilon)^{-d_Y-d_Z} + \mathrm{o}(1), \end{aligned}$$

where $\delta_{v_0}(v)$ stands for the kronecker delta function, which can be thought of as the limiting pdf of a random variable with all mass at the point v_0, and

$$\begin{aligned} (2\varepsilon)^{-d_Y-d_Z}E_{W_j}[f_{X,Y}(x_j,y_j)I_{ij}^{YZ}] &= \int f_{X,Y}(x_j,y_j)\delta_{y_j,z_j}(y_i,z_i)f_{X,Y,Z}(x_j,y_j,z_j)\,\mathrm{d}x_j\,\mathrm{d}y_j\,\mathrm{d}z_j + \mathrm{o}(1) \\ &= \int f_{X,Y}(x_j,y_i)f_{X,Y,Z}(x_j,y_i,z_i)dx_j + \mathrm{o}(1). \end{aligned}$$

Adapting from Powell and Stoker (1996), we assume the following three conditions:

Condition 1: *(rate of convergence of pointwise bias of $r(w_i,\varepsilon)$). The functions $r(w_i,\varepsilon)$ satisfy*

$$r(w_i,\varepsilon) - r_0(w_i) = s(w_i)\varepsilon^\alpha + s^*(w_i,\varepsilon),$$

for some $\alpha > 0$, and the remainder term $s^*(\cdot)$ satisfies $E\|s^*(W_i,h)\|^2 = \mathrm{o}(h^{2\alpha})$.

For our kernel the bias in each of the contributions to the kernel converges to zero at rate $\alpha = 2$. Therefore Condition 1 holds with $\alpha = 2$. In fact it might be possible to replace the local bias Condition 1 by a global version, involving $E[r(W_i,\varepsilon) - r_0(W_i)]$, which may tend to zero faster than the local bias. However, for our purposes the local assumption with $\alpha = 2$ suffices.

Condition 2: *(series expansion for second moment of $K_2(W_1,W_2)$). The function $K_2(w_1,w_2)$ satisfies*

$$E\left[(K_2(W_1,W_2))^2\right] = q_2\varepsilon^{-\gamma} + q_2^*(\varepsilon)$$

for some $\gamma > 0$, where the remainder term q_2^* satisfies $(q_2^*(\varepsilon))^2 = \mathrm{o}(\varepsilon^{-\gamma})$.

This is a weaker version of Powell and Stoker's (1996) Assumption 2, which required a series expansion locally. For our purposes the weaker assumption suffices, since T_n is a global functional of the distribution of W.

Condition 3: *(series expansion for second moment of $K(W_1,W_2,W_3)$). The function $K(w_1,w_2,w_3)$ satisfies*

$$E\left[(K(W_1,W_2,W_3))^2\right] = q_3\varepsilon^{-\delta} + q_3^*(\varepsilon)$$

for some $\delta > 0$, where the remainder term q_3^* satisfies $(q_3^*(\varepsilon))^2 = \mathrm{o}(\varepsilon^{-\delta})$.

For our kernel Condition 3 is satisfied with $\delta = d_X + 2d_Y + d_Z$, since none of the contributions to the kernel have a variance increasing faster in ε than at the rate $\varepsilon^{d_X+2d_Y+d_Z}$. Finding an appropriate value for γ in Condition 2 is somewhat more involved. We examine the rate at which each of the contributions to the kernel function depend on ε. For example, for the term $(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ik}^{XYZ} I_{ij}^{Y}$ we find $E_{W_k}[(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ik}^{XYZ} I_{ij}^{Y}] = (2\varepsilon)^{-d_Y} f_{X,Y,Z}(X_i, Y_i, Z_i) I_{ij}^{Y} + \mathrm{o}(1)$ from which one obtains

$$
\begin{aligned}
E\left[\left((2\varepsilon)^{-d_Y} E_{W_k}\left[I_{ik}^{XYZ} I_{ij}^{Y}\right]\right)^2\right] &= (2\varepsilon)^{-2d_Y} E\left[f_{X,Y,Z}^2(X_i, Y_i, Z_i) I_{ij}^{Y} + \mathrm{o}(\varepsilon^{d_Y})\right] \\
&= (2\varepsilon)^{-d_Y} E\left[f_{X,Y,Z}^2(X_i, Y_i, Z_i) f_Y(Y_i)\right] + \mathrm{o}(\varepsilon^{-d_Y}).
\end{aligned}
$$

Proceeding in this way for each of the terms in the kernel, one finds that the dominant contributions are given by the terms $(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ij}^{XYZ} I_{ik}^{Y}$ and $(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ji}^{XYZ} I_{jk}^{Y}$. For the first of these one finds $E_{W_k}[(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ij}^{XYZ} I_{ik}^{Y}] = (2\varepsilon)^{-d_X-d_Y-d_Z} I_{ij}^{XYZ} f_Y(Y_i) + \mathrm{o}(1)$, giving

$$
\begin{aligned}
E\left[\left((2\varepsilon)^{-d_X-d_Y-d_Z} E_{W_k}\left[I_{ij}^{XYZ} I_{ik}^{Y}\right]\right)^2\right] &= (2\varepsilon)^{-2d_X-2d_Y-2d_Z} E\left[I_{ij}^{XYZ} f_Y^2(Y_i)\right] + \mathrm{o}(\varepsilon^{-d_X-d_Y-d_Z}) \\
&= (2\varepsilon)^{-d_X-d_Y-d_Z} E\left[f_{X,Y,Z}(X_i, Y_i, Z_i) f_Y^2(Y_i)\right] \\
&\quad + \mathrm{o}(\varepsilon^{-d_X-d_Y-d_Z}).
\end{aligned}
$$

All other terms increase with vanishing ε slower, which demonstrates that Condition 2 holds with $\gamma = d_X + d_Y + d_Z$ and a constant q_2 given by $q_2 = \frac{4}{36} \times 2^{-d_X-d_Y-d_Z} E[f_{X,Y,Z}(X_i, Y_i, Z_i) f_Y^2(Y_i)]$. The factor 4 enters due to the fact that there are two terms, $E_{W_k}[(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ij}^{XYZ} I_{ik}^{Y}]$ and $E_{W_k}[(2\varepsilon)^{-d_X-2d_Y-d_Z} I_{ij}^{XYZ} I_{jk}^{Y}]$, which are asymptotically perfectly correlated if ε tends to zero sufficiently slowly with the sample size.

It follows from Condition 1 that

$$
\mathrm{Var}\left[r(W_i, \varepsilon)\right] = \mathrm{Var}\left[r_0(W_i)\right] + C_0 \varepsilon^{\alpha} + \mathrm{o}(\varepsilon^{\alpha}),
$$

where $C_0 = 2\mathrm{Cov}\left[r_0(W_i), s(W_i)\right]$. We can thus express the mean squared error of T_n as

$$
\mathrm{MSE}[T_n] = \left(E[s(W_i)]\right)^2 \varepsilon^{2\alpha} + \frac{9}{n} C_0 \varepsilon^{\alpha} + \frac{9}{n} \mathrm{Var}\left[r_0(W_i)\right] + \frac{18}{n^2} q_2 \varepsilon^{-\gamma} + \frac{6}{n^3} q_3 \varepsilon^{-\delta}. \qquad (16)
$$

T_n is asymptotically $N(0, \sigma^2/n)$ distributed with $\sigma^2 = 9\mathrm{Var}[r_0(W_i)]$, provided that each of the ε-dependent terms in the MSE of T_n are $\mathrm{o}(n^{-1})$. If we let $\varepsilon \sim n^{-\beta}$, this implies the following four conditions should hold:

$$
-2\alpha\beta < -1, \qquad -\alpha\beta < 0, \qquad \gamma\beta < 1, \qquad \delta\beta < 2.
$$

The first two of these imply $\beta > \frac{1}{2\alpha} = \frac{1}{4}$ and $\beta > 0$, respectively, while the last two imply $\beta < \frac{1}{\gamma} = \frac{1}{d_X+d_Y+d_Z}$ and $\beta < \frac{2}{\delta} = \frac{2}{d_X+2d_Y+d_Z}$. Because $\frac{1}{d_X+d_Y+d_Z} < \frac{2}{d_X+2d_Y+d_Z}$, the conditions can be summarized as: $\frac{1}{4} < \beta < \frac{1}{d_X+d_Y+d_Z}$. Therefore, for the case $d_X = d_Y = d_Z = 1$, and a sequence of bandwidths $\varepsilon_n \sim n^{-\beta}$ for some $\beta \in (\frac{1}{4}, \frac{1}{3})$, the test statistic is asymptotically normal:

$$\sqrt{n}\frac{T_n(\varepsilon_n) - q}{\sigma} \xrightarrow{d} N(0,1)$$

with $\sigma^2 = 9\mathrm{Var}[r_0(W_i)]$.

Note that it might also be possible to derive appropriate values for the rate β for $d_X+d_Y+d_Z > 3$, but only provided that the overall bias $E[s(W_i)]$ tends to zero faster than ε^2.

A.2 Optimal Bandwidth

The MSE optimal bandwidth balances the dominating squared bias and variance terms (the first and fourth term on the right-hand-side of Eq. (16)), the other bandwidth dependent terms being of smaller order. The optimal bandwidth which asymptotically minimizes the sum of these terms is given by

$$\varepsilon^* = \left(\frac{18 \cdot 3q_2}{4(E[s(W)])^2}\right)^{\frac{1}{7}} n^{-\frac{2}{7}}. \tag{17}$$

To guide the choice of the multiplicative factor C in $\varepsilon = Cn^{-\frac{2}{7}}$, it is illustrative to examine the optimal choice $C^* = \left(\frac{18\cdot 3q_2}{4(E[s(W)])^2}\right)^{\frac{1}{7}}$ in specific cases. Above an expression for q_2 was found already in terms of the joint density of W. A similar expression for $E[s(W_i)]$ can be found by using local taylor expansions of the density of w, locally near w_k. As each of the 6 terms in T_n have the same expectation, to determine the bias we consider the first of these only:

$$(2\varepsilon)^{-d_X-2d_Y-d_Z}(I_{ik}^{XYZ} I_{ij}^{Y} - I_{ik}^{XY} I_{ij}^{YZ}).$$

Taking averages over j and k for a fixed vector w_i leads to an expression involving plug-in estimators of local densities

$$\frac{(2\varepsilon)^{-d_X-2d_Y-d_Z}}{(n-1)(n-2)}\sum_{j\neq i}\sum_{k\neq i,j}(I_{ik}^{XYZ} I_{ij}^{Y} - I_{ik}^{XY} I_{ij}^{YZ}) = \widehat{f}_{X,Y,Z}(x_i,y_i,z_i)\widehat{f}_Y(y_i) - \widehat{f}_{X,Y}(x_i,y_i)\widehat{f}_{Y,Z}(y_i,z_i).$$

An expression for the local bias can be obtained by examining the bias of each of the estimated densities in this expression.

For a general density $f_V(v)$ of a random vector $V = (V^1,\ldots,V^m)$, of which a sample $\{V_i\}_{i=1}^n$ is available, the bias of $\widehat{f}(\tilde{v}) = (2\varepsilon)^m \frac{1}{n}\sum_{i=1}^n I(\|V_i - \tilde{v}\| \leq \varepsilon)$ locally at $\tilde{v}$ can be found from a taylor expansion of the density of $f_V(v)$ around $\tilde{v}$:

$$f(v) - f(\tilde{v}) = \sum_{i=1}^{m} a_i(\tilde{v})(v^i - \tilde{v}^i) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij}(\tilde{v})(v^i - \tilde{v}^i)(v^j - \tilde{v}^j) + O(\|v - \tilde{v}\|^3),$$

with $a_i(\tilde{v}) = \frac{\partial}{\partial v^i}\big|_{v=\tilde{v}} f(v)$ and $b_{ij}(\tilde{v}) = \frac{\partial^2}{\partial v^i \partial v^j}\,|_{v=\tilde{v}} f(v)$. The local bias of $\widehat{f}(\tilde{v})$ is given by

$$\begin{aligned}
E[\widehat{f}(\tilde{v})] - f(\tilde{v}) &= \frac{1}{2}(2\varepsilon)^{-m}\sum_{i=1}^{m}\sum_{j=1}^{m}\int_{\tilde{v}^1-\varepsilon}^{\tilde{v}^1+\varepsilon}\cdots\int_{\tilde{v}^m-\varepsilon}^{\tilde{v}^m+\varepsilon} b_{ij}(\tilde{v})(v^i - \tilde{v}^i)(v^j - \tilde{v}^j)\,\mathrm{d}v^1\ldots\mathrm{d}v^m + \mathrm{o}(\varepsilon^2)\\
&= \frac{1}{2}(2\varepsilon)^{-1}\sum_{i=1}^{m}\int_{\tilde{v}^i-\varepsilon}^{\tilde{v}^i+\varepsilon} b_{ii}(\tilde{v})(v^i - \tilde{v}^i)^2 \mathrm{d}v^i + \mathrm{o}(\varepsilon^2)\\
&= (2\varepsilon)^{-1}\frac{1}{3}\varepsilon^3\sum_{i=1}^{m} b_{ii}(\tilde{v}) + \mathrm{o}(\varepsilon^2)\\
&= \frac{1}{6}\varepsilon^2\nabla^2 f(\tilde{v}) + \mathrm{o}(\varepsilon^2).
\end{aligned}$$

Up to leading order in ε, the bias of products of estimated densities follows from identities such as $E[\widehat{f}_V\widehat{f}_W] = E[(f_V + (\widehat{f}_V - f_V))(f_W + (\widehat{f}_W - f_W))] = f_V f_W + f_V E[\widehat{f}_W - f_W] + f_W E[\widehat{f}_V - f_V] + \mathrm{o}(\varepsilon^2)$. In this way the local bias of $\widehat{f}_{X,Y,Z}(x_i, y_i, z_i)\widehat{f}_Y(y_i) - \widehat{f}_{X,Y}(x_i, y_i)\widehat{f}_{Y,Z}(y_i, z_i)$ can be written as

$$\begin{aligned}
r(w_i, \varepsilon) - r_0(w_i) &= \tfrac{1}{6}\varepsilon^2\left[f_Y(y_i)\nabla^2 f_{X,Y,Z}(x_i, y_i, z_i) - f_{X,Y}(x_i, y_i)\nabla^2 f_{Y,Z}(y_i, z_i)\right.\\
&\quad \left. + f_{X,Y,Z}(x_i, y_i, z_i)\nabla^2 f_Y(y_i) - f_{Y,Z}(y_i, z_i)\nabla^2 f_{X,Y}(x_i, y_i)\right] + \mathrm{o}(\varepsilon^2),
\end{aligned} \tag{18}$$

which shows that Condition 1 holds for $\alpha = 2$ and $s(w)$ equal to one 6th of the term between square brackets. Suppressing the subscripts for convenience, one may write

$$\begin{aligned}
s(w) &= \tfrac{1}{6}f(y)[\nabla_x^2 f(x, y, z) + \nabla_y^2 f(x, y, z) + \nabla_z^2 f(x, y, z)]\\
&\quad -\tfrac{1}{6}f(x, y)[\nabla_y^2 f(y, z) + \nabla_z^2 f(y, z)]\\
&\quad -\tfrac{1}{6}f(y, z)[\nabla_x^2 f(x, y) + \nabla_y^2 f(x, y)]\\
&\quad +\tfrac{1}{6}f(x, y, z)\nabla_y^2 f(y),
\end{aligned}$$

where $\nabla_x^2 = \sum_{j=1}^{d_X}\frac{\partial^2}{\partial x^{j2}}$ and ∇_y^2 and ∇_z^2 are defined analogously. Upon taking expectations with respect to W one obtains the coefficient of the leading bias term $E[s(W)]$, which enters expression (17) for the optimal bandwidth.

Under the null hypothesis the leading bias term can be simplified by rewriting it in terms of conditional densities:

$$\begin{aligned} s(w) &= \tfrac{1}{6} f(y) f(x,z) \nabla_y^2 f(y|x,z) - \tfrac{1}{6} f(x,y) f(z) \nabla_y^2 f(y|z) \\ &\quad -\tfrac{1}{6} f(y,z) f(x) \nabla_y^2 f(y|x) + \tfrac{1}{6} f(x,y,z) \nabla_y^2 f(y) \\ &\quad +\tfrac{1}{6} f(y) f(y,z) \nabla_x^2 [f(x|y,z) - f(x|y)] \\ &\quad +\tfrac{1}{6} f(x,y) f(y) \nabla_z^2 [f(z|x,y) - f(z|y)]. \end{aligned}$$

The terms within square brackets are zero if the null hypothesis holds. The remaining terms can be expressed as:

$$\begin{aligned} s(w) &= \tfrac{1}{6} \nabla_y^2 [f(y) f(x,y,z) - f(x,y) f(y,z)] \\ &\quad -\tfrac{1}{3} \nabla_y f(y) \cdot \nabla_y f(x,y,z) + \tfrac{1}{3} \nabla_y f(x,y) \cdot \nabla_y f(y,z), \end{aligned}$$

where ∇_y is the gradient operator, and the dot denotes the usual vector inner product. Again the term in square brackets vanishes under the null, and the remaining terms reduce to

$$s(w) = \frac{1}{3} f^2(y) \nabla_y f(x|y) \cdot \nabla_y f(z|y).$$

Finally, the following expression for $E[s(W)]$ under the null is obtained by taking expectations of this local expression with respect to the random vector W:

$$E[s(W)] = \frac{1}{3} E_Y \left[f_Y^2(Y) \nabla H_X(Y) \cdot \nabla H_Z(Y) \right].$$

A.3 Dependence

According to Denker and Keller (1983), for weakly dependent data T_n is still asymptotically $N(q, \frac{\sigma^2}{n})$ distributed, provided that the covariance among the $r_0(W_i)$ is taken into account in the asymptotic variance σ^2:

$$\sigma^2 = 9 \left[\mathrm{Var}(r_0(W_1)) + 2 \sum_{k \geq 2} \mathrm{Cov}(r_0(W_1), r_0(W_{1+k})) \right].$$

If we estimate $r_0(W_i)$ as

$$\widehat{r}_0(W_i) = \frac{(2\varepsilon)^{-d_X - 2d_Y - d_Z}}{(n-1)(n-2)} \sum_{j, j \neq i} \sum_{k, k \neq i} K(W_i, W_j, W_k),$$

an autocorrelation consistent estimator for σ^2 is given by (Newey and West, 1987):

$$S_n^2 = \sum_{k=1}^{K} R_k \omega_k,$$

where $R_k = \frac{1}{n-k}\sum_{i=1}^{n-k}(\widehat{r}_0(W_i) - T_n)(\widehat{r}_0(W_{i+k}) - T_n)$ is the sample autocovariance function of $\widehat{r}_0(W_i)$, and ω_k a decreasing weight function as in Hiemstra and Jones (1994). It follows that

$$\sqrt{n}\frac{(T_n - q)}{S_n} \xrightarrow{d} N(0,1),$$

which proves Theorem 1.

Although T_n is a third order U-statistic, both T_n and the asymptotic variance S_n^2 can be determined in $\mathcal{O}(n^2)$ computational time. For each i, the calculation of $\widehat{f}_W(W_i)$ and the I_{ij}^W is $\mathcal{O}(n)$. A second $\mathcal{O}(n)$ calculation then provides $\widehat{r}_0(W_i)$ through

$$\begin{aligned}\widehat{r}_0(W_i) \quad = \quad & \tfrac{1}{3}\left(\widehat{f}_{X,Y,Z}(X_i,Y_i,Z_i)\widehat{f}_Y(Y_i) - \widehat{f}_{X,Y}(X_i,Y_i)\widehat{f}_{Y,Z}(Y_i,Z_i)\right) \\ & + \tfrac{1}{3n}\sum_j \Big(\widehat{f}_{X,Y,Z}(X_j,Y_j,Z_j)I_{ij}^Y(2\varepsilon)^{-d_Y} + I_{ij}^{XYZ}\widehat{f}_Y(Y_j)(2\varepsilon)^{-d_X-d_Y-d_Z} \\ & \qquad - \widehat{f}_{X,Y}(X_j,Y_j)I_{ij}^{YZ}(2\varepsilon)^{-d_Y-d_Z} - I_{ij}^{XY}\widehat{f}_{Y,Z}(Y_j,Z_j)(2\varepsilon)^{-d_X-d_Y}\Big),\end{aligned}$$

a result which follows from straightforward calculation from the definition of $\widehat{r}_0(W_i)$. C-code can be obtained from the authors upon request.

References

Abhyankar, A. (1998). Linear and non-linear granger causality: Evidence from the U.K. stock index futures market. *Journal of Futures Markets*, **18,** 519–540.

Baek, E. and Brock, W. (1992). A general test for granger causality: Bivariate model. Technical Report. Iowa State University and University of Wisconsin, Madison.

Bell, D., Kay, J. and Malley, J. (1996). A non-parametric approach to non-linear causality testing. *Economics Letters*, **51,** 7–18.

Brock, W. A., Dechert, W. D., Scheinkman, J. A. and B, LeBaron (1996). A test for independence based on the correlation dimension. *Econometrics Review*, **15,** number 3, 197–235.

Brooks, C. (1998). Predicting stock index volatility: Can market volume help? *Journal of Forecasting*, **17,** 59–80.

Chen, A.-S. and Lin, J. W. (2004). Cointegration and detectable linear and nonlinear causality: analysis using the London Metal Exchange lead contract. *Applied Economics*, **36,** 1157–1167.

Ciner, C. (2001). Energy shocks and financial markets: Nonlinear linkages. *Studies in Nonlinear Dynamics and Econometrics*, **5,** number 3, 203–212.

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

Diks, C. and Panchenko, V. (2005). A note on the Hiemstra-Jones test for Granger non-causality. *Studies in Nonlinear Dynamics and Econometrics*; forthcoming.

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

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37,** 424–438.

Hiemstra, C. and Jones, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance*, **49,** number 5, 1639–1664.

Ma, Y. and Kanas, A. (2000). Testing for a nonlinear relationship among fundamentals and the exchange rates in the erm. *Journal of International Money and Finance*, **19,** 135–152.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica,* **55,** 703–708.

Okunev, J., Wilson, P. and Zurbruegg, R. (2000). The causal relationship between real estate and stock markets. *Journal of Real Estate Finance and Economics*, **21,** number 3, 251–261.

Okunev, J., Wilson, P. and Zurbruegg, R. (2002). Relationships between Australian real estate and stock market prices - a case of market inefficiency. *Journal of Forecasting*, **21,** 181–192.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, **89,** 1303–1313.

Pompe, B. (1993). Measuring statistical dependences in time series. *Journal of Statistical Physics*, **73,** 587–610.

Powell, J. L. and Stoker, T. M. (1996). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, **75,** 219–316.

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

Silvapulla, P. and Moosa, I. A. (1999). The relationship between spot and futures prices: Evidence from the crude oil market. *Journal of Futures Markets*, **19,** 157–193.

Skaug, H. J. and Tjøstheim, D. (1993). Nonparametric tests of serial independence. In *Developments in time series analysis* (ed. T. Subba Rao), chapter 15. Chapman and Hall, London.

Su, L. and White, H. (2003). A nonparametric Hellinger metric test for conditional independence. Technical Report. Department of Economics, UCSD.