Detecting Granger Causality with a Nonparametric Information-based Statistic

Cees G. H. Diks[1 2] and Hao Fang [*1 2]

[1]CeNDEF, Amsterdam School of Economics, University of Amsterdam

[2]Tinbergen Institute

Abstract

Testing causal effects has attracted much attention in the domains of many disciplines since Granger's pioneering work. The recent literature shows an increasing interest in testing for Granger non-causality in a general sense by nonparametric evaluation of conditional dependence. This paper introduces a novel nonparametric test based on the first order Taylor expansion of an information theoretic measure: transfer entropy. The new test statistic is shown to have an information-based interpretation for Granger non-causality. The proposed test avoids the lack of power problem in the frequently-used test proposed by Diks and Panchenko (2006), which is not a consistent test under some alternatives. Attributed to the U-statistic representation, the asymptotic normality of our test statistic is achieved when all densities are estimated with appropriate sample-size dependent bandwidth. Simulation result confirms the usefulness of this test. Finally two applications to financial data of daily and intraday frequency conclude this paper.

Keywords: Granger causality, Nonparametric test, U-statistic, Financial time series, High frequency data

JEL codes: C12, C14, C58, G10

*Corresponding author: Center for Nonlinear Dynamics in Economics and Finance, Faculty of Economics and Business, University of Amsterdam, Roetersstraat 11, 1018WB, Amsterdam, The Netherlands. E-mail: h.fang@uva.nl

1 Introduction

Characterizing causal interactions between time series was a challenging issue until Granger (1969) brought forward the concept later known as Granger causality in his pioneering work. Since then, testing causal effect has attracted massive attention not only in Economics and Econometrics research, but also in the domains of neuroscience (Bressler and Seth (2011), Ding, Chen, and Bressler (2006)), biology (Guo, Ladroue, and Feng (2010)) and physics(Barrett, Barnett, and Seth (2010)) among others.

The vector autoregressive (VAR) modeling based test has become a popular methodology for decades, with repeated debates on its validity. As we see it, there are at least two critical problems with the parametric causality tests. First, basing on a classical linear VAR model, traditional Granger causality tests may overlook a significant nonlinear relationship between two processes. As Granger (1989) puts it, nonlinear models represent the proper way to model the real world which is 'almost certainly nonlinear'. Secondly, parametric causality testing approaches bear the risk of model mis-specification. The conclusion draw from a wrong regression model could be either misleading or lacking power. For example, Baek and Brock (1992) show a constructed example where nonlinear causal relations cannot be detected by traditional linear causality test.

A series of studies tried to relax the influence of parametric model assumptions and provide many nonparametric versions of Granger causality tests. Among all nonparametric approaches, the Hiemstra and Jones (1994) test is one of the most far-reaching tests. By modifying the Baek and Brock (1992) nonparametric methodology, Hiemstra and Jones firstly developed a nonparametric test to detect nonlinear causality in weakly dependent stochastic data. However, the Hiemstra-Jones test is suffering from an over-rejection problem, which why it was then modified by Diks and Panchenko (2006), who proposed a new test statistic (hereinafter referred to as DP test). Other alternative tests include additive models by Bell, Kay, and Malley (1996), the Hellinger distance measure by Su and White (2008), and empirical likelihood ratio based test by Su and White (2014).

The scope of this paper is to provide a novel test for Granger causality, based on an information theoretic notion *transfer entropy*, which was coined by Schreiber (2000). Transfer entropy is initially used to measure asymmetric information exchanged in a bivariate system. By using appropriate conditional densities, transfer entropy is able to distinguish information that is actually transferred from shared information due to common history. This property makes it attractive for detecting conditional dependence, i.e. Granger causality. We refer to Hlaváčková-Schindler, Paluš, Vejmelka, and Bhattacharya (2007), Amblard and Michel (2012) for detailed reviews of the relation between Granger causality and directed information theory.

The application of concepts from information theory into time series analysis has been proved difficult, though attractive, due to the lack of asymptotic theory. For example, Granger and Lin (1994) normalize entropy to detect serial dependence using critical values obtained by

simulation. Later Hong and White (2005) achieve asymptotic Normality for entropy measure, but the asymptotic only holds for a specific kernel function. Barnett and Bossomaier (2012) establish an asymptotic χ^2 distribution for transfer entropy estimator under finite Markov chain assumption. Establishing asymptotic distribution theory for fully nonparametric transfer entropy measure is challenging, if not impossible.

In this research, we propose a test statistic based on first order Taylor expansion of transfer entropy, which follows Normal distribution asymptotically. Instead of deriving the limiting distribution of the transfer entropy itself, we bypass the problem by testing an implication of the null hypothesis. Further we show that this new test statistic is closely related to the DP test, but with an appealing property of non-negative definiteness.

This paper is organized as follows. Section 2 first provides a short introduction to nonparametric test on Granger non-causality and DP test, followed by an example to show that DP test is impotent against some alternatives because of the improper usage of non-negativity. Afterwards, information theory, and the testing framework of series expansion on transfer entropy (hereafter TE) statistic is introduced. The close linkage of this novel test statistic with DP test would be present, and the asymptotical normality is further proved by an order three U-statistic representation. The optimal bandwidth selection rule is also discussed in section 2. Section 3 deals with Monte Carlo simulations. Three different data generating processes are considered, direct comparison of size and power is presented between the modified DP test and DP test. Section 4 considers two financial applications. In the first case, we apply our test on the stock volume and return data to make a direct comparison with DP test; in the second application, high frequency exchange rates of main currencies are tested. Finally, section 5 summaries.

2 A Transfer Entropy-Based Test Statistic for Granger non-Causality

2.1 Nonparametric Granger non-Causality Tests

This subsection provides some basic concepts and definitions for Granger causality, and the idea of nonparametric test on the conditional independence has been shown as well. In this paper, we restrict ourselves to the bivariate setting as it is the most-accepted implementation, although the generalization to multivariate density is possible in the context of transfer entropy measure, which is defined as the expectation of log-likelihood functions.

Intuitively, in a bivariate system $\{X_t, Y_t\}$, given two strictly stationary process $\{X_t\}$ and $\{Y_t\}$, $t \in \mathbb{Z}$, it is claimed that $\{X_t\}$ Granger causes $\{Y_t\}$ if current and past values of $\{X_t\}$ contain some additional information beyond current and past values of $\{Y_t\}$ on the prediction of future values of $\{Y_t\}$. The linear Granger causality tests based on a parametric VAR model can be seen as an example, where restrictions on conditional means, as a special case for conditional

distributions, are tested.

In a more general manner, nonparametric tests for the null of Granger causality could be rephrased in terms of conditional dependence between two series: $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if the distribution of $\{Y_t\}$ conditional on its own history is not the same as conditional on the histories of both $\{X_t\}$ and $\{Y_t\}$. If we denote the information set of $\{X_t\}$ and $\{Y_t\}$ until time $t-1$ by $\mathscr{F}_{Y,t-1}$ and $\mathscr{F}_{X,t-1}$, respectively, and use '$\sim$' to show equivalence in distribution, we may give a formal and general definition for Granger causality. For a strictly stationary bivariate system $\{X_t, Y_t\}$, $t \in \mathbb{Z}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if, given lags l_X, l_Y and k,

$$(Y_t, ..., Y_{t+k}) \mid (\ \mathscr{F}_{Y,t-l_Y},\ \mathscr{F}_{X,t-l_X}\ ) \ \nsim\ (Y_t, ..., Y_{t+k}) \mid \mathscr{F}_{Y,t-l_Y}.$$

In the absence of Granger causality, i.e. $(Y_t, ..., Y_{t+k}) \mid (\ \mathscr{F}_{Y,t-l_Y},\ \mathscr{F}_{X,t-l_X}\ ) \ \sim\ (Y_t, ..., Y_{t+k}) \mid \mathscr{F}_{Y,t-l_Y}$ has no influence on the distribution of future $\{Y_t\}$. This is also referred to as Granger non-causality and often expressed as conditional independence between $\{X_t\}$ and $\{Y_t\}$:

$$(Y_t, ..., Y_{t+k}) \ \perp\ (X_{t-1}, ..., X_{t-l_X}) \mid \mathscr{F}_{Y,t-l_Y}\ . \tag{1}$$

Granger non-causality, or eq. (1), lays the first stone for a nonparametric test without imposing any parametric assumptions about the data generating process or underlying distributions for $\{X_t\}$ and $\{Y_t\}$. We only assume two things here. First, $\{X_t, Y_t\}$ is a strictly stationary bivariate process. Second, the process has a finite memory, i.e. $p, q \ll \infty$. The second assumption is needed in a nonparametric setting to allow conditioning on the past, as a finite Markov property.

The null hypothesis of a Granger causality test is that H_0 *:* $\{X_t\}$ *is not a Granger cause of* $\{Y_t\}$, and Granger non-Causality under the null is statistically tested by investigating the process $\{X_t, Y_t\}$. Strictly speaking, we should refer to Granger non-causality test instead of the well-accepted name Granger causality test. But in this paper, we will not distinguish the latter from the former. Both terms are used to stand for the test on the conditional independence between $\{X_t\}$ and $\{Y_t\}$.

For simplicity, we limit ourselves to detect only one-step-ahead effect by setting $l_X = l_Y = 1$ and $k = 0$, which is the case of most practical interest. Further we define a three-variate vector W_t as $W_t = (X_t, Y_t, Z_t)$, where $Z_t = Y_{t+1}$; and $W = (X, Y, Z)$ is used when there is no danger of confusion. Within the bivariate setting, W is a three dimensional continuous vector. By using density functions $f(.)$ and given $l_X = l_Y = 1$ and $k = 0$, eq. (1) can be phrased as

$$H_0 : \quad \frac{f_{X,Y,Z}(x, y, z)}{f_Y(y)} = \frac{f_{Y,Z}(y, z)}{f_Y(y)} \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{2}$$

for (x, y, z) in the support of W, or equivalently as

$$H_0: \quad \frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} - \frac{f_{Y,Z}(y,z)}{f_Y(y)}\frac{f_{X,Y}(x,y)}{f_Y(y)} = 0 \tag{3}$$

for (x, y, z) in the support of W. A nonparametric test on Granger non-causality seeks to find statistical evidence of violation ofeq. (2) or eq. (3). There are many nonparametric measures available for this purpose, some are mentioned previously. However, as far as we know, the DP test, to be introduced below, is the only fully nonparametric test which is proved to have correct size under the null hypothesis of no Granger causality.

2.2 The DP Test

Hiemstra and Jones (1994) first propose to test on eq. (2) by calculating correlation integrals for each density and measuring the discrepancy between two sides of the equation. However their test has been shown by literatures to suffer from severe size distortion due to a simple fact that measuring each density separately needs not to deliver the same quantity implied by eq. (2). To overcome the problem, Diks and Panchenko (2006) suggests to use a conditional dependence measure by incorporating a local weighting function $g(x, y, z)$ and formulate eq. (3) in such a way:

$$E\left[\left(\frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \frac{f_{Y,Z}(Y,Z)}{f_Y(Y)}\frac{f_{X,Y}(X,Y)}{f_Y(Y)}\right) g(X,Y,Z)\right] = 0 \tag{4}$$

Under the null hypothesis of no Granger causality, the term within the round bracket vanish, and the expectation goes to zero. eq. (4) can be treated as an infinite moment restrictions, as commented by Diks and Wolski (2015). Although testing on eq. (4) instead of eq. (2) or eq. (3) may lead to a loss of power against some specific alternatives, there is also an advantage to do so. For example, the weighting function $g(x, y, z)$ is assumed to be $g(x, y, z) = f_Y^2(y)$ in DP test as it delivers a U-statistic representation of the corresponding estimator, which enables the analytically asymptotic distribution for the test statistic. In principal, other choices for $g(x, y, z)$ will also do as long as the test has satisfactory power against the alternatives. By plug in $g(x, y, z) = f_Y^2(y)$, DP tests on the implication of H_0:

$$H_0': \quad q \equiv E\left[f_{X,Y,Z}(X,Y,Z)f_Y(Y) - f_{X,Y}(X,Y)f_{Y,Z}(Y,Z)\right] = 0 \tag{5}$$

Given a local density estimator of a d_W-variate random vector W at W_i as

$$\hat{f}_W(W_i) = ((n-1)h)^{-d_W} \sum_{j,j\neq i}^{n} \mathbb{K}\left(\frac{W_i - W_j}{h}\right) \tag{6}$$

where $\mathbb{K}$ is a kernel function and h is the bandwidth, the DP test develops a third order

U-statistic estimator for the functional q:

$$T_n(h) = \frac{(n-1)}{n(n-2)} \sum_i \left(\hat{f}_{X,Y,Z}(X_i, Y_i, Z_i)\hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i, Y_i)\hat{f}_{Y,Z}(Y_i, Z_i) \right) \tag{7}$$

where the normalization factor $(n-1)/(n(n-2))$ is inherited from the U-statistic representation of $T_n(h)$. It is worthwhile to mention that a second order uniform kernel $\mathbb{K}$ is adopted in Diks and Panchenko (2006), however there are two main drawbacks of uniform kernel as we see it. First, uniform kernel will yield a discontinuous density estimate $\hat{f}(\cdot)$, which is not attractive from practice perspective. Second, it weights all neighbor points W_j equally, overlooking their relative distance to the estimation point W_i. Therefore, a smooth kernel function, Gaussian kernel, is applied in this paper instead. Namely, $\mathbb{K}(.)$ is a product kernel function defined as $\mathbb{K}(W) = \prod_{s=1}^{d_W} \kappa(w^s)$, where w^s is s^{th} element in W. Using standard univariate Gaussian kernel, $\kappa(w^s) = (2\pi)^{-1/2} e^{-\frac{1}{2}(w^s)^2}$, $\mathbb{K}(.)$ is the standard multivariate Gaussian kernel as described in Wand and Jones (1994) and Silverman (1986).

For $l_X = l_Y = 1$, Diks and Panchenko (2006) proves the asymptotic normality of $T_n(h)$. Namely, if the bandwidth h depends on the sample size in such way that $h = Cn^{-\beta}$ for $C > 0$ and $\beta \in (\frac{1}{4}, \frac{1}{3})$, the test statistic in eq. (7) satisfies:

$$\sqrt{n}\frac{T_n(h) - q}{S_n} \xrightarrow{d} N(0,1) \tag{8}$$

where $\xrightarrow{d}$ denotes convergence in probability and S_n^2 is a consistent estimator of the asymptotic variance of $T_n(h)$. Diks and Panchenko (2006) suggests to implement a one-sided version of the test, although q is not (almost) positive definite, rejecting H_0' against the alternative $H_a : q > 0$ if eq. (7) is too large. In other words, given the asymptotic critical value $z_{1-\alpha}$, the null hypothesis H_0' is rejected if $\sqrt{n}(T_n(h) - q)/S_n > z_{1-\alpha}$ at significance level α.

The crucial defect of DP test arise from the fact that H_0' in eq. (5) need not to be identical to H_0 in eq. (2) and eq. (3). If H_0' is equivalent to H_0, then q forms a suitable basis for DP test on Granger non-causality. However this equivalence depends on a subtle property of q:

Property 1. *q is non-negative in such a way that $q \geq 0$ with equality if and only if X_t and Z_t are conditionally independent given Y_t.*

From previous reasoning, it is obvious that eq. (5) is implied by eq. (3), and Prop. 1 states that a strictly positive q is achieved only if H_0 is violated. In other words, the null hypothesis of Granger non-causality requites that X_t and Z_t are independent conditional on Y_t, which is just a sufficient, but not necessary, condition for $q = 0$. With Prop. 1, H_0' coincides with H_0 and a consistent estimator of q, i.e. $T_n(h)$ as suggested by DP, will have unit asymptotic power. If this property is not satisfied, a test on $q = 0$ could deviate from the test on H_0. Although Diks and Wolski (2015) proves for some specific classes of processes that the property holds, we could easily construct a counterexample to show that DP test has no power even X_t strongly

Granger causes Z_t.

Inspired by the example in Skaug and Tjøstheim (1993) where a closely related test for unconditional independence are performed, we may come up with its conditional counterpart to illustrate that q is not positive definite. Hence the one-sided DP test will suffer from lacking power as a consequence. As we show below, in an extreme case when $q = 0$, this drawback cannot be overcome even with a two sided DP test.

Consider the process $\{X_t, Y_t, Z_t\}$ where $Z_t \equiv Y_{t+1}$ as before. We assume that $X_t \in [-1, 1]$ $i.i.d$, with probability $1-d$ being positive, where $0 < d < 1$. Further, there are no instantaneous dependence between X_t and Y_t; i.e. Z_t does not depend on Y_t but on X_t in such a way that the joint density of (X_t, Z_t) is given by

$$p(X_t, Z_t) = \begin{cases} 1-2d, & \text{if } 0 \le X_t \le 1, \quad 0 \le Z_t \le 1 \\ d, & \text{if } 0 \le X_t \le 1, \ -1 \le Z_t < 0 \\ d, & \text{if } -1 \le X_t < 0, \quad 0 \le Z_t \le 1 \\ 0, & \text{if } -1 \le X_t < 0, \ -1 \le Z_t < 0 \end{cases} \tag{9}$$

Equivalently, the conditional density of Z_t on $\{X_t, Y_t\}$ is determined as $p(0 < Z_t \le 1|(X_t, Y_t)) = 1$ if $-1 \le X_t < 0$ and $p(0 \le Z_t \le 1|(X_t, Y_t)) = (1-2d)/(1-d)$ if $0 \le X_t \le 1$. In this setting, as long as $d \neq 0$, X_t is a Granger cause of Y_t since it has an impact on the distribution of future values of Y_t. Then it is not difficult to analytically calculate q defined in eq. (5). Particularly one finds that $q = d^2\left((1-d)^3 + d^3\right)(4d-1)$. For $0 < d < \frac{1}{4}$, q ends up with negative value. In this situation, the one-sided DP test, which rejects significantly positive q, is not a proper test on Granger non-causality. One may argue this is not a problem if we use a two-sided test at the price of losing some power. However, the severe inconsistence between DP test on H_0' and the initial purpose of testing H_0 is illustrated by considering $q = 0$, which is the case when $d = \frac{1}{4}$. Therefore X_t is clearly a Granger cause of Y_t, however DP test fails to detect this specific conditional independence.

Figure 1 reports the test power of the one-sided DP test against sample size at different significance levels, based on 10,000 independent replications. Three nominal sizes are illustrated here: 5%, 10% and 15%, and the sample size goes from 100 to 20,000. It is striking from fig. 1 that the DP does not have power at all to reject the alternative for this artificial process. Same conclusion can be made based on fig. 2, where the power-size plots are given. For almost all sub-panels with different sample sizes, the power of DP test is slightly lower if not close enough to the theoretical size for this particular example when $q = 0$, which indicate that DP test has only trivial power to detect Granger causality from X_t to Y_t.

[Figure 1 about here.]

[Figure 2 about here.]

The complete failure of the one-sided DP test in this example is hardly to be alleviated by its two-sided counterpart, as a result of the absence of equivalence between $q = 0$ and conditional independence. Without Prop. 1, the inconsistence between H_0' and H_0 gives rise to the defect of DP test. In the next subsection, a new test statistic based on an information theoretical concept, transfer entropy, is introduced and the test statistic is shown to be almost positive definite, which in return overcomes the inherited defect of DP test. In fact, this new test statistic shares many similarities with DP test statistic, but provided with a concrete interpretation for its non-negativity based on information theory.

2.3 Information Theory-Based Interpretation

In a very different field from econometrics research on conditional dependence, the problem of feedback and impact information theory also has drawn many attentions since 1950. Information theory, as a branch of applied mathematical theory of probability and statistics, studies the transmission of information over a noisy channel. Entropy, also refers to Shannon entropy, is one key measure in the field of information theory brought by Shannon (1948, 1951). Entropy measures the uncertainty and randomness associated with a random variable. Suppose S is a random vector with density $f_S(s)$, Shannon entropy is defined as:

$$H(s) = -\int f_S(s) \log\{f_S(s)\} ds.$$

There is a long history of applying information measures in econometrics. For example, Robinson (1991) applies Kullback and Leibler (1951) information criterion to construct a one-sided testing for serial independence. Since then, nonparametric testing using entropy measure for independence between two time series are becoming prevalent. Granger and Lin (1994) normalize entropy measure to identify the lags in a nonlinear bivariate model. Granger, Maasoumi, and Racine (2004) study dependence with a transformed metric entropy, which turns out to be a proper measure of distance. Hong and White (2005) provide a new entropy-based test for serial dependence, and the test statistic follows standard normal distribution asymptotically.

Although those research are heuristic, their methodologies cannot be applied directly to measure conditional dependence, i.e., Granger causality. Transfer entropy(TE) named by Schreiber (2000), although appeared in literatures earlier under different names, is a suitable measure to serve this purpose. TE quantifies the amount of information explained in one series at k steps ahead from the state of another series, given the current state of itself. It would be helpful to briefly introduce TE and Kullback-Leibler criterion before we further discuss its relation with the modified DP test.

Suppose we have two series $\{X_t\}$ and $\{Y_t\}$, for brevity put $X = \{X_t\}$, $Y = \{Y_t\}$ and $Z = \{Y_{t+k}\}$. Again we limit lag period $k = 1$ for simplicity, and define the three-dimensional vector $W = (X, Y, Z)$ same as before. Transfer entropy $TE_{X\to Y}$ is a nonlinear measure for the

amount of information explained in Z by X, accounting for the contribution of Y. Although TE defined by Schreiber (2000) applies for discrete variable, it is easily generalized to continuous variables. Conditional on Y, $TE_{X\to Y}$ is defined as

$$\begin{aligned} TE_{X\to Y} &= E_W\left(\log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}\right) \\ &= \int\int\int f_{X,Y,Z}(X,Y,Z)\log\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}dxdydz \\ &= E_W\left(\log\frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \log\frac{f_{X,Y}(X,Y)}{f_Y(Y)} - \log\frac{f_{Y,Z}(Y,Z)}{f_Y(Y)}\right) \\ &= E_W\left(\log f_{X,Y,Z}(X,Y,Z) + \log f_Y(Y) - \log f_{X,Y}(X,Y) - \log f_{Y,Z}(Y,Z)\right). \end{aligned} \tag{10}$$

Using conditional mutual information $I(Z,X|Y=y)$, TE can be equivalently formulated with four Shannon entropy terms:

$$\begin{aligned} TE_{X\to Y} &= I(Z,X|Y) \\ &= H(Z|Y) - H(Z|X,Y) \\ &= H(Z,Y) - H(Y) - H(Z,X,Y) + H(X,Y). \end{aligned}$$

In order to construct a test for Granger causality based on TE measure, one needs first to show quantitatively TE is a proper basis for detecting whether null hypothesis is satisfied. Namely, it has to be proved that there exists an analogue property for TE such as Prop. 1 for q in DP test. The following theorem, as a direct application of Kullback-Leibler criterion, lays the quantitative foundation for testing on TE.

Theorem 1. *$TE_{X\to Y} \geq 0$ with equality if and only if $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$.*

Proof. The proof to Thm. 1 can be given by generalizing Theorem 3.1 in Chapter 2 of Kullback (1968), where the divergence between two different densities has been measured. An alternative proof is given in appendix A.1 by using Jensen's inequality and concavity of log function. □

It is not difficult to verify that the condition for $TE_{X\to Y} = 0$ coincide with eqs. (2) and (3) defined for Granger non-causality under the null hypothesis. This non-negative property makes $TE_{X\to Y}$ a desirable statistic for constructing a one-sided test of Granger causality: any divergence from zero is a sign of conditional dependence of Y on X. To estimate TE, one may follow the recipe of Kraskov, Stögbauer, and Grassberger (2004) by measuring k-nearest neighbour distances. A more natural method, as we applied is this paper, is to use the plug-in kernel estimates eq. (6) of densities, and replace the expectation by sample average.

However, the direct use of TE to test Granger non-causality is doomed due to the lack of asymptotic theory for the test statistic. As Granger and Lin (1994) put it, very few asymptotic distribution results for entropy based estimator is available. Although over these years several break-throughs have been made with application of entropy to testing serial independence, like

Robinson (1991) obtained an asymptotic $N(0,1)$ distribution for an entropy measure by a data splitting device and Hong and White (2005) derived an asymptotic normal distribution under bounded support and quartic kernel assumptions, the limiting distribution of TE statistic is still in myth.

One may argue to use simulation techniques to overcome the problem of lacking asymptotic distribution. However, as suggested by Su and White (2008), there exists estimation biases of TE statistics for non-parametric dependence measuring under smoothed bootstrap procedure. Even with a parametric test statistics as defined in Barnett and Bossomaier (2012), the authors noticed that the TE based estimator is generally biased. Surrogate data is also applied widely by, among others, Wibral, Pampu, Priesemann, Siebenhühner, Seiwert, Lindner, Lizier, and Vicente (2013) and Papana, Kyrtsou, Kugiumtzis, and Diks (2016) to detect information transfers, though the time-shifted technique potentially may destroy the conditional dependence structure in the data set. Direct usage of TE for non-parametric Granger non-causality test is considered difficult, if not impossible. First order Taylor expansion on TE, on the other hand, provides a way-out to construct the asymptotic distribution of this meaningful information measure. In the next section, we will show that the first order Taylor expansion of TE is identical to a modified DP test statistic. This equivalence not only helps to circumvent the problem of asymptotic distribution for entropy based statistic, but also endows the modified DP statistics t non-negativity, which is missing in DP test statistic. We simply kill two birds with one stone.

The remaining part of this section will introduce the first order Taylor expansion of TE, and the non-negative definiteness of the statistic will be given afterwards. Starting with eq. (10), we perform the first order Taylor expansion locally at $TE_{X\to Y} = 0$:

$$\begin{aligned} TE_{X\to Y} &= E_W\left[\log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}\right] \\ &= E_W\left[\log\left(1+\left(\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}-1\right)\right)\right] \\ &= E_W\left[\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}-1\right] + \text{h.o.t.}, \end{aligned} \tag{11}$$

By ignoring higher order terms, we define the first order expansion as the test statistic $t = E_W\left[\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}-1\right]$. The following theorem proves that t inherits the non-negative definiteness of TE.

Theorem 2. *$t \geq 0$ with equality if and only if $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$.*

Proof. See appendix A.2 □

Thm. 2 indicates that the statistic t has the desirable property of non-negativity, which is absent for the original DP test. However, direct estimation of eq. (11) does not deliver a meaningful test statistic because the asymptotic distribution is missing. In the next subsection, by

incorporating a different weight function $v(x, y, z)$, a modified DP test statistics actually reproduces the first order Taylor expansion of transfer entropy measure, which provides theoretical interpretation for the directional information interaction between $\{X_t\}$ and $\{Y_t\}$.

2.4 A Modified DP Test

In comparing eq. (3) and eq. (4), it can be seen that the discrepancy between H_0' and H_0 arises from incorporation of the weighting function $g(x, y, z) = f_Y^2(y)$ to the null hypothesis. In principal, a different positive function $g(x, y, z)$ will also work, such as those discussed in Diks and Panchenko (2006). As long as the weighting function does not sabotage the U-statistic representation, the asymptotic Normal distributions is maintained. Particularly, we propose to modify DP test by dividing all terms in the expectation of eq. (5) by another weighting function $v(x, y, z)$ defined as $v(x, y, z) = (f_{X,Y}(x, y)f_{Y,Z}(y, z))$. Similar to the discussion in Section 2.2, we test on the implication of H_0 instead of the null directly, and the new implication of H_0 has the form of

$$H_0'': \quad E\left[(f_{X,Y,Z}(X,Y,Z)f_Y(Y) - f_{X,Y}(X,Y)f_{Y,Z}(Y,Z))\,\frac{1}{v(X,Y,Z)}\right] = 0 \tag{12}$$

One can also think of eq. (12) as a result of plugging in a different weighting function in eq. (4). By defining $g(x, y, z) = f_Y^2(y)/(f_{X,Y}(x, y))(f_Y, Z(y, z))$ instead of $g(x, y, z) = f_Y^2(y)$ suggested by DP, eq. (12) simplifies to

$$H_0'': \quad t \equiv E\left[\frac{f_{X,Y,Z}(X,Y,Z)f_Y(Y)}{f_{X,Y}(X,Y)f_{Y,Z}(Y,Z)} - 1\right] = 0, \tag{13}$$

which is equivalent to the first order Taylor expansion in eq. (11). To estimate this t, we propose to use the following statistic with density estimator defined in eq. (6):

$$T_n'(h) = \frac{(n-1)}{n(n-2)}\sum_i\left[\left(\hat{f}_{X,Y,Z}(X_i,Y_i,Z_i)\hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i,Y_i)\hat{f}_{Y,Z}(Y_i,Z_i)\right)\frac{1}{\hat{v}(X_i,Y_i,Z_i)}\right] \tag{14}$$

The reason not estimating t directly is that, with the sample statistic $T_n'(h)$, we could restore an order three U-statistic representation of t, similar to the one for DP test statistic, from where the asymptotic normality can be derived. Thm. 3 states this formal result. The proof of Thm. 3 relies on the following two lemmas.

Lemma 1. *Let $\{W_i = (X_i, Y_i, Z_i)\}, i \in N$ be a sequence of k-dimensional random variables with Lebesgue density f. For the estimation of f, we use the kernel estimator $\hat{f}$ with kernel function $\mathbb{K}(w)$ defined in eq. (6). If $f(w)$ is continuous at $w \in \mathbb{R}^k$ and $\mathbb{K}(w)$ is of bounded variation, then*

$$\sup_{w\in R^k}|\hat{f}(w) - f(w)| \to 0 \ a.s.,$$

provided with any of the following two conditions:

(A1) W_i *is an independent sequence and either*

$$\sum_{i=1}^{\infty} e^{-\gamma i h^{2k}}, \text{ for all } \gamma > 0$$

$$or\ (\frac{\log\log i}{i})^{1/2k} = o(h)$$

(A2) W_i *is* ϕ*-mixing,* $A_l(\phi) < \infty$ *(for definition of* $A_l(\phi)$ *see Sen et al. (1974) (2.1)) and*

$$\sum_{i=1}^{\infty} (\frac{\gamma}{h^k i^{1/2}})^{2(l+1)} < \infty, \text{ for all } \gamma \in R_+.$$

Proof. Lemma 1 is Theorem 1 in Rüschendorf (1977) and the proof is given there. □

Lemma 1 shows the uniformly consistent with probability one for a class of kernel estimators of multivariate density functions. This is a generalization of the consistency result of the univariate density estimation of Nadaraya (1965) and Schuster (1969) to the multivariate case. Note that we impose a stronger consistency than the pointwise weak consistency. The pointwise convergency, which cannot assure global convergency after aggregation over each point, may not sufficient to serve our purpose here. We refer to Wegman (1972) and Wied and Weißbach (2012) for detailed discussion between different types of convergency. Based on lemma 1, we have not only local convergence as $\hat{f}(.) \to f(.)$ a.s. when $n \to \infty$, but also $\hat{v}(x, y, z) \to v(x, y, z)$ a.s. globally as a result.

Lemma 2. *Suppose* $\hat{v}(x_i, y_i, z_i)$ *is a consistent estimate of* $v(x_i, y_i, z_i)$ *a.s.,* T'_n *has the same limiting distribution as the one with* $\hat{v}(.)$ *replaced by* $v(.)$ *if the asymptotic distribution exists. Formally, if we denote the statistic containing a true* $v(.)$ *by* $\tilde{T}'_n$*, then we have*

$$(T'_n - \tilde{T}'_n) = o_P(\frac{1}{\sqrt{n}}).$$

Proof. See appendix A.3 □

Theorem 3. *Given the gaussian kernel, let the bandwidth tends to zero with* n *as* $h = Cn^{-\beta}, C > 0, \beta \in (\frac{1}{4}, \frac{1}{3})$*, the statistic* T'_n *satisfies*

$$\sqrt{n}\frac{T'_n(h) - t}{S_n} \xrightarrow{d} N(0, 1),$$

where S_n^2 *is a consistent estimator of the asymptotic variance* $\sigma^2 = 9\text{var}[\lim_{h\to\infty}(\tilde{K}_1(W_i, h)/v(W_i, h))]$.

Proof. See appendix A.4 □

Thm. 4 assures that the modified DP statistic $T'_n(h)$ is not degenerated thus may be used for performing a statistical test.

Theorem 4. *The limiting distribution of the modified DP test statistic is not degenerated.*

Proof. See appendix A.5 □

Final comment on eq. (14) goes to the treatment of marginals. Although our testing framework does not depend crucially on the restrict assumption of uniform distribution for the time series as in Pompe (1993) and Hong and White (2005), we recommend to use the probability integral transformation (PIT), the synonymy of uniform transformation, which usually improves the performance of statistical dependence test, as Diks and Panchenko (2006) suggest. The reason as we see it is that, contrary to direct testing on the original data, the bounded support after marginal transformation avoids calculating indefinite integral in the sense of bias and variance evaluation, which helps to stabilize the test statistic. There are more ways to transform the marginal variables into a bounded support than PIT, for example, by using Epanechnikov quadratic kernel. However for practical convenience, we would only apply PIT. The procedure is to transform the original series $\{X_t\}$ ($\{Y_t\}$) to $\{U_t^X\}$ ($\{U_t^Y\}$) such that $\{U_t^X\}$ ($\{U_t^Y\}$) is the empirical CDF of $\{X_t\}$ ($\{Y_t\}$) and the empirical distribution of $\{U_t^X\}$ ($\{U_t^Y\}$) is uniform. Assuming X_t and Y_t are continuous, the PIT transformation is almost everywhere invertible and the dependence structure between X_t and Y_t remains intact in the pair of U_t^X and U_t^Y. An extra merit of our test is that T_n' is constructed to be invariant under PIT transformation due to the division form; the DP test, on the other hand, is not invariant since marginal transformation has a direct impact on the value of T_n in eq. (7).

As proved by Thm. 2, the transfer entropy based statistic is typically positively deviate from zero under H_a. Thus tests based on the statistic $T_n'(h)$ can be implemented as one-sided test, namely, for a given significance level α, the null hypothesis is rejected if $\sqrt{n}(T_n'(h) - t)/S_n > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of standard Normal distribution. Although DP test also exploits one-sided significance level to achieve a higher power, as we argued before, the theoretical ground for its non-negativity is missing.

2.5 Bandwidth Selection

For nonparametric test, there is no uniformly most powerful test against all alternatives. Hence it is less likely that a uniformly optimal bandwidth exists. As long as the bandwidth tends to zero with as $h = Cn^{-\beta}, C > 0, \beta \in (\frac{1}{4}, \frac{1}{3})$, asymptotically our test has uniform power. Yet, we may define the optimal bandwidth in the sense of minimal asymptotical mean squared error (MSE). When balancing the first and forth leading terms in eq. (A.7) to minimize squared bias and variance, for second order kernel, it is easy to show that the optimal bandwidth for DP test is given by

$$h_{DP} = Cn^{-2/7}, \text{ where } C = \left(\frac{18 * 3q_2}{4(\mathrm{E}[s(W)])^2}\right)^{1/7}, \tag{15}$$

with q_2 and $\mathrm{E}[s(W)]$ the series expansion for second moment of kernel function and expectation of bias, respectively. Since the convergence rate of MSE, derived in appendix A.4, is not affect

by the way we construct the new test statistic, the derivation for eq. (15) remain intact and we simply calibrate the optimal bandwidth for our new test as

$$h^* \approx 0.6 h_{DP}, \tag{16}$$

where the scale factor 0.6 involved as a result of bias and variance adjustment for replacing Square kernel by Gaussian kernel. Intuitively, the q_2 and $\mathrm{E}[s(W)]$ terms are different from the ones in DP test, more details can be found in appendix A.6.

The optimal value for C is process-dependent and difficult to track analytically. For example, Diks and Panchenko (2006) demonstrated that for (G)ARCH process, the optimal bandwidth is approximately given by $h_{DP} = Cn^{-2/7}$ where $C \approx 8$. After calibration, we proceed with $h^* = 4.8n^{-2/7}$ for (G)ARCH process. To gain some insights about this bandwidth, we illustrate the test size and power with a 2-variate ARCH process

$$\begin{aligned} X_t &\sim N(0, 1 + aY_{t-1}^2) \\ Y_t &\sim N(0, 1 + aY_{t-1}^2). \end{aligned} \tag{17}$$

We let $0 < a < 0.4$ and run 5,000 Monte Carlo simulations for time series length varies from 200 to 5000. The size assessment is delivered based on testing Granger non-causality from $\{X_t\}$ to $\{Y_t\}$, and for the power we use the same process but testing on Granger non-causality from $\{Y_t\}$ to $\{X_t\}$. The result is presented in table 1, from where we see that the MDP test is conservative in the sense that empirical size is lower than the nominal 0.05 in all cases, while the power converges to unit when a increases or sample size goes larger.

[Table 1 about here.]

3 Monte Carlo Simulations

This section investigates the performance of the modified DP test. Before proceeding with new data generating processes, we first revisit the example illustrated in eq. (9) where DP test fails to detect any impact of X on Y. The modified DP test is performed with 10,000 replications again, with the same bandwidth. The counterpart of power-size plots for DP test in fig. 2 is delivered in fig. 3. Contrast with the impotence of DP test, for time series length $n = 500$ and larger, the modified DP test already has a very high power in this artificial experiment, as expected.

[Figure 3 about here.]

Next, we use numerical simulations to study the behavior of modified DP test, while direct comparisons between the modified DP test T_n' with the DP test T_n are also given. Three

processes have been considered. In the first experiment, we consider a simple bivariate VAR process:

$$\begin{aligned} X_t &= aY_{t-1} + \varepsilon_{x,t},\ \ \varepsilon_{x,t} \sim N(0,1) \\ Y_t &= aY_{t-1} + \varepsilon_{y,t},\ \ \varepsilon_{y,t} \sim N(0,1). \end{aligned} \tag{18}$$

[Figure 4 about here.]

The second experiment is designed as a nonlinear VAR process in eq. (19). Again the size and power are investigated by testing on Granger non-causality from two different directions.

$$\begin{aligned} X_t &= 0.6X_{t-1} + aX_{t-1}Y_{t-1} + \varepsilon_{x,t},\ \ \varepsilon_{x,t} \sim N(0,1) \\ Y_t &= 0.6Y_{t-1} + \varepsilon_{y,t},\ \ \varepsilon_{y,t} \sim N(0,1). \end{aligned} \tag{19}$$

[Figure 5 about here.]

The last experiment is same as the example we used for illustrating the performance of the bandwidth selection rule, which is a bivariate ARCH process as in eq. (17):

$$\begin{aligned} X_t &\sim N(0, 1 + aY_{t-1}^2) \\ Y_t &\sim N(0, 1 + aY_{t-1}^2). \end{aligned} \tag{20}$$

[Figure 6 about here.]

Results in figs. 4 to 6 are obtained with 5,000 iterations for each simulation. We present DP test and modified DP test with both empirical size-size plot and size-power for three processes in eqs. (18) to (20) for sample size $n = 500$ and $n = 5000$, respectively. The controlling parameter a is considered to take values 0.1 and 0.4. As before, the empirical size is obtained on testing Granger non-causality from $\{X_t\}$ to $\{Y_t\}$, and the empirical power is the rejection rate of test on Granger non-causality from $\{Y_t\}$ to $\{X_t\}$.

It can be seen from the figures that the modified DP test is slightly more conservative than DP test under the null hypotheses. However the size distortion is relived when sample size increases. The modified DP test in eqs. (18) and (19) is more powerful than DP test in the linear and nonlinear VAR settings. Overall, we see that the larger sample size and stronger causal effect, the better asymptotic performance for modified DP test.

4 Empirical Research

4.1 Stock Volume and Return Relation

In this section, we first re-visit the stock return-volume relation as Hiemstra and Jones (1994) and Diks and Panchenko (2006) did. There has been a long research history on this topic. Early

empirical works mainly focused on the positive correlation between volume and stock price change, see Karpoff (1987). Later literatures exposed the directional relations, for example, Gallant, Rossi, and Tauchen (1992) found that large price movements are followed by high volume; Gervais, Kaniel, and Mingelgrin (2001) observed the high-volume return premium, namely, periods of extremely high (low) volume tend to be followed by positive (negative) excess returns. More recently, Podobnik, Horvatic, Petersen, and Stanley (2009) investigated the power law cross-correlations between price changes and volume changes of the S&P 500 Index over a long period.

We use daily volume and returns data for the three most-followed indices in US stock markets, the Standard and Poor's 500 (S&P), the NASDAQ Composite (NASDAQ) and the Dow Jones Industrial Average (DJIA), between January 1985 and October 2016. Yahoo Finance is the source of the daily volume and adjusted daily closing prices. All data is converted by taking logs and log returns are multiplied by 100. In order to adjust for the day-of-the-week and month-of-the-year seasonal effects in both mean and variance of stock returns and volumes, we perform a two stage-adjustment process, similar to the procedure applied in Hiemstra and Jones (1994)[1]. We apply our test on not only the raw data, but also the VAR filtered residuals and the EGARCH(1,1,1) filtered residuals.[2] The idea of filtration is to remove linear dependence and the effect of heteroscedasticity to test on the nonlinear, high-moment relationships among series.

Tables 2 to 4 report the resulting T statistics for both DP test and our modified DP test in both directions. The linear Granger F-values based on the optimal VAR models are also given. Two bandwidth values are used: 1.5 and 0.6, where the latter one 0.6 roughly corresponds to the optimal bandwidth ($h = 0.6138$) and the larger bandwidth 1.5 also used in Diks and Panchenko (2006) checks for the robustness.

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

Generally speaking, the results indicate the effect in the return-volume direction is stronger than the other way around. For the test results on the raw data, the F-tests based on the linear VAR model and both nonparametric tests suggest evidence of Return affecting Volume for all three Indexes. For the other direction, causality from Volume to Return, linear Granger test has no power at all while the nonparametric tests claim strong causal effect except for the

[1]We replace Akaike's information criterion used by Hiemstra and Jones (1994) with Schwarz et al. (1978) information criterion to be more stringent on picking up variables. Having no intention to provoke the debate over the two criteria, we simply prefer a more parsimonious liner model to avoid potential overfitting.

[2]We have tried different error distributions like Normal, Students't, GED and Hansen (1994)'s skewed t. The difference caused by different distribution assumptions are ignorable. Thus we only report for the result based on Students't for simplicity.

DJIA case where only the modified DP test sticks to the affection of Volume on Return. As argued before, the linear test is doomed since it only examines linear causal effect in the mean, information exchange from higher moments are completely ignored.

The direct comparison between the DP test and the modified DP test shows that the new test is more powerful overall. For the unfiltered data, both tests find strong causal effect in two directions for S&P and NASDAQ, but for DJIA series, the t-statistics of DP test are weaker than that of the modified DP test. The bi-directional causality between Return and Volume remain unchanged after linear VAR filtration, though DP test again shows weaker evidence. The result also suggests that the causality is strictly nonlinear. The linear test (F-test) either overlooks the nonlinear inkages, or is unable to spot its nonlinear nature.

Further, in the direction of Volume to Return, these nonlinear causality tend to vanish after EGARCH filtering. Thus the bi-directional linkage is reduced to an one-directional relation from Return to Volume. The modified DP statistics, however, are in general larger than that of DP t-values, and finds more causal relations than DP test does. On the contrary to DP test, our test suggests that the nonlinear causality cannot be completely attributed to the second moment effect. Heteroscedasticity modeling may reduce the nonlinear feature to some extent, but its impact is not as strong as DP test claims. As argued in previous sections, the DP test may suffer from the problem of lacking power due to its vulnerable non-negativity property.

4.2 Application to Intraday Exchange Rate

In the second application, we apply the modified DP test to intraday exchange rates. We consider five major currency: JPY/USD, AUD/USD, GBP/USD, EUR/USD and CHF/USD. The data, obtained from Dukascopy Historical Data Feed, contain 5-minute bid and ask quotes for the third quarter of 2016; from July 1 to September 30, with a total of 92 trading days and 26496 high frequency observations. We use 5-minute data, corresponding to the sampling frequency of 288 times per day, which is high enough to avoid measurement errors (see Andersen and Bollerslev (1998)) but also low enough such that the microstructure is not the major concern.

Although exchange market is one of the most active financial market in the world, where trading takes 24 hours a day, the intraday trading is not always active. Thus we delete the thin trading period, from Friday 21:00 GMT until Sunday 20:55 GMT, also to keep the intraday periodicity intact. We calculate the exchange rate returns as in Diebold, Hahn, and Tay (1999). First the average log bid and log ask prices are calculated, then the difference between the log prices at consecutive time stamps are obtained. Next, we remove the conditional mean dynamics by fitting a MA(1) model and using the residuals as our return series following Bollerslev and Domowitz (1993). Finally the intraday seasonal effect is filtered out by using the estimated

time-of-day dummies in a way similar to Diebold, Hahn, and Tay (1999), i.e.

$$r_{i,n,t} = d_{i,t} z_{i,n,t}. \tag{21}$$

Here $r_{i,n,t}$ denotes the continuously intraday log returns after MA(1) filtration. Subscript $i = 1, ..., 5$ indicates five different currency and n, t stands for time t in day n. The first component of return series $d_{i,t}$ refers to a deterministic intraday seasonal component while $z_{i,n,t}$ is the nonseasonal return portion, which assumed to be independent of $d_{i,t}$. To distinguish $d_{i,t}$ from $z_{i,n,t}$, we fit the time-of-day dummies to $2 \log |r_{i,n,t}|$ and use the estimated $\hat{d}_{i,t}$ to standardize the return $r_{i,n,t}$ with the restriction $\sum_{t=1}^{T} d_{i,t} = 1$. Figures 7 to 9 report the first 200 autocorrelations of returns, absolute returns and square returns, when checking on the raw series, MA(1) filtration residuals and EGARCH filtration residuals, respectively.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

We perform the pairwise nonparametric Granger causality test on the conditional mean and seasonal effect removed data, as well as on the standardized residuals after EGARCH(1,1,1) filtration. We use Hansen (1994)'s skewed t distribution to model the innovation terms. We choose bandwidth at 0.2768, by applying eq. (16).

[Table 5 about here.]

The testing results are shown in table 5 for both MA(1) demeaned and de-seasoned data, as well as EGARCH filtered data. Although not reported here, there exist strong bi-directional causality among all currency pairs on raw return data at 5-minute lag. These bi-directional causalities do not die out after removing MA(1) component and seasonal component. However, the observed information spillover is significantly weaken after the EGARCH filtration. When testing on the EGARCH variance standardized residuals, only few pairs still show a sign of a strong causal relation. Especially, the directional relation of EUR $\rightarrow$ CHF is the only one spot by both DP test and modified DP test at 1% level. An graphic representation is depicted in fig. 10, where clearly we see most causal links are gone after EGARCH filtration. The modified DP test exposes five uni-directional linkages on the EGARCH filtered returns at 5% level. EUR and GBP are the most important driving currencies. While DP test also admits the importance of JPY and particularly AUD which has bi-directional causality with JPY and GBP.

[Figure 10 about here.]

To sum up, we find the evidence of strong causalities among exchange returns on intraday high-frequency level. Each currency has predictive power on other currencies, implying the high co-movement in the international exchange market. Although those directional linkages are not affected by demeaning procedure, we may reduce most of them by taking the volatility dynamic into account. When filtering out heteroscedasticity by EGARCH estimation, there only exist few pairs containing spillover effect.

5 Summary and Conclusions

Borrowing the concept of transfer entropy from Information Theory, this paper develops a novel non-parametric test statistic for Granger non-causality. The asymptotical Normality is achieved by taking advantage of an order-three U-statistic representation initially applied in DP test. The modified DP statistic, however, overwhelms the DP statistic in at least the two aspects: firstly, non-negativity, while missing for DP statistic, is a natural property for our statistic, which paves the way for properly testing difference between conditional densities; secondly, the weighting function in our test is justified by the theoretical information representation, while the weighting function in DP test is an arbitrary selection. Simulation studies show that our modified DP test provides a reasonable size and power for different data generating processes. In the first application, direct comparison with DP test again suggests that the lacking of rejections in DP test may be suspicious, while the second application to high frequency exchange return data may help us better understand whether the spillover channel in exchange market arises from mean, variance or higher moments. These applications of different data frequency indicate possibilities for future empirical research. One possible extension to current work is incorporating more lags under multivariate setting, in which case the asymptotical theory is needed.

Acknowledgments

We would like to thank seminar participants at University of Amsterdam and Tinbergen Institute, as well as participants at the 10th International Conference on Computational and Financial Econometrics (Seville, December 09-11, 2016), the 25th Annual Symposium of the Society for Nonlinear Dynamics and Econometrics (Paris, Mar 30-31, 2017). The authors also wish to extend particular gratitude to Simon A. Broda and Chen Zhou for their suggestions on an early version of this paper. We also acknowledge the support of SURFsara for providing the environment of LISA system.

References

Amblard, P.-O., and O. J. Michel (2012): "The relation between granger causality and directed information theory: a review," *Entropy*, 15(1), 113–143.

Andersen, T. G., and T. Bollerslev (1998): "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International economic review*, pp. 885–905.

Baek, E. G., and W. A. Brock (1992): "A general test for nonlinear Granger causality: bivariate model," *Working paper, Iowa State University and University of Wisconsin, Madison*.

Barnett, L., and T. Bossomaier (2012): "Transfer entropy as a log-likelihood ratio," *Physical review letters*, 109(13), 138105.

Barrett, A. B., L. Barnett, and A. K. Seth (2010): "Multivariate Granger causality and generalized variance," *Physical Review E*, 81(4), 041907.

Bell, D., J. Kay, and J. Malley (1996): "A non-parametric approach to non-linear causality testing," *Economics Letters*, 51(1), 7–18.

Bollerslev, T., and I. Domowitz (1993): "Trading patterns and prices in the interbank foreign exchange market," *The Journal of Finance*, 48(4), 1421–1443.

Bressler, S. L., and A. K. Seth (2011): "Wiener–Granger causality: a well established methodology," *Neuroimage*, 58(2), 323–329.

Diebold, F. X., J. Hahn, and A. S. Tay (1999): "Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange," *Review of Economics and Statistics*, 81(4), 661–673.

Diks, C. (2009): "Nonparametric tests for independence," in *Encyclopedia of complexity and systems Science*, pp. 6252–6271. Springer.

Diks, C., and V. Panchenko (2006): "A new statistic and practical guidelines for non-parametric Granger causality testing," *Journal of Economic Dynamics and Control*, 30(9), 1647–1669.

Diks, C., and M. Wolski (2015): "Nonlinear granger causality: Guidelines for multivariate analysis," *Journal of Applied Econometrics*.

Ding, M., Y. Chen, and S. L. Bressler (2006): *Granger causality: basic theory and application to neuroscience*chap. 17, pp. 437–460. John Wiley & Sons.

Gallant, A. R., P. E. Rossi, and G. Tauchen (1992): "Stock prices and volume," *Review of Financial studies*, 5(2), 199–242.

Gervais, S., R. Kaniel, and D. H. Mingelgrin (2001): "The high-volume return premium," *The Journal of Finance*, 56(3), 877–919.

Granger, C., and J.-L. Lin (1994): "Using the mutual information coefficient to identify lags in nonlinear models," *Journal of time series analysis*, 15(4), 371–384.

Granger, C., E. Maasoumi, and J. Racine (2004): "A dependence metric for possibly nonlinear processes," *Journal of Time Series Analysis*, 25(5), 649–669.

Granger, C. W. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438.

Granger, C. W. J. (1989): *Forecasting in business and economics*. Academic Press.

Guo, S., C. Ladroue, and J. Feng (2010): "Granger causality: theory and applications," in *Frontiers in Computational and Systems Biology*, pp. 83–111. Springer.

Hansen, B. E. (1994): "Autoregressive conditional density estimation," *International Economic Review*, pp. 705–730.

——— (2009): "Lecture notes on nonparametrics," .

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

Hlaváčková-Schindler, K., M. Paluš, M. Vejmelka, and J. Bhattacharya (2007): "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, 441(1), 1–46.

Hong, Y., and H. White (2005): "Asymptotic distribution theory for nonparametric entropy measures of serial dependence," *Econometrica*, 73(3), 837–901.

Karpoff, J. M. (1987): "The relation between price changes and trading volume: A survey," *Journal of Financial and quantitative Analysis*, 22(01), 109–126.

Kraskov, A., H. Stögbauer, and P. Grassberger (2004): "Estimating mutual information," *Physical review E*, 69(6), 066138.

Kullback, S. (1968): *Information theory and statistics*. Courier Corporation.

Kullback, S., and R. A. Leibler (1951): "On information and sufficiency," *The annals of mathematical statistics*, 22(1), 79–86.

Nadaraya, E. (1965): "On non-parametric estimates of density functions and regression curves," *Theory of Probability & Its Applications*, 10(1), 186–190.

Papana, A., C. Kyrtsou, D. Kugiumtzis, and C. Diks (2016): "Detecting Causality in Non-stationary Time Series Using Partial Symbolic Transfer Entropy: Evidence in Financial Data," *Computational Economics*, 47(3), 341–365.

Podobnik, B., D. Horvatic, A. M. Petersen, and H. E. Stanley (2009): "Cross-correlations between volume change and price change," *Proceedings of the National Academy of Sciences*, 106(52), 22079–22084.

Pompe, B. (1993): "Measuring statistical dependences in a time series," *Journal of Statistical Physics*, 73(3), 587–610.

Powell, J. L., and T. M. Stoker (1996): "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics*, 75(2), 291–316.

Robinson, P. M. (1991): "Consistent nonparametric entropy-based testing," *The Review of Economic Studies*, 58(3), 437–453.

Rüschendorf, L. (1977): "Consistency of estimators for multivariate density functions and for the mode," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 243–250.

Schreiber, T. (2000): "Measuring information transfer," *Physical review letters*, 85(2), 461.

Schuster, E. F. (1969): "Estimation of a probability density function and its derivatives," *The Annals of Mathematical Statistics*, pp. 1187–1195.

Schwarz, G., et al. (1978): "Estimating the dimension of a model," *The annals of statistics*, 6(2), 461–464.

Sen, P. K., et al. (1974): "Weak convergence of multidimensional empirical processes for stationary ϕ-mixing processes," *The Annals of Probability*, 2(1), 147–154.

Shannon, C. E. (1948): "A mathematical theory of communication," *Bell System Technical Journal*, 27, 379–423; 623–656.

Shannon, C. E. (1951): "Prediction and entropy of printed English," *Bell system technical journal*, 30(1), 50–64.

Silverman, B. W. (1986): *Density estimation for statistics and data analysis*, vol. 26. CRC press.

Skaug, H. J., and D. Tjøstheim (1993): *Nonparametric tests of serial independence*chap. 15, pp. 207–209. Chapman & Hall: London.

Su, L., and H. White (2008): "A nonparametric Hellinger metric test for conditional independence," *Econometric Theory*, 24(04), 829–864.

——— (2014): "Testing conditional independence via empirical likelihood," *Journal of Econometrics*, 182(1), 27–44.

Wand, M. P., and M. C. Jones (1994): *Kernel smoothing*. Crc Press.

Wegman, E. J. (1972): "Nonparametric probability density estimation: I. A summary of available methods," *Technometrics*, 14(3), 533–546.

Wibral, M., N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente (2013): "Measuring information-transfer delays," *PloS one*, 8(2), e55809.

Wied, D., and R. Weissbach (2012): "Consistency of the kernel density estimator: a survey," *Statistical Papers*, 53(1), 1–21.

A Appendix

A.1 Proof for Non-Negativity of Transfer Entropy

According to the definition of TE in eq. (14), essentially the expectation over the logarithm of the density ratio is evaluated: $TE_{X\to Y} = E_W\left(\log\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}\right)$. Define the reciprocal of the density ratio in the logarithm as a random variable R in such a way: $R = \frac{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}{f_{Z,X|Y}(Z,X|Y)}$, we could rewrite TE as $TE_{X\to Y} = E\left(-\log(R)\right)$. Since $\log(R)$ is a concave function of R, following Jensen's inequality we have

$$E\left(\log(R)\right) \leq \log\left(E(R)\right). \tag{A.1}$$

Next, as random variable R is nonnegative since it is defined as a fraction of densities. For any realization of $R = r > 0$, $\log(r) \leq r - 1$. This is because as a concave function, $\log(r)$ is bounded from above by the tangent line at point (1,0), which is given by $r - 1$. It follows that

$$\log\left(E(R)\right) \leq E(R) - 1. \tag{A.2}$$

Combining eqs. (A.1) and (A.2), we have $E\left(\log(R)\right) \leq E(R) - 1 = 0$, where the last equality holds simply as a result of integral of the *pdf* over its full support delivering 1. A similar argument can be found in Diks (2009). Thus, we has proved that $TE_{X\to Y} \equiv -E\left(\log(R)\right) \geq 0$. It is obvious that the equity holds if and only if $R = 1$, which is equivalent to claim that $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$. This completes the proof for Thm. 3.

A.2 Proof for Non-Negativity of the First Order TE Statistic

Starting from eq. (11) and the definition of t,

$$\begin{aligned}
t &= E_W\left(\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)} - 1\right) \\
&= \int\int\left(\frac{f^2_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - f_{Z,X|Y}(z,x|y)\right)d_{x|y}d_{z|y} \\
&= \int\int f_{X|Y}(x|y)f_{Z|Y}(z|y)\left(\frac{f^2_{Z,X|Y}(z,x|y)}{f^2_{X|Y}(x|y)f^2_{Z|Y}(z|y)} - \frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)}\right)d_{x|y}d_{z|y} \\
&= \int\int f_{X|Y}(x|y)f_{Z|Y}(z|y)\left(\frac{f^2_{Z,X|Y}(z,x|y)}{f^2_{X|Y}(x|y)f^2_{Z|Y}(z|y)} - 2\frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} + 1\right)d_{x|y}d_{z|y} \\
&= \int\int f_{X|Y}(x|y)f_{Z|Y}(z|y)\left(\frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} - 1\right)^2 d_{x|y}d_{z|y} \geq 0,
\end{aligned} \tag{A.3}$$

with equality if and only if $\frac{f_{Z,X|Y}(z,x|y)}{f_{X|Y}(x|y)f_{Z|Y}(z|y)} = 1$ for any (z,x|y). In the fourth step we completed the square by using the fact that the integral over the whole support of *pdf* is 1. Finally, $t \geq 0$

follows naturally from the integrand is non-negative.

A.3 Proof for lemma2

Suppose we denote the asymptotic distribution of $\tilde{T}'_n$ in such a way that:

$$\sqrt{n}\frac{\tilde{T}'_n(h)-t}{\sigma} \xrightarrow{d} N(0,1), \tag{A.4}$$

with appropriate variance σ. To prove lemma 2, it is sufficient to show that

$$\sqrt{n}\frac{T'_n(h)-t}{\sigma} \xrightarrow{d} N(0,1),$$

because the two limiting distributions imply that $\sqrt{n}(T'_n - \tilde{T}'_n) = o_P(1)$.

To construct the equivalence of the limiting distributions, we need to prove that if $\hat{v}(.)$ consistent, the limiting distribution in eq. (A.4) remains for T'_n. Lemma 1 guarantees the estimate of a density is consistent. For a consistent density estimate, it can always be expressed as the corresponding true density plus an extra error term which vanishes in limit, for example $\hat{f}_{X,Y}(x_i,y_i) = f_{X,Y}(x_i,y_i) + \varepsilon_{X,Y}(x_i,y_i)$. Following this manner and ignoring the subscripts for simplicity, we can show that an estimated $\hat{v}(.)$ differs from $v(.)$ only in the higher order:

$$\begin{aligned}
&\frac{1}{\hat{v}(x_i,y_i,z_i)} \\
&= \frac{1}{\hat{f}(x_i,y_i)\hat{f}(y_i,z_i)} \\
&= \frac{1}{(f(x_i,y_i)+\varepsilon(x_i,y_i))(f(y_i,z_i)+\varepsilon(y_i,z_i))} \\
&= \frac{1}{f(x_i,y_i)f(y_i,z_i)+\varepsilon(x_i,y_i)f(y_i,z_i)+\varepsilon(y_i,z_i)f(x_i,y_i)+\varepsilon(x_i,y_i)\varepsilon(y_i,z_i)} \\
&= \frac{1}{f(x_i,y_i)f(y_i,z_i)}\frac{1}{\left(1+\frac{\varepsilon(x_i,y_i)f(y_i,z_i)+\varepsilon(y_i,z_i)f(x_i,y_i)+\varepsilon(x_i,y_i)\varepsilon(y_i,z_i)}{f(x_i,y_i)f(y_i,z_i)}\right)} \\
&= \frac{1}{f(x_i,y_i)f(y_i,z_i)}\left(1-\frac{\varepsilon(x_i,y_i)f(y_i,z_i)+\varepsilon(y_i,z_i)f(x_i,y_i)+\varepsilon(x_i,y_i)\varepsilon(y_i,z_i)}{f(x_i,y_i)f(y_i,z_i)}+h.o.t\right) \\
&= \frac{1}{f(x_i,y_i)f(y_i,z_i)}+h.o.t \\
&= \frac{1}{v(x_i,y_i,z_i)}+h.o.t,
\end{aligned} \tag{A.5}$$

where the last equation but one uses the approximation $\frac{1}{(1+x)} \approx 1-x+O(x^2)$, and in the last step we make a simplification by omitting higher order terms. Clearly, if eq. (A.4) holds, asymptotically it makes no difference if $v(.)$ is replaced by $\hat{v}(.)$.

A.4 Proof for the Asymptotic distribution of Modified DP Statistic

Applying lemma 2, we only need to prove that

$$\sqrt{n}\frac{\tilde{T}'_n(h)-t}{S_n} \xrightarrow{d} N(0,1).$$

According to the definition, $\tilde{T}'_n(h)$ is a re-scaled DP statistic with the scaling factor $(1/v(.))$. In a similar manner of Theorem 1 in Diks and Panchenko (2006), we could obtain the asymptotic behavior of $\tilde{T}'_n(h)$ by making use of the optimal mean squared error (MSE) bandwidth developed by Powell and Stoker (1996) for point estimator. The test statistic $\tilde{T}'_n(h)$ can be expressed by an order three U-statistic $\tilde{K}(W_i, W_j, W_k)$ by symmetrization with respect to the indices i, j, k. Further define two kernel functions as $\tilde{K}_1(w_i) = E[\tilde{K}(w_i, W_j, W_k)]$ and $\tilde{K}_2(w_i, w_j) = E[\tilde{K}(w_i, w_j, W_k)]$, we assume the three mild conditions adapting from Powell and Stoker (1996) for controlling the rate of convergence of the point-wise bias as well as the serial expansions of the kernel functions:

$$\begin{aligned}
\tilde{K}_1(w_i,h) - \lim_{h\to 0}\tilde{K}_1(w_i,h) &= s(w_i)h^{\alpha} + s^*(w_i,h),\ \alpha>0 \\
E[(\tilde{K}_2(W_i,W_j))^2] &= q_2 h^{-\gamma} + q_2^*(h),\ \gamma>0 \\
E[(\tilde{K}(W_i,W_j,W_k))^2] &= q_3 h^{-\delta} + q_3^*(h),\ \delta>0
\end{aligned} \tag{A.6}$$

where all remainder terms are of higher orders, i.e. $E||s^*(W_i,h)||^2 = o_P(h^{2\alpha})$, $q_2^*(h) = o_P(h^{-\gamma})$ and $q_3^*(h) = o_P(h^{-\delta})$ and the convergence rate is controlled by the parameters α, γ and δ. Conditions eq. (A.6) are satisfied if α is set as the order of kernel function $\mathbb{K}(.)$, in our case for Gaussian kernel which is 2, and γ, δ depend on the dimensions of the variables under consideration in such a way: $\gamma = d_X + d_Y + d_Z$ and $\delta = d_X + 2d_Y + d_Z$. Define $C_0 = 2\text{cov}[\lim_{h\to 0}\tilde{K}_1(W_i,h), s(W_i)]$, we can show the mean squared error of DP statistic as a function of sample size dependent bandwidth:

$$\text{MSE}[T_n(h)] = (E[s(W_i)])^2 h^{2\alpha} + \frac{9}{n}C_0 h^{\alpha} + \frac{9}{n}\text{Var}\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right] + \frac{18}{n^2}q_2 h^{-\gamma} + \frac{6}{n^3}q_3 h^{-\delta} + \text{h.o.t} \tag{A.7}$$

The scaling factor $(1/v(.))$ in $\tilde{T}'_n(h)$ would enter the MSE in eq. (A.7) by mainly changing the variance term. For other bandwidth-dependent terms, $(1/v(.))$ just re-scale their coefficients without affecting the convergence rates. Thus, we may still let all those h-dependent terms to be $o_P(n^{-1})$ to make the 9Var $\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right]$ dominant as in Diks and Panchenko (2006). Therefore, adopting a sample size n-dependent bandwidth $h = Cn^{\beta}$, with C, $\beta > 0$, one finds

$$\sqrt{n}\frac{\tilde{T}'_n(h)-t}{S_n} \xrightarrow{d} N(0,1) \text{ if } \frac{1}{2\alpha} < \beta < \frac{1}{d_X+d_Y+d_Z} \tag{A.8}$$

where S_n^2 is a consistent estimator of the asymptotic variance $9\text{Var}\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right]$. In our bivariate case, $\alpha = 2$ and $d_X + d_Y + d_Z = 3$, and we would have $\beta \in (1/4, 1/3)$.

A.5 Proof for Non-Degeneracy of the Limiting Distribution of the Modified DP Test Statistic

To show the asymptotic normality in Thm. 3 is a non-degenerate distribution, it is sufficient to prove that with the plug-in weighting function $v(\cdot)$, the modified DP test U-statistic kernel is not degenerated.

The symmetrized U-statistic representation of the modified DP test statistic defined in eq. (14) is given by

$$\begin{aligned} K(w_1,w_2,w_3) = &\left[\left(\kappa_{XYZ}(w_1-w_2)\kappa_Y(w_1-w_3) - \kappa_{XY}(w_1-w_2)\kappa_{YZ}(w_1-w_3)\right)/v(w_1)\right]/6 \\ &+ \text{permutations of } w_1, w_2, w_3, \end{aligned} \tag{A.9}$$

where κ is a (bandwidth h dependent) density estimation kernel function, and $w_i = (x_i, y_i, z_i)'$, $i \in \{1,2,3\}$.

Let r_1 be the Hájek projection

$$r_1(w_1;h) = E(K(w_1, W_2, W_3))$$

of the U-statistic kernel.

Define

$$r_1(w_1) = \lim_{h\to 0} r_1(w_1;h).$$

The U-statistic is degenerate (in that the variance is of higher order than in the derivation of the modified DP test statistic) if $r_1(w_1)$ is constant as a function of $w_1 = (x_1, y_1, z_1)'$. Combining

the above equations and eq. (A.9), we obtain

$$\begin{aligned}
r_1(w_1;h) =& E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(w_1-W_3)\kappa_Y(w_1-W_2)-\kappa_{XY}(w_1-W_3)\kappa_{YZ}(w_1-W_2)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-w_1)\kappa_Y(W_3-W_2)-\kappa_{XY}(W_3-w_1)\kappa_{YZ}(W_3-W_2)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/6 \\
=& E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]/3 \\
&+E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]/3 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/3 \\
\equiv& E_1(w_1;h)/3+E_2(w_1;h)/3+E_3(w_1;h)/3,
\end{aligned} \tag{A.10}$$

where in the last step we used the fact that the terms with W_2 and W_3 swapped are identical.

We next consider

$$\begin{aligned}
r_1(w_1) &= \lim_{h\to 0} r_1(w_1;h) \\
&= \lim_{h\to 0} E_1(w_1;h)/3+\lim_{h\to 0} E_2(w_1;h)/3+\lim_{h\to 0} E_3(w_1;h)/3 \\
&\equiv E_1(w_1)/3+E_2(w_1)/3+E_3(w_1)/3.
\end{aligned}$$

For $E_1(w_1)$ we find

$$\begin{aligned}
E_1(w_1) =& \lim_{h\to 0} E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right] \\
=& \lim_{h\to 0}\int\int f_W(w_2)f_W(w_3)\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)\right.\right. \\
&\left.\left.-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]\mathrm{d}w_2\mathrm{d}w_3 \\
=& \frac{1}{v(w_1)}\int\int f_{XYZ}(w_2)f_{XYZ}(w_3)\left(\delta_{XYZ}(w_1-w_2)\delta_Y(w_1-w_3)\right. \\
&\left.-\delta_{XY}(w_1-w_2)\delta_{YZ}(w_1-w_3)\right)\mathrm{d}w_2\mathrm{d}w_3 \\
=& \frac{1}{v(w_1)}\left(f_{XYZ}(w_1)f_Y(w_1)-f_{XY}(w_1)f_{YZ}(w_1)\right),
\end{aligned} \tag{A.11}$$

where in the third step, $\delta(\cdot)$ is the Dirac delta function, also referred to as the unit impulse symbol. Using convolution, we have the last equality. Under H_0 for all w_1 in the support of W, eq. (A.11) is zero by construction.

However, the other terms, $E_2(w_1)$ and $E_3(w_1)$, need not be constant even under H_0. For

instance,

$$
\begin{aligned}
E_2(w_1,h) =& E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]\\
E_2(w_1) =& \lim_{h\to 0} E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]\\
=& \lim_{h\to 0} E\left[\left(\kappa_{XYZ}(W_2-w_1)f_Y(W_2)-\kappa_{XY}(W_2-w_1)f_{YZ}(W_2)\right)/v(w_1)\right]\\
=& \lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XYZ}(w_2-w_1)f_Y(w_2)/v(w_1)\,\mathrm{d}w_2\\
& -\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& \int f_{XYZ}(w_2)\delta_{XYZ}(w_2-w_1)f_Y(w_2)/v(w_1)\,\mathrm{d}w_2\\
& -\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)-\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)\\
& -\lim_{h\to 0}\int f_{XYZ}(x_2,y_2,z_2)\kappa_{XY}(x_2-x_1,y_2-y_1)f_{YZ}(y_2,z_2)/v(w_1)\,\mathrm{d}x_2\mathrm{d}y_2\mathrm{d}z_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)\\
& -\frac{1}{v(w_1)}\int f_{XYZ}(x_2,y_2,z_2)\delta_{XY}(x_2-x_1,y_2-y_1)f_{YZ}(y_2,z_2)\,\mathrm{d}x_2\mathrm{d}y_2\mathrm{d}z_2\\
=& \frac{1}{v(w_1)}(f_{XYZ}(w_1)f_Y(w_1)-\int f_{XYZ}(x_1,y_1,z_2)f_{YZ}(y_1,z_2)\,\mathrm{d}z_2).
\end{aligned}
\tag{A.12}
$$

Since the last term in the bracket does not depend on z_1, while the first typically depends on x_1, y_1 and z_1 under H_0, $E_2(w_1)$ typically isn't constant. A similar argument also can be used to show that $E_3(w_1)$ is not constant (and neither is $E_2(w_1)+E_3(w_1)$, because $E_3(w_1)$ is a function of (y_1, z_1) only, while $E_2(w_1)$ also depends on x_1 typically).

For completeness, $E_3(w_1, h)$ and $E_3(w_1)$ are given below:

$$\begin{aligned}
E_3(w_1, h) =& E\left[(\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1))/v(w_1)\right] \\
E_3(w_1) =& \lim_{h\to 0} E\left[(\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1))/v(w_1)\right] \\
=& \lim_{h\to 0} \int f_{XYZ}(w_3)\kappa_Y(w_3 - w_1) f_{XYZ}(w_3)/v(w_1)\,\mathrm{d}w_3 \\
& - \lim_{h\to 0} \int f_{XY}(w_3)\kappa_{YZ}(w_3 - w_1) f_{XYZ}(w_3)/v(w_1)\,\mathrm{d}w_3 \\
=& \lim_{h\to 0} \int f_{XYZ}(x_3, y_3, z_3)\kappa_Y(y_3 - y_1) f_{XYZ}(x_3, y_3, z_3)/v(w_1)\,\mathrm{d}x_3\mathrm{d}y_3\mathrm{d}z_3 \\
& - \lim_{h\to 0} \int f_{XY}(x_3, y_3)\kappa_{YZ}(y_3 - y_1, z_3 - z_1) f_{XYZ}(x_3, y_3, z_3)/v(w_1)\,\mathrm{d}x_3\mathrm{d}y_3\mathrm{d}z_3 \\
=& \frac{1}{v(w_1)}\left(\int f_{XYZ}(x_3, y_1, z_3) f_{XYZ}(x_3, y_1, z_3)\,\mathrm{d}x_3\mathrm{d}z_3 - \int f_{XY}(x_3, y_1) f_{XYZ}(x_3, y_1, z_1)\,\mathrm{d}x_3\right)
\end{aligned} \tag{A.13}$$

Because $E_2(w_1)$ and $E_3(w_1)$ are not constant, $r_1(w_1)$ cannot be constant, hence the U-statistic defined in eq. (14) is non-degenerate.

To illustrate that $E_2(w_1)$ and $E_3(w_1)$ are typically not constant, consider the example where $W \sim N(0, I_3)$. In this case H_0 holds, so we have $E_1 = 0$, while

$$v(w_1) = f_{X,Y}(x_1, y_1) f_{Y,Z}(y_1, z_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^4 \mathrm{e}^{-(x_1^2 + 2y_1^2 + z_1^2)/2},$$

plug into eqs. (A.12) and (A.13) we have

$$\begin{aligned}
E_2(w_1) =& \left[\left(\frac{1}{\sqrt{2\pi}}\right)^4 \mathrm{e}^{-(x_1^2+2y_1^2+z_1^2)/2} - \left(\frac{1}{\sqrt{2\pi}}\right)^5 \int \mathrm{e}^{-(x_1^2+2y_1^2+2z_2^2)/2}\,\mathrm{d}z_2\right]/v(w_1) \\
=& \left[\left(\frac{1}{\sqrt{2\pi}}\right)^4 \mathrm{e}^{-(x_1^2+2y_1^2+z_1^2)/2} - \left(\frac{1}{\sqrt{2\pi}}\right)^5 \mathrm{e}^{-(x_1^2+2y_1^2)/2} \int \mathrm{e}^{-2z_2^2/2}\,\mathrm{d}z_2\right]/v(w_1) \\
=& \left[\left(\frac{1}{\sqrt{2\pi}}\right)^4 \mathrm{e}^{-(x_1^2+2y_1^2+z_1^2)/2} - \left(\frac{1}{\sqrt{2\pi}}\right)^4 \frac{1}{\sqrt{2}} \mathrm{e}^{-(x_1^2+2y_1^2)/2}\right]/v(w_1) \\
=& \left[\left(\frac{1}{\sqrt{2\pi}}\right)^4 \left(\mathrm{e}^{-z_1^2/2} - \frac{1}{\sqrt{2}}\right) \mathrm{e}^{-(x_1^2+2y_1^2)/2}\right]/v(w_1) \\
=& \left(\mathrm{e}^{-z_1^2/2} - \frac{1}{\sqrt{2}}\right) \mathrm{e}^{z_1^2/2} \\
=& 1 - \frac{1}{\sqrt{2}} \mathrm{e}^{z_1^2/2}
\end{aligned}$$

$$
\begin{aligned}
E_3(w_1) &= \left[\int f_{XYZ}^2(x,y_1,z)\,\mathrm{d}x\mathrm{d}z - \int f_{XY}(x,y_1)f_{XYZ}(x,y_1,z_1)\,\mathrm{d}x\right]/v(w_1) \\
&= \left[\left(\frac{1}{\sqrt{2\pi}}\right)^6 \int \left(\mathrm{e}^{-(x^2+y_1^2+z^2)/2}\right)^2 \mathrm{d}x\mathrm{d}z - \left(\frac{1}{\sqrt{2\pi}}\right)^5 \int \mathrm{e}^{-(2x^2+2y_1^2+z_1^2)/2}\,\mathrm{d}x\right]/v(w_1) \\
&= \left[\left(\frac{1}{\sqrt{2\pi}}\right)^6 \mathrm{e}^{-y_1^2} \int \mathrm{e}^{-(x^2+z^2)}\,\mathrm{d}x\mathrm{d}z - \left(\frac{1}{\sqrt{2\pi}}\right)^5 \mathrm{e}^{-(2y_1^2+z_1^2)/2} \int \mathrm{e}^{-x^2}\,\mathrm{d}x\right]/v(w_1) \\
&= \left[\left(\frac{1}{\sqrt{2\pi}}\right)^6 \mathrm{e}^{-y_1^2}\pi - \left(\frac{1}{\sqrt{2\pi}}\right)^5 \mathrm{e}^{-(2y_1^2+z_1^2)/2}\sqrt{\pi}\right]/v(w_1) \\
&= \left[\left(\frac{1}{\sqrt{2}}\right)^5 \frac{1}{\pi^2}\left(\frac{1}{\sqrt{2}}\mathrm{e}^{-y_1^2} - \mathrm{e}^{-y_1^2}\mathrm{e}^{-z_1^2/2}\right)\right]/v(w_1) \\
&= \left[\left(\frac{1}{\sqrt{2}}\right)^5 \frac{1}{\pi^2}\mathrm{e}^{-y_1^2}\left(\frac{1}{\sqrt{2}} - \mathrm{e}^{-z_1^2/2}\right)\right]/v(w_1) \\
&= \frac{1}{\sqrt{2}}\left(\frac{1}{\sqrt{2}} - \mathrm{e}^{-z_1^2/2}\right)\mathrm{e}^{(x_1^2+z_1^2)/2},
\end{aligned}
$$

clearly both $E_2(w_1)$ and $E_3(w_1)$ are not constant.

A.6 Optimal Bandwidth for Modified DP Test

The optimal bandwidth should balance the squared bias and variance of the test statistic, given in eq. (A.7). Particularly, the first and fourth terms are leading, and all reminders are of higher order. The optimal bandwidth should have a similar form as the one for DP test in eq. (15).

In fact, the difference between two bandwidth is up to a scalar as a result of replacing square kernel by Gaussian one. Assuming the product kernel in eq. (6), the bias and variance of the density estimator are described following Wand and Jones (1994) and Hansen (2009),

$$
\begin{aligned}
\mathrm{Bias}(\hat{f}(x)) &= \frac{\mu_\nu(\kappa)}{\nu!}\sum_{j=1}^{k}\frac{\partial^\nu}{\partial x_j^\nu}f(x)h_j^\nu + o_P(h_1^\nu + ... + h_k^\nu), \\
\mathrm{Var}(\hat{f}(x)) &= \frac{f(x)R(\kappa)^k}{(n-1)h_1h_2...h_k} + o_P((n-1)h_1h_2...h_k),
\end{aligned}
\tag{A.14}
$$

where $\mu_\nu(\kappa) = \int_{-\infty}^{\infty} t^\nu\kappa(t)dt$ is the νth moment of a kernel function, with ν the corresponding order of the kernel. For Gaussian kernel $\kappa(.)$, $\nu = 2$. The function $R(\kappa) = \int_{-\infty}^{\infty}\kappa(t)^2dt$ is the so called roughness function of the kernel. For a k-dimensional vector, the multivariate density estimation is carried out with a bandwidth vector $\mathbf{H} = (h_1, ..., h_k)'$. It is not difficult to see that $\mathrm{E}[s(W)]$ and q_2 defined in eq. (A.6) depend on the used kernel function trough functions $\mu_\nu(\kappa)$ and $R(\kappa)$.

Using the superscript 'G' and 'SQ' to denote Gaussian and square kernels, Hansen (2009)

shows

$$\begin{aligned} \mu_\nu^{SQ}(\kappa) &= 1/3, \quad R^{SQ}(\kappa) = 1/2, \\ \mu_\nu^{G}(\kappa) &= 1, \quad R^{G}(\kappa) = 1/2\sqrt{\pi}. \end{aligned} \tag{A.15}$$

In this research, when we substitute the square kernel by Gaussian kernel, the squared bias-related $\mathrm{E}[s(W)]$ and the variance-related q_2 will change correspondingly. Directly applying eq. (A.15), we have $q^G(\kappa)_2 = 3q^{SQ}(\kappa)_2$, $R^G(\kappa) = R^{SQ}(\kappa)/\sqrt{\pi}$. Plug this into eq. (15) and do some calculations, one may find

$$h^* \approx 0.6 h_{DP}. \tag{A.16}$$

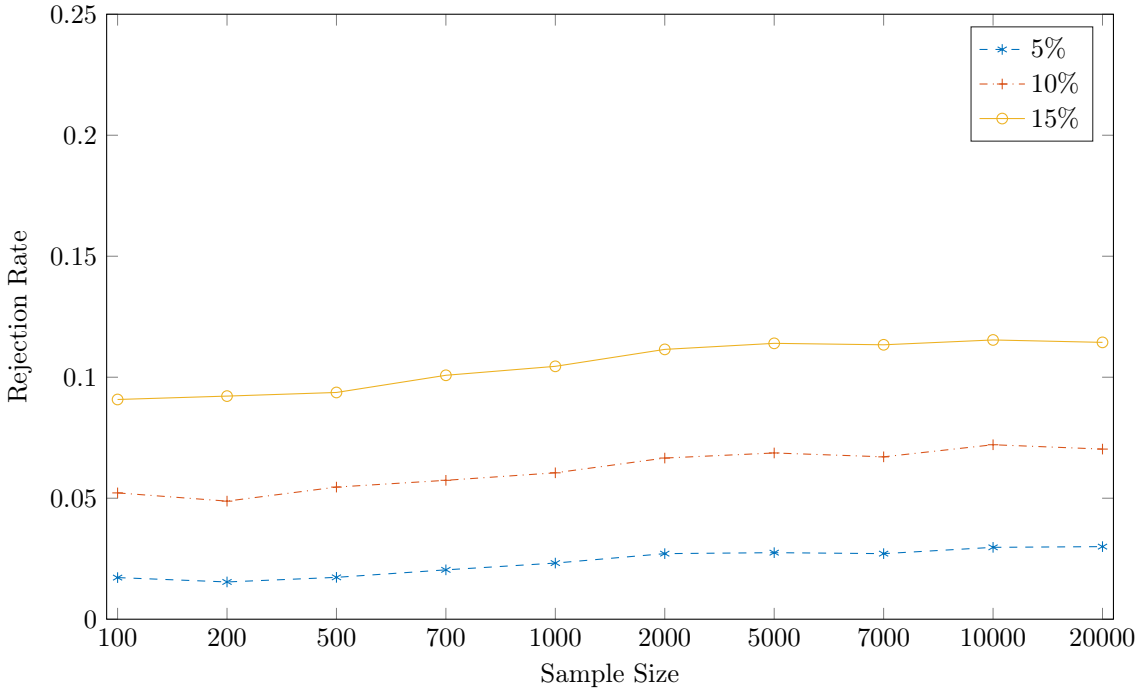


Figure 1: Power for the one-sided DP test for the artificial process $\{X_t, Y_t, Z_t\}$ given $q = 0$ at nominal size, from the bottom to the top, 5%, 10% and 15%, respectively, based on 10,000 independent replications. The sample size runs from 100 to 20,000.

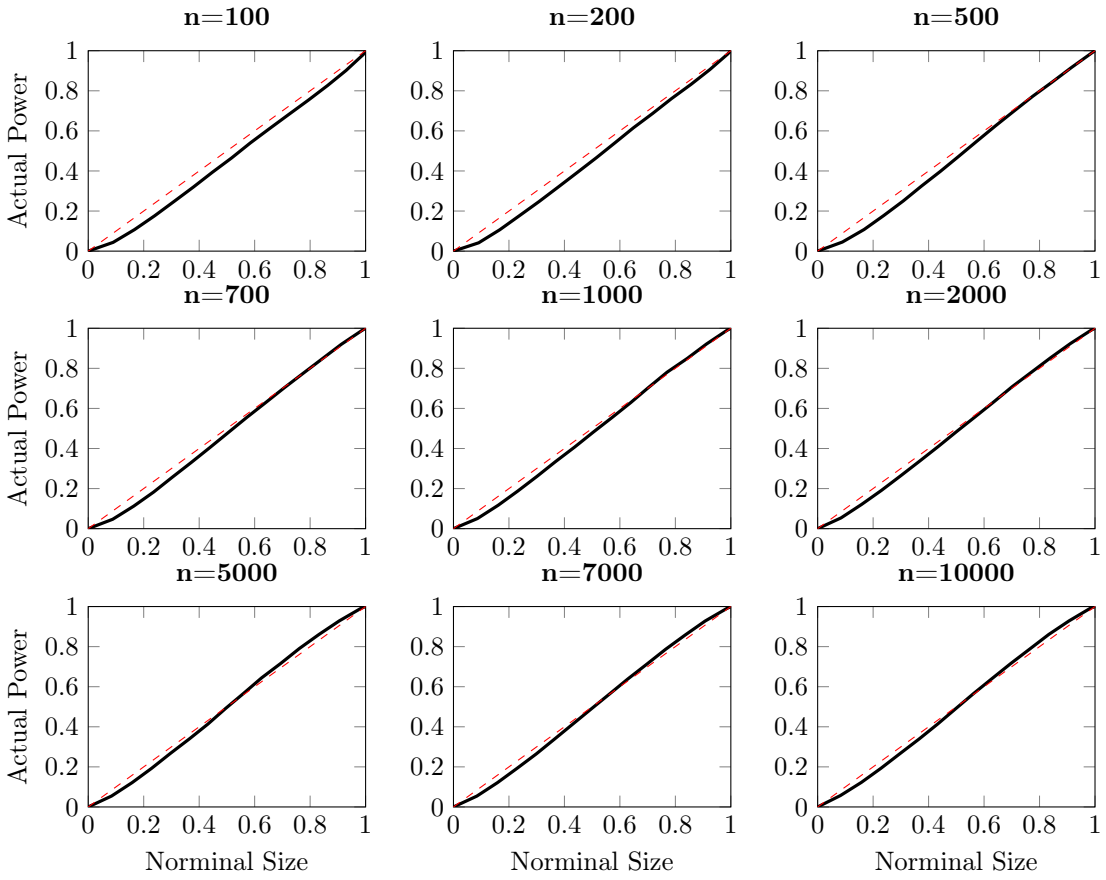


Figure 2: Size-power plots for the one-sided DP test for the artificial process $\{X_t, Y_t, Z_t\}$ given $q = 0$, based on 10,000 independent replications. Each subplot draws the actual power against the nominal size for different sample sizes, ranging from 100 to 10,000. The solid curve represents the actual power and the red dash line indicate the 45-degree line, also been treated as the theoretical size of a test.

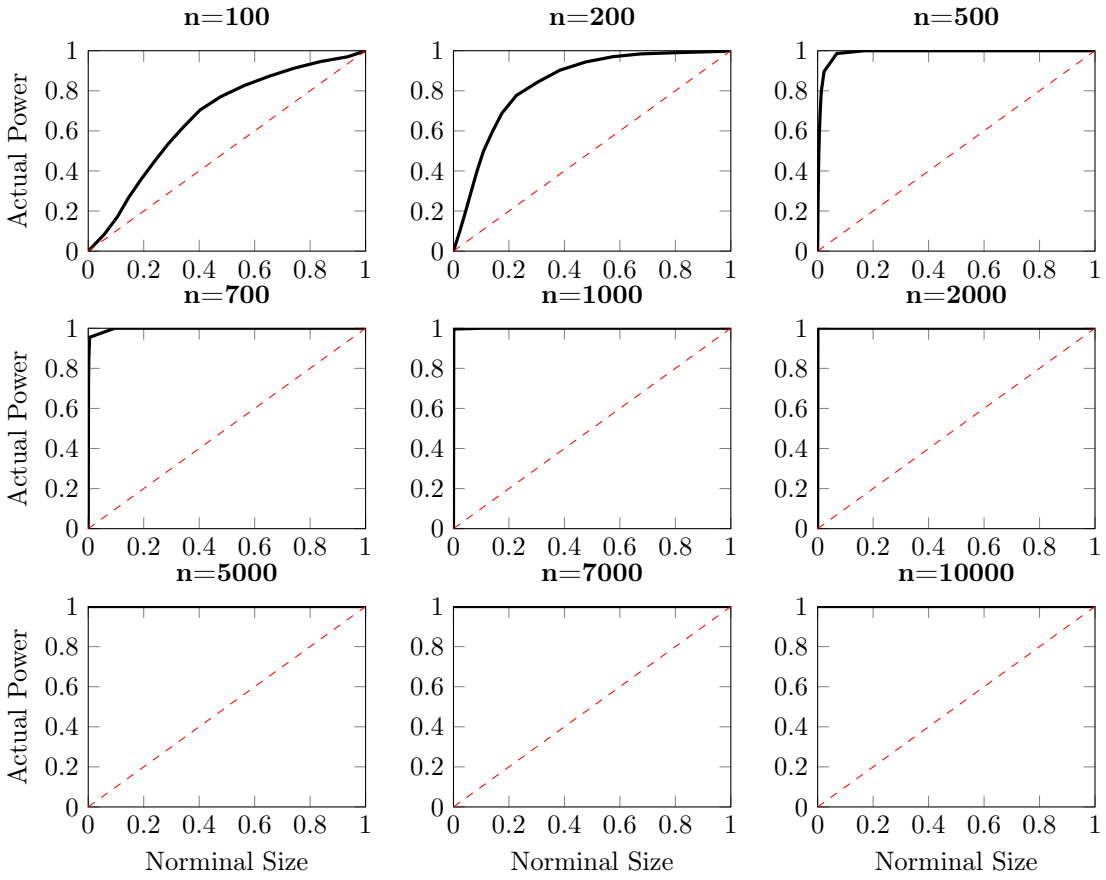


Figure 3: Size-power plots for the one-sided modified DP test for the artificial process $\{X_t, Y_t, Z_t\}$ given $q = 0$, based on 10,000 independent replications. Each subplot draws the actual power against the nominal size for different sample sizes, ranging from 100 to 10,000. The solid curve represents the actual power and the red dash line indicate the 45-degree line, also been treated as the theoretical size of a test.

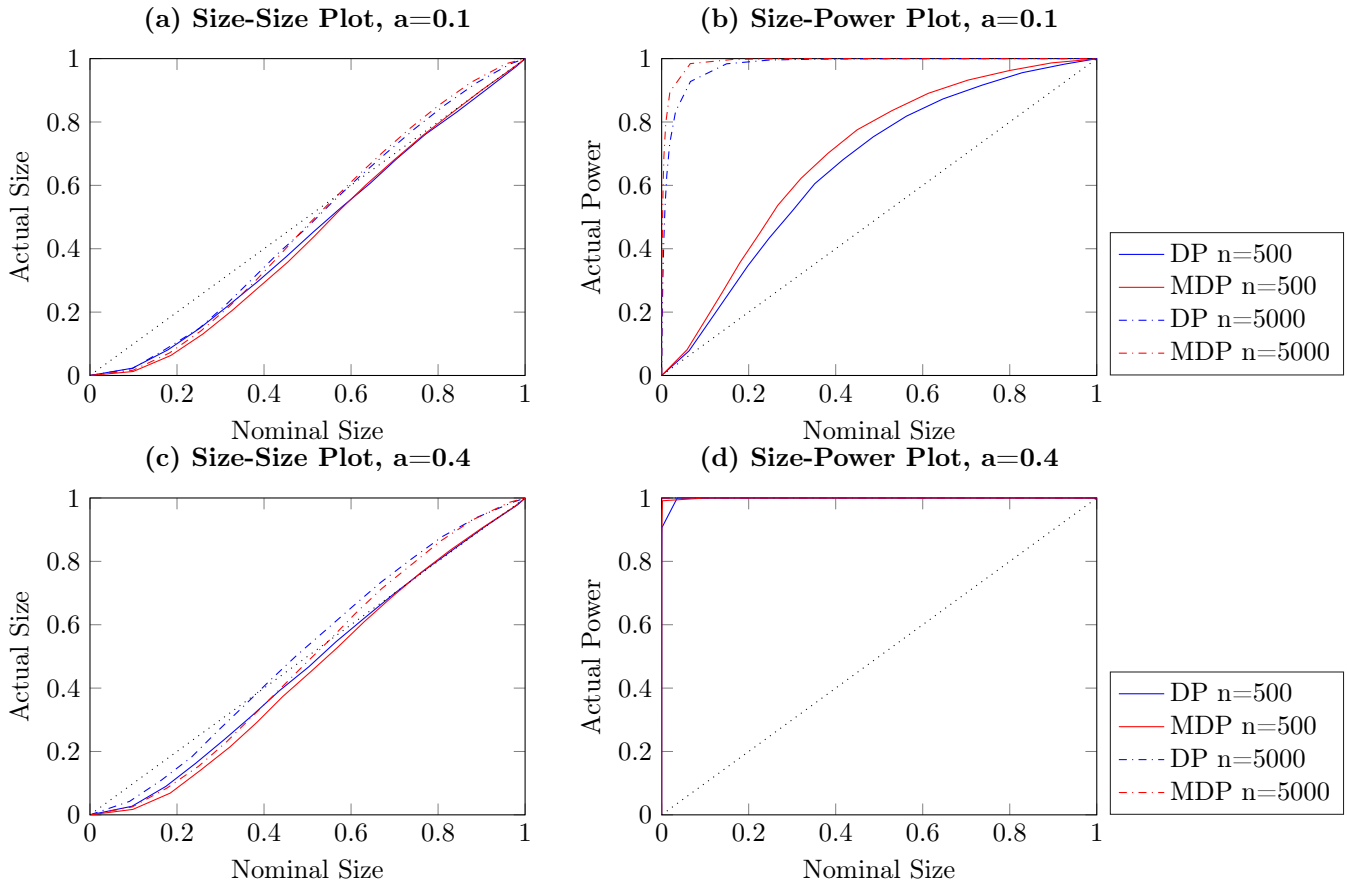


Figure 4: Size-size and size-power plots of Granger non-causality tests, based on 5,000 replications. The DGP is bivariate linear VAR as in eq. (18), with Y affecting X. The left (right) column shows observed rejection rates under the null(alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5000$, respectively.

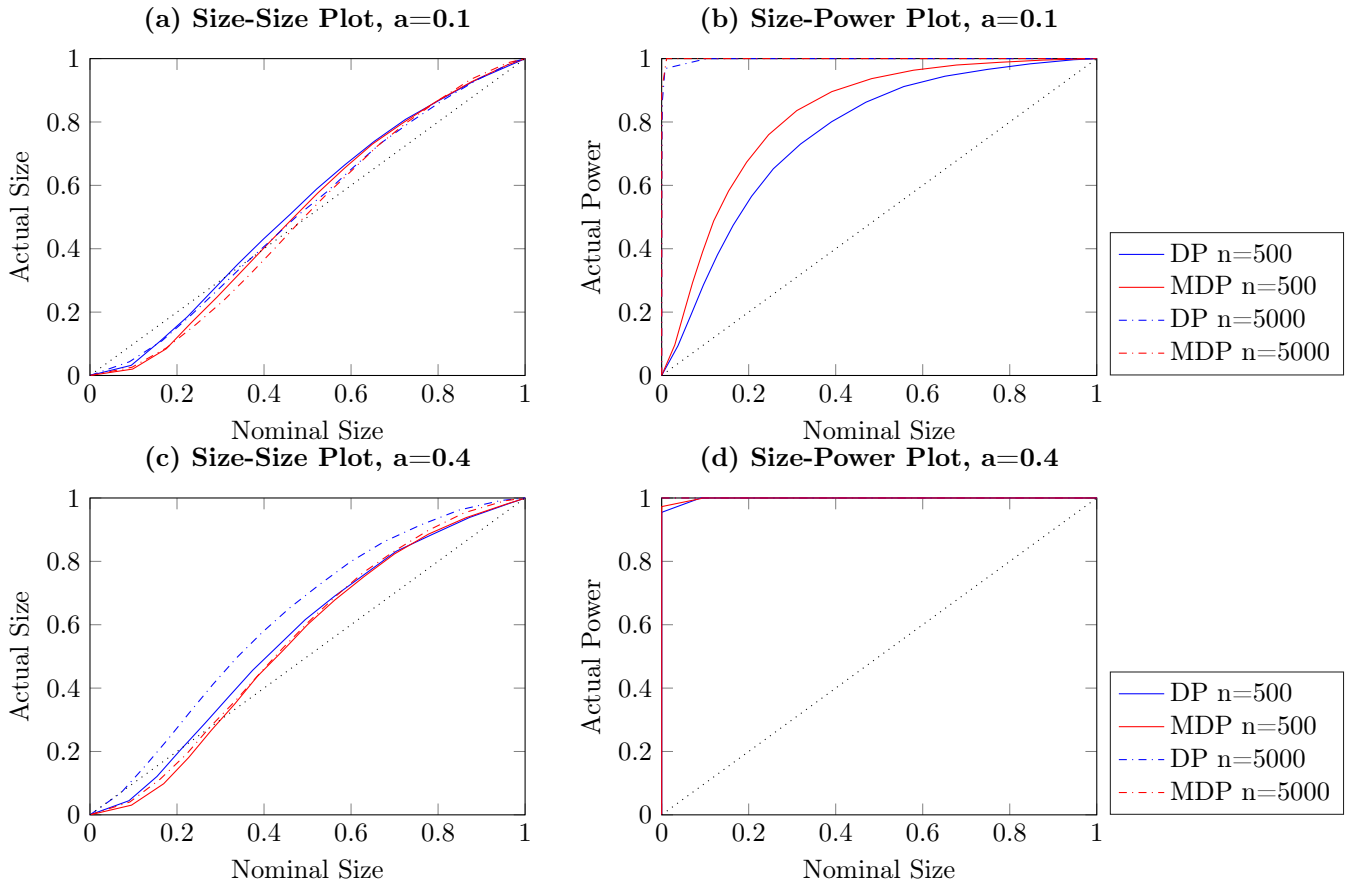


Figure 5: Size-size and size-power plots of Granger non-causality tests, based on 5,000 replications. The DGP is bivariate Non-linear VAR as in eq. (19), with Y affecting X. The left (right) column shows observed rejection rates under the null(alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5000$, respectively.

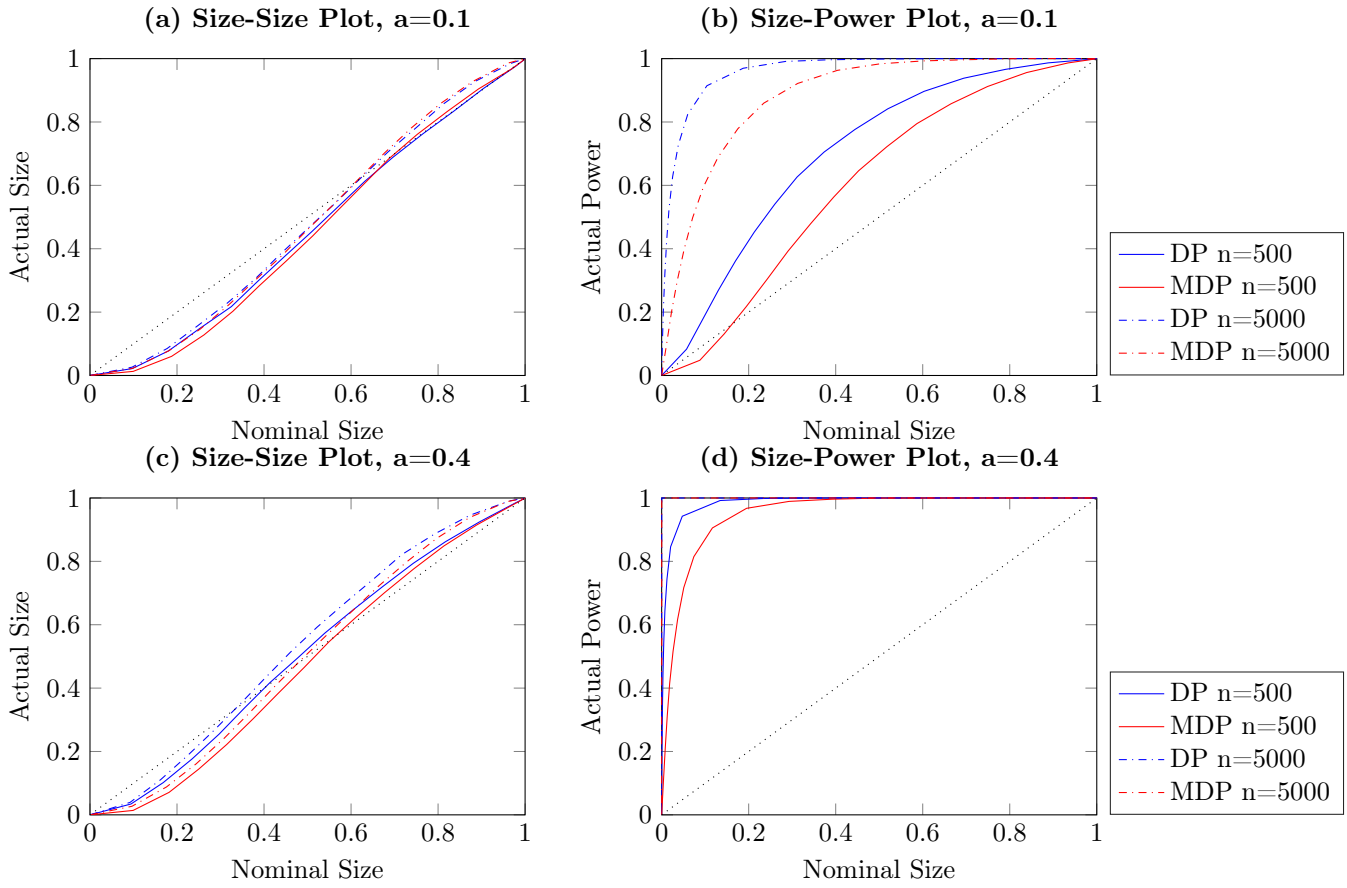


Figure 6: Size-size and size-power plots of Granger non-causality tests, based on 5,000 replications. The DGP is bivariate ARCH as in eq. (20), with Y affecting X. The left (right) column shows observed rejection rates under the null(alternative), blue color stands for DP test while red line indicates modified DP test. Real line and dash line present results with sample size $n = 500$ and $n = 5000$, respectively.

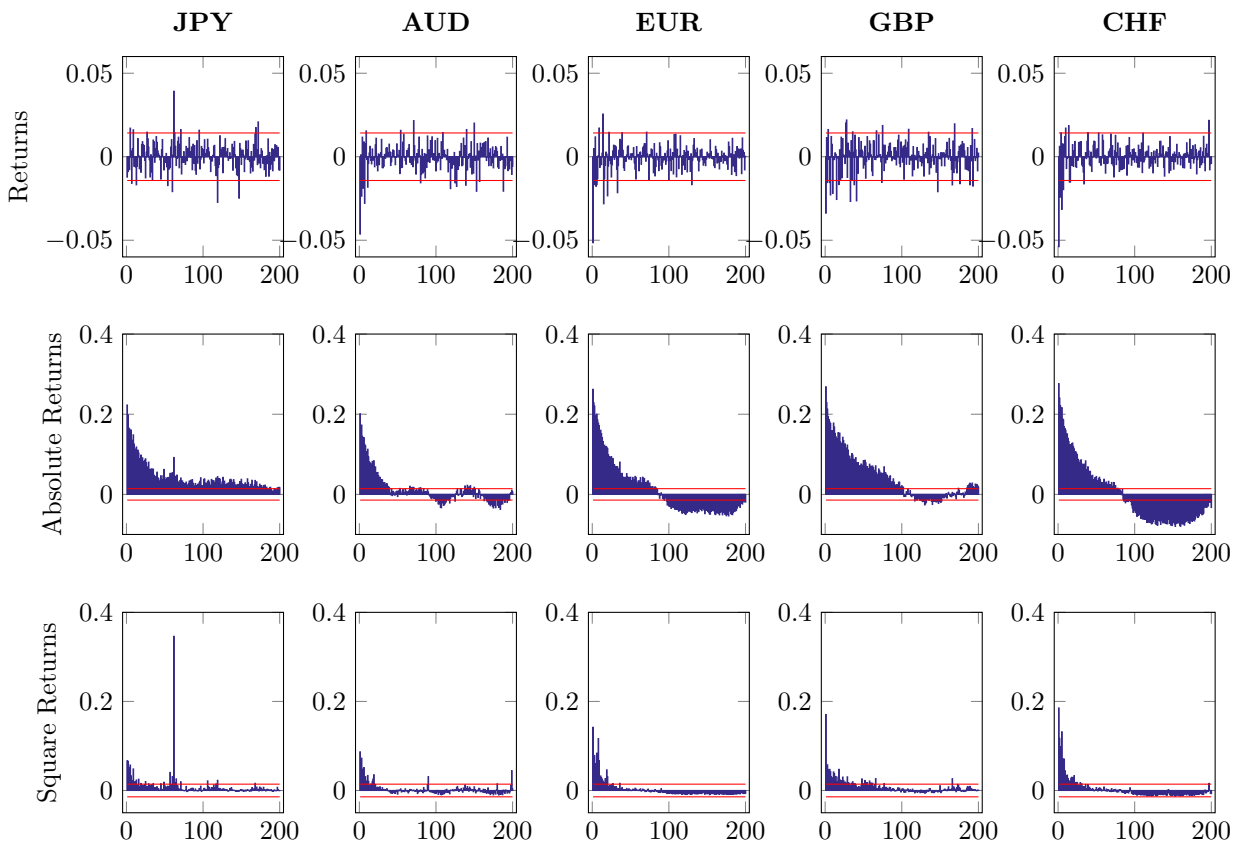


Figure 7: Autocorrelations of Returns, Absolute Returns and Square Returns, up to 200 lags.

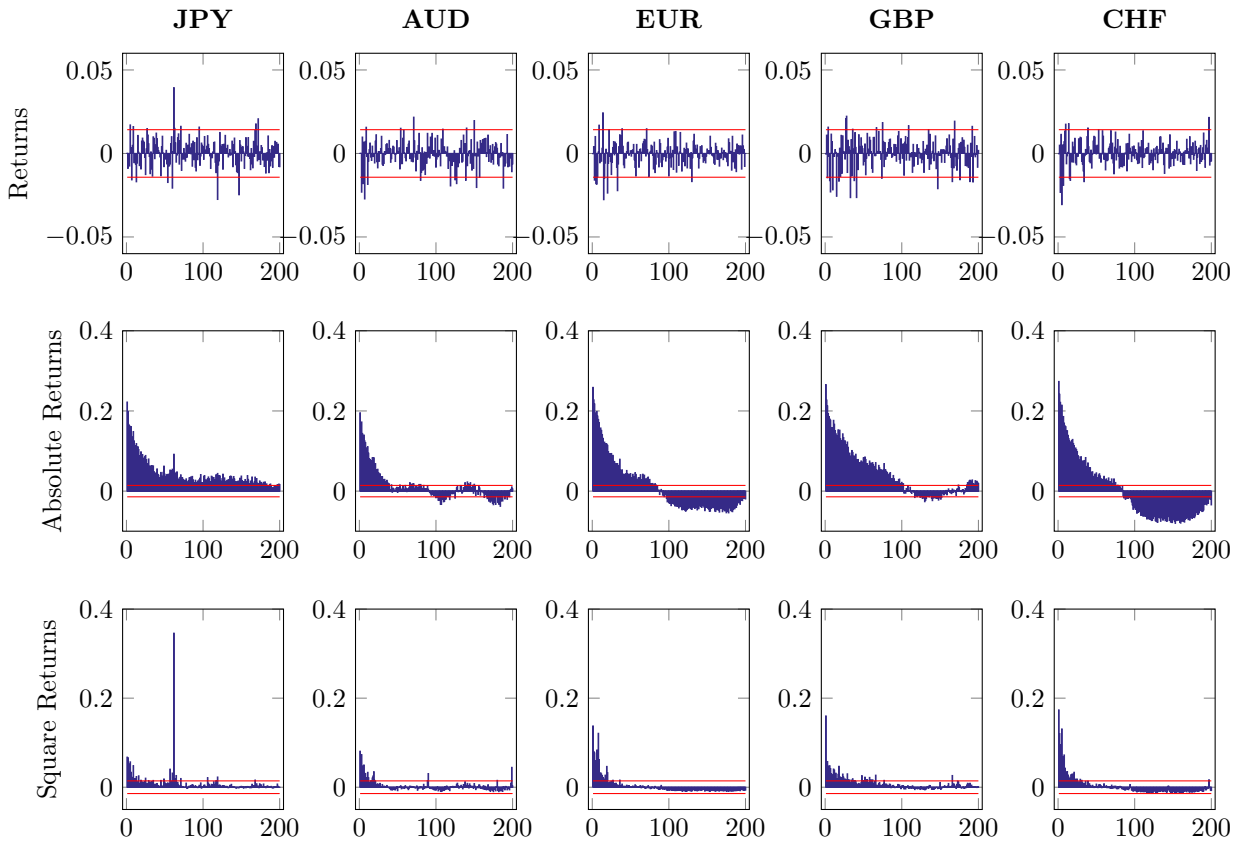


Figure 8: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) Component removed, up to 200 lags.

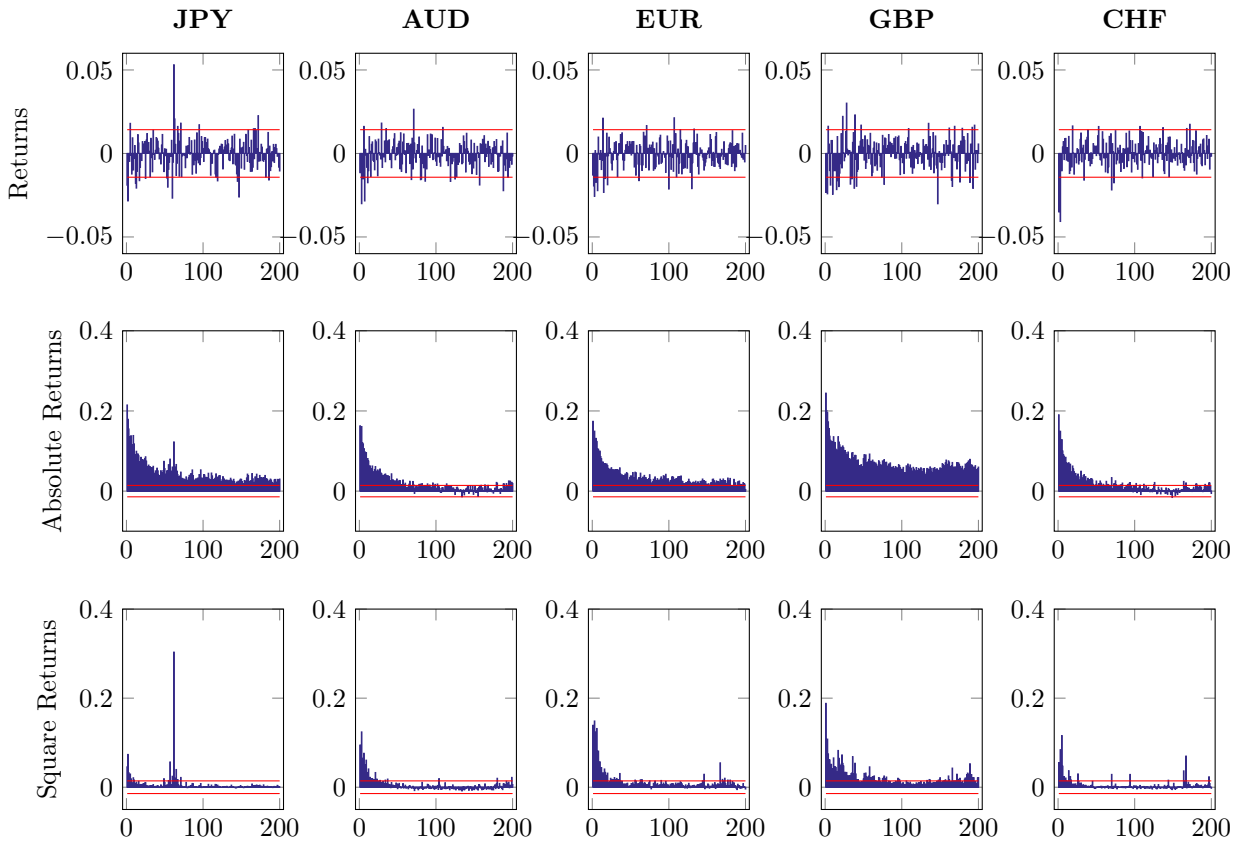
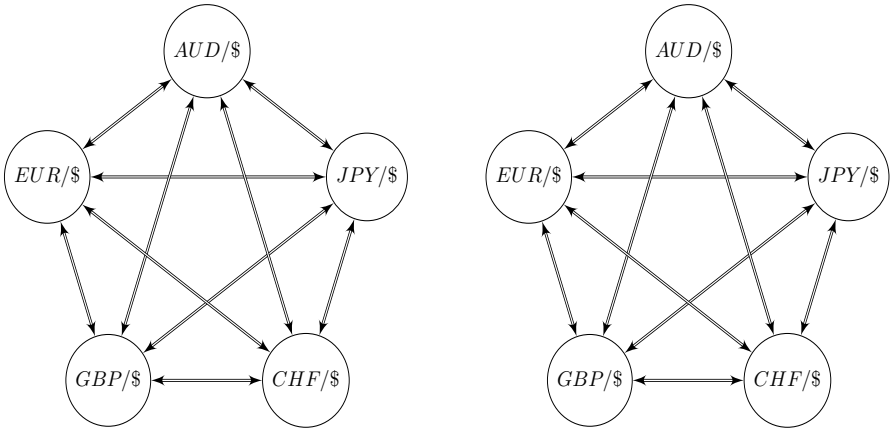


Figure 9: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) and GARCH Component removed, up to 200 lags.



(a) DP test on MA residuals

(b) Modified DP test on MA residuals

(c) DP test on EGARCH residuals

(d) Modified DP test on EGARCH residuals

Figure 10: Graphic representation of pairwise causalities on MA and seasonal component filtered residuals, as well as EGARCH filtered residuals. All "$\rightarrow$" in the graph indicate a significant directional causality at 5% level.

Table 1: Observed Size and Power of the $T'_n(h)$ test for bivariate ARCH process eq. (17)

	n	200	500	1000	2000	5000
	h	1.0563	0.8130	0.6670	0.5471	0.4211
$a = 0.1$	Size	0.0020	0.0016	0.0036	0.0016	0.0048
	Power	0.0032	0.0128	0.0288	0.0792	0.4000
$a = 0.2$	Size	0.0020	0.0008	0.0032	0.0016	0.0048
	Power	0.0208	0.0932	0.2824	0.7292	0.9992
$a = 0.3$	Size	0.0020	0.0012	0.0028	0.0032	0.0044
	Power	0.0816	0.3668	0.7916	0.9972	1.0000
$a = 0.4$	Size	0.0020	0.0016	0.0032	0.0028	0.0084
	Power	0.1928	0.6968	0.9848	1.0000	1.0000

Note: Table 1 presents the empirical size and power of modified DP test for process eq. (17) for different sample sizes and parameter a. The values represent observed rejection rates over 5000 realizations for nominal size 0.05.

Table 2: Test Statistics for the S&P500 returns and volume data

	Volume → Return				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	0.8503	2.8769**	2.9952**	3.8526**	2.7929**
VAR residuals	-	3.6880**	3.5683**	4.2696**	3.5769**
EGARCH residuals	-	1.4403	1.2347	1.2672	2.4143**
	Return → Volume				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	18.8302**	4.9309**	4.5138**	4.9253**	3.8359**
VAR residuals	-	5.2732**	5.1692**	5.2239**	3.7835**
EGARCH residuals	-	3.0067**	3.1214**	3.1176**	3.5101**

Note: Table 2 presents the test statistics for the Granger causality between S&P500 returns and volume data. Results are shown for the linear Granger test, the DP test and the modified DP test. Bandwidth values are . The asterisks indicate significance at the 5% (*) and 1% (**) levels.

Table 3: Test Statistics for the NASDAQ returns and volume data

	Volume → Return				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	0.0979	3.5894**	3.3751**	4.1311**	3.3532**
VAR residuals	-	4.3932**	4.2931**	5.3026**	3.7300**
EGARCH residuals	-	0.8282	0.5604	1.0430	1.2531
	Return → Volume				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	11.1736**	5.1100**	5.0201**	4.9980**	4.5935**
VAR residuals	-	5.6293**	6.4855**	5.5750**	5.0233**
EGARCH residuals	-	3.5959**	4.0693**	4.0745**	4.4522**

Note: Table 3 presents the test statistics for the Granger causality between NASDAQ returns and volume data. Results are shown for the linear Granger test, the DP test and the modified DP test. Bandwidth values are . The asterisks indicate significance at the 5% (*) and 1% (**) levels.

Table 4: Test Statistics for the DJIA returns and volume data

	Volume → Return				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	0.9761	1.2557	1.8384*	1.9450*	2.2697*
VAR residuals	-	1.8711*	2.0951*	2.0998*	2.7207**
EGARCH residuals	-	1.4543	1.4317	1.0566	1.4801
	Return → Volume				
	Linear	DP		MDP	
		$h = 1.5$	$h = 0.6$	$h = 1.5$	$h = 0.6$
Raw data	17.0779**	2.3076*	1.7236*	2.4972**	3.3222**
VAR residuals	-	2.3086*	2.0454*	2.6734**	3.1056**
EGARCH residuals	-	0.8033	1.0589	1.8989*	3.0707**

Note: Table 4 presents the test statistics for the Granger causality between DJIA returns and volume data. Results are shown for the linear Granger test, the DP test and the modified DP test. Bandwidth values are . The asterisks indicate significance at the 5% (*) and 1% (**) levels.

Table 5: Test Statistics for the Pairwise Granger Causality on Raw Exchange Returns

Pair		MA residuals		EGARCH residuals	
		DP	MDP	DP	MDP
JPY	AUD	4.0180**	3.0086**	1.9843*	1.2611
JPY	EUR	4.4724**	3.7586**	0.9336	0.5734
JPY	GBP	4.4096**	3.9775**	0.4305	0.3480
JPY	CHF	4.3236**	3.9867**	1.6542*	1.5474
AUD	JPY	4.4872**	3.6505**	2.0162*	1.4173
AUD	EUR	4.5398**	3.5291**	2.7414**	1.9737*
AUD	GBP	3.9936**	2.8616**	1.6532*	0.6208
AUD	CHF	3.2458**	3.2727**	1.5546	1.5006
EUR	JPY	4.0257**	3.2913**	1.1139	0.3133
EUR	AUD	3.7456**	3.1796**	1.5543	1.0551
EUR	GBP	5.3053**	4.3236**	3.0613**	2.2103*
EUR	CHF	5.5101**	4.6634**	3.8299**	3.5006**
GBP	JPY	4.2506**	3.4310**	0.3216	0.0284
GBP	AUD	4.7248**	4.0036**	2.3418**	1.9653*
GBP	EUR	4.7092**	3.9164**	1.3648	0.5487
GBP	CHF	2.7094**	2.2224*	2.0109*	1.6580*
CHF	JPY	4.0033**	3.4545**	0.9972	0.3293
CHF	AUD	3.3506**	2.5622**	0.8823	0.0981
CHF	EUR	3.8227**	2.6958**	1.6378	0.2864
CHF	GBP	3.6522**	3.0242**	1.8381*	1.5102

Note: Table 5 presents the statistics for pairwise Granger non-causality tests on high-frequency returns of five major currencies. The data are first cleaned by MA(1) component and seasonal component, and then standardized by EGARCH variance. Results are shown both for the DP test and the modified DP test with bandwidth $h = 0.2877$. The asterisks indicate significance at the 5% (*) and 1% (**) levels.