A Consistent Nonparametric Test for Granger Non-Causality Based on the Transfer Entropy

Cees G. H. Diks * [1 2 3] and Hao Fang[1 2 4]

[1]Center for Nonlinear Dynamics in Economics and Finance (CeNDEF), University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

[2]Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS, Amsterdam, The Netherlands

[3]The UvA Institute for Advanced Study, Oude Turfmarkt 147, 1012 GC, Amsterdam, The Netherlands

[4]Independent View, Strawinskylaan 1123, 1077 XX, Amsterdam, The Netherlands

Abstract

To date, testing for Granger non-causality using kernel density-based nonparametric estimates of the transfer entropy has been hindered by the intractability of the asymptotic distribution of the estimators. We overcome this by shifting from the transfer entropy to its first-order Taylor expansion near the null hypothesis, which is also non-negative and zero if and only if Granger causality is absent. The estimated Taylor expansion can be expressed in terms of a U-statistic, demonstrating asymptotic normality. After studying its size and power properties numerically, the resulting test is illustrated empirically with applications to stock indices and exchange rates.

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, Amsterdam School of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB, Amsterdam, The Netherlands. E-mail: C.G.H.Diks@uva.nl

1 Introduction

Characterizing causal interactions between time series has been challenging until Granger (1969) in his pioneering work brought forward the concept later known as Granger causality. Since then, testing causal effects has attracted attention not only in Economics and Econometrics, 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 over the last decades, with repeated debates on its validity. As we see it, there are at least two critical problems with parametric causality tests. First, being based on a classical linear VAR model, traditional Granger causality tests may overlook significant nonlinear dynamical relationships between variables. As Granger (1989) puts it, nonlinear models represent the proper way to model the real world which is 'almost certainly nonlinear'. Secondly, parametric approaches to causality testing bear the risk of model mis-specification. A wrong regression model could lead to a lack of power, or worse, unjustified conclusions. For example, Baek and Brock (1992) construct an example where nonlinear causal relations cannot be detected by a traditional linear causality test.

In a series of studies authors have tried to relax parametric model assumptions and provide nonparametric versions of Granger causality tests. Hiemstra and Jones (1994) were among the first to propose a formal nonparametric approach. By modifying the Baek and Brock (1992) nonparametric method, and developing asymptotic theory, Hiemstra and Jones proposed a nonparametric test to detect nonlinear causality in weakly dependent stochastic data. However, the Hiemstra-Jones test is suffering from a fundamental inconsistency problem, leading Diks and Panchenko (2006) to propose a new nonparametric test (hereafter referred to as the DP test) in a similar spirit. Alternative semiparametric and nonparametric tests include additive models used by Bell, Kay, and Malley (1996), the Hellinger distance measure used by Su and White (2008), and the empirical likelihood ratio-based test proposed by Su and White (2014).

The scope of this paper is to provide a novel test for Granger causality, based on the information theoretical notion of *transfer entropy* (hereafter TE), coined by Schreiber (2000). The transfer entropy was initially used to measure asymmetric information exchange in a bivariate system. By using appropriate conditional densities, the transfer entropy is able to measure

information transfer from one variable to another. This property makes it attractive for detecting conditional dependence in dynamical settings, and in fact corresponds to Granger causality in a general (distributional) sense. 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.

Despite the attractive properties of the transfer entropy and related information theoretical notions such as the mutual information, the application of concepts from information theory to time series analysis has proved difficult due to the lack of asymptotic theory for these information theoretical measures. For example, Granger and Lin (1994) utilize entropy to detect serial dependence using critical values obtained by simulation. Hong and White (2005) prove asymptotic normality for an entropy-based statistic, but the asymptotics only hold for a specific kernel function. Barnett and Bossomaier (2012) establish an asymptotic χ^2 distribution for transfer entropy estimators in parametric settings. Establishing asymptotic distribution theory for a fully nonparametric transfer entropy measure is challenging, if not impossible.

In this paper, we propose a test statistic based on a first order Taylor expansion of the transfer entropy, which is shown to be asymptotically normally distributed. Instead of deriving the limiting distribution of the transfer entropy – which is hard to track – directly, we bypass the problem by focusing on a quantity that locally (near the null hypothesis) is similar, but globally different, while still sharing the global positive-definiteness property with the transfer entropy. Furthermore, we show that this new test statistic is closely related to the DP test, and follow a similar approach to finding the asymptotic normal distribution of the estimator of the Taylor expansion.

This paper is organized as follows. Section 2 provides a short introduction to the nonparametric DP test and its lack of power against certain alternatives. Subsequently, the transfer entropy and a nonparametric test based on its first order Taylor expansion near the null hypothesis are introduced. The close linkage of this novel test statistic with the DP test is shown, and asymptotic normality is proved by using a U-statistic representation of the test statistic. Section 2 also discusses the optimal bandwidth selection rule for specific cases. Section 3 deals with Monte Carlo simulations; three different data generating processes are considered, enabling a direct comparison of size and power between the modified DP test and the DP test. Section 4 considers two financial applications. In the first, we apply the new test to stock volume and return data to make a direct comparison with the DP test; in the second application high

frequency exchange rates of main currencies are tested. Finally, Section 5 summarizes.

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 nonparametrically testing for conditional independence. We restrict ourselves to the bivariate setting as it is the most common implementation, although generalization to multivariate densities is possible.

Intuitively, for a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $t \in \mathbb{Z}$, it is said that $\{X_t\}$ Granger causes $\{Y_t\}$ if current and past values of $\{X_t\}$ contain some additional information, beyond that in current and past values of $\{Y_t\}$, about future values of $\{Y_t\}$. A linear Granger causality test based on a parametric VAR model can be seen as a special case where testing for conditional independence is equivalent to testing a restriction in the conditional mean specification.

In a more general setting, the null hypothesis of Granger non-causality can be rephrased in terms of conditional dependence between two time 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 that 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 by $\mathscr{F}_{Y,t}$ and $\mathscr{F}_{X,t}$, respectively, and use '$\sim$' to denote equivalence in distribution, we may give a formal and general definition for Granger causality. For a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $t \in \mathbb{Z}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if, for all $k \in 1, 2, \ldots$,

$$(Y_{t+1}, \ldots, Y_{t+k}) \mid (\ \mathscr{F}_{Y,t},\ \mathscr{F}_{X,t}\ ) \ \nsim\ (Y_{t+1}, \ldots, Y_{t+k}) \mid \mathscr{F}_{Y,t}.$$

In the absence of Granger causality, i.e.

$$(Y_{t+1}, \ldots, Y_{t+k}) \mid (\ \mathscr{F}_{Y,t},\ \mathscr{F}_{X,t}\ ) \ \sim\ (Y_{t+1}, \ldots, Y_{t+k}) \mid \mathscr{F}_{Y,t},$$

$\{X_t\}$ 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 $\{Y_t\}$ as

$$(Y_{t+1}, \ldots, Y_{t+k}) \perp (X_t, \ldots, X_{t-m}) \mid \mathscr{F}_{Y,t}, \tag{1}$$

for $m = 0, 1, 2, \ldots$ Granger non-causality, as expressed in Eq. (1), lays the first stone for a nonparametric test without imposing any parametric assumptions, apart from strict stationarity and weak dependence, about the data generating process or underlying distributions for $\{X_t\}$ and $\{Y_t\}$. The orthogonality here concerns not only the conditional mean, but also higher conditional moments. We assume two things here. First, $\{(X_t, Y_t)\}$ is a strictly stationary bivariate process. In practice it is infeasible in nonparametric settings to condition on the entire past of X_t and Y_t. We therefore implicitly consider the process to be of finite Markov orders, $l_X < \infty$ and $l_Y < \infty$ in the past of X_t and Y_t, respectively.

The null hypothesis of Granger non-causality is that H_0 : $\{X_t\}$ *is not a Granger cause of* $\{Y_t\}$. To keep focus on the main contribution of this paper – the Taylor expansion of the transfer entropy and its asymptotic distribution – in this paper we limit ourselves to the bivariate case with single lags in the past such that $k = l_X = l_Y = 1$, which so far has been the case considered most in the literature on nonparametric Granger non-causality.[1] We define the three-variate vector $W_t = (X_t, Y_t, Z_t)$, where $Z_t = Y_{t+1}$; and $W = (X, Y, Z)$ indicates a random variable W with distribution equal to the invariant distribution of W_t. Within the bivariate setting, W is a three-dimensional continuous vector. In terms of density functions $f_{\cdot}(\cdot)$ (which are assumed to exist), and given $k = l_X = l_Y = 1$, 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 all (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 all (x, y, z) in the support of W. A nonparametric test for Granger non-causality seeks to find statistical evidence of violation of Eq. (2) or Eq. (3). There are many nonparametric measures available for this purpose, some of which are mentioned above. However, as far as we

[1] Extensions to higher lags and/or higher-variate processes are feasible, but require reduction of the bias order by data sharpening or higher-order density estimation kernels (see e.g. Diks and Wolski, 2016).

know, the DP test, to be described below, currently is the only fully nonparametric test that is known to have correct asymptotic size under the null hypothesis of Granger non-causality.

2.2 The DP Test

Hiemstra and Jones (1994) proposed to test the condition expressed by Eq. (2) by calculating correlation integrals for each density and measuring the discrepancy between two sides of the equation. However their test is known to suffer from severe size distortion due to the fact that the quantity on which the test is based is inconsistent with Eq. (2). To overcome this problem, Diks and Panchenko (2006) suggest to use a conditional dependence measure by incorporating a local weight function $g(x, y, z)$ and formulating Eq. (3) as

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

Under the null hypothesis of no Granger causality, the term within the large round brackets vanishes, and the expectation goes to zero. As noted by Diks and Wolski (2016), Eq. (4) can be treated as an infinite number of moment restrictions. Although testing for Eq. (4) for a specific function g instead of testing 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, in the DP test the weight function $g(x, y, z)$ is taken to be $g(x, y, z) = f_Y^2(y)$, as this leads to a U-statistic representation of the corresponding estimator, which enables the analytical derivation of the asymptotic normality of the test statistic. In principle, other choices for $g(x, y, z)$ will also do as long as the test has satisfactory power against alternatives of interest. Since in the DP test, $g(x, y, z) = f_Y^2(y)$, it tests the implication

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

of H_0, rather than H_0 itself.

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 density function and h is the bandwidth, the DP test develops a third order

U-statistic estimator for the functional q, given by

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

where the normalization factor $(n-1)/(n(n-2))$ is inherited from the U-statistic representation of $T_n(h)$. It is worth mentioning that a second order square kernel $\mathbb{K}$ is adopted by Diks and Panchenko (2006). However there are two main drawbacks of using a square kernel. First, a square kernel will yield a discontinuous density estimate $\hat{f}(\cdot)$, which is not attractive from a practical perspective. Second, it weighs all neighbor points W_j equally, overlooking their relative distance to the estimation point W_i. Therefore, a smooth kernel function — the Gaussian kernel — is used here, namely the 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 a 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) prove the asymptotic normality of $T_n(h)$. Namely, if the bandwidth $h = h_n$ depends on the sample size as $h_n = Cn^{-\beta}$ for constants $C > 0$ and $\beta \in (\frac{1}{4}, \frac{1}{3})$, then the test statistic in Eq. (7) satisfies

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

where S_n^2 is a consistent estimator of the asymptotic variance of $T_n(h)$. Diks and Panchenko (2006) suggest to implement an one-sided version of the test, rejecting H_0': $q = 0$ against the alternative H_a: $q > 0$ if $T_n(h)$ is too large. That is, given the asymptotic critical value $z_{1-\alpha}$, the null hypothesis H_0' is rejected at significance level α if $\sqrt{n} T_n(h)/S_n > z_{1-\alpha}$.

The drawback of the DP test arises from the fact that H_0' in Eq. (5), obtained for a specific weight function $g(x, y, z)$, need not be equivalent to H_0 in Eqs. (2) and (3); it merely is an implication of H_0. For consistently testing H_0, an analogue of q is desirable that satisfies the positive definiteness property stated next, which q does not satisfy.

Property 1. *A functional s of the distribution of W is positive definite if $s \geq 0$ with $s = 0$ if and only if X_t and Z_t are conditionally independent given Y_t.*

From the previous reasoning, it is obvious that Eq. (5) is implied by Eq. (3), and Prop. 1 states that a strictly positive q is achieved if and only if H_0 is violated. In other words, the

null hypothesis of Granger non-causality requires that X_t and Z_t are independent conditionally on Y_t, which is just a sufficient, but not a 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 for $q = 0$ could deviate from the test on H_0. Although Diks and Wolski (2016) identified specific sub-classes of processes for which q is positive definite, we can easily construct a counterexample where the DP test has no power even if X_t strongly Granger causes Z_t. For completeness, such a counterexample is given next.

Inspired by the example in Skaug and Tjøstheim (1993), where a closely-related test for unconditional independence is proposed, we consider a conditional counterpart to illustrate that q is not positive definite. The one-sided DP test will be seen to suffer from a lack of power for this example process. As we show below, in the case where $q = 0$, this drawback cannot be overcome even with a two-sided DP test.

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

$$f(x_t, z_t|y_t) = f(x_t, z_t) = \begin{cases} 1 - 2d, & \text{if } 0 \le x_t \le 1, \quad 0 \le z_t \le 1, \\ d, & \text{if } 0 \le x_t \le 1, \ -1 \le z_t < 0, \\ d, & \text{if } -1 \le x_t < 0, \quad 0 \le z_t \le 1, \\ 0, & \text{if } -1 \le x_t < 0, \ -1 \le z_t < 0, \end{cases} \tag{9}$$

for $0 \le d \le \frac{1}{2}$.

Given Eq. (9), the marginal densities of X_t, Y_t and Z_t can be calculated to be all equal, with $P(-1 \le X_t < 0) = d$ and $P(0 \le X_t \le 1) = 1 - d$, while the conditional probability of Z_t being larger than zero given $\{(X_t = x_t, Y_t = y_t)\}$ is given by $P(0 < Z_t \le 1|(x_t, y_t)) = 1$ for $-1 \le x_t < 0$, and $P(0 \le Z_t \le 1|(x_t, y_t)) = (1 - 2d)/(1 - d)$ for $0 \le x_t \le 1$. Hence, for $0 < d \le \frac{1}{2}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ since X_t has an impact on the distribution of $Z_t = Y_{t+1}$, given Y_t. For this example we can explicitly calculate q defined in Eq. (5), which is found to be $q = d^2\left((1 - d)^3 + d^3\right)(4d - 1)$. For $0 < d < \frac{1}{4}$, q has a negative value. In this situation, the one-sided DP test, which rejects for large q, is not a consistent test for Granger

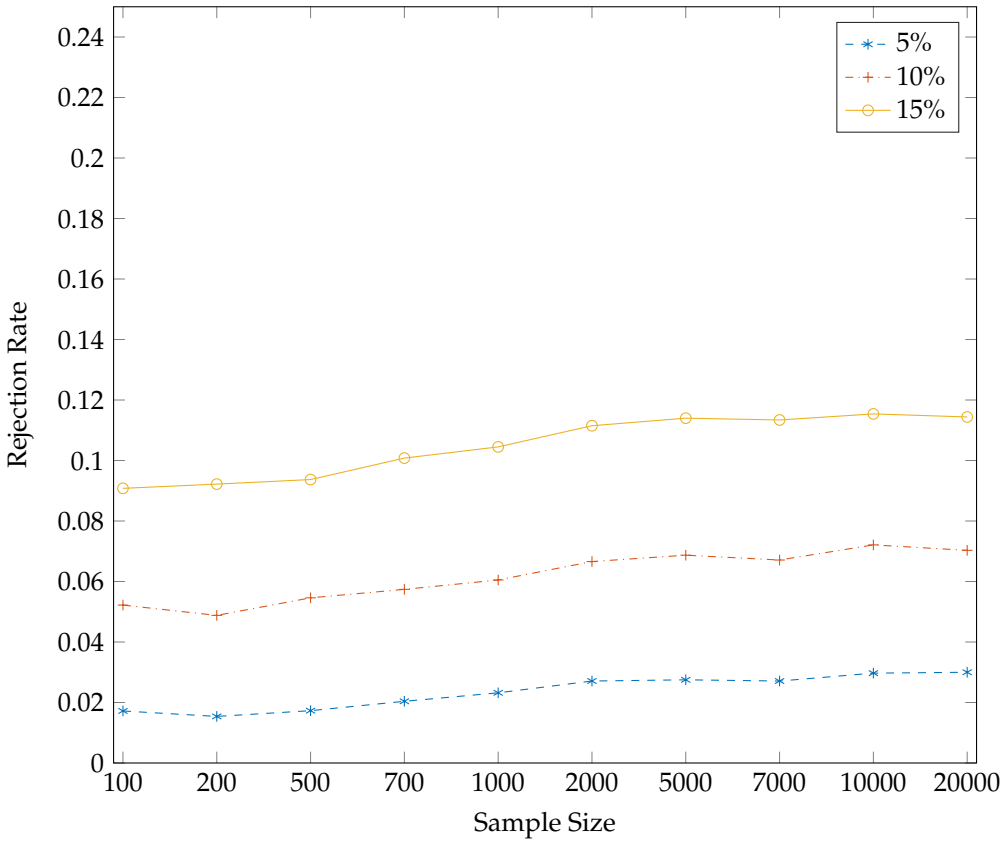


Figure 1: Power for the one-sided DP test for the artificial process $\{(X_t, Y_t, Z_t)\}$ given $q = 0$ at nominal size, from the bottom to the top, 5%, 10% and 15%, respectively, based on $10,000$ independent simulations.

non-causality. One may argue that this is not a problem if we use a two-sided test at the price of losing some power. However, the inconsistency of the DP test – which tests H_0' rather than H_0 – then would still be illustrated by the example if $d = \frac{1}{4}$, for which $q = 0$ exactly, while $\{X_t\}$ is clearly a Granger cause of $\{Y_t\}$; the DP test will only have trivial power against this alternative.

Fig. 1 reports the power of the one-sided DP test as a function of the sample size for different significance levels, based on $10,000$ independent simulations. Three nominal sizes are illustrated here: 5%, 10% and 15%, and the sample size ranges from 100 to $20,000$. It is striking from Fig. 1 that the DP test hardly has power against this alternative. The same conclusion can be drawn from Fig. 2, where the size-power plots are given. For almost all sub-panels with different sample sizes, the power of the DP test is around the diagonal line for this particular example when $q = 0$, which indicates that the DP test has only trivial power to detect Granger causality from X_t to Y_t.

The lack of power of the one-sided DP test in this example is hardly alleviated by its two-

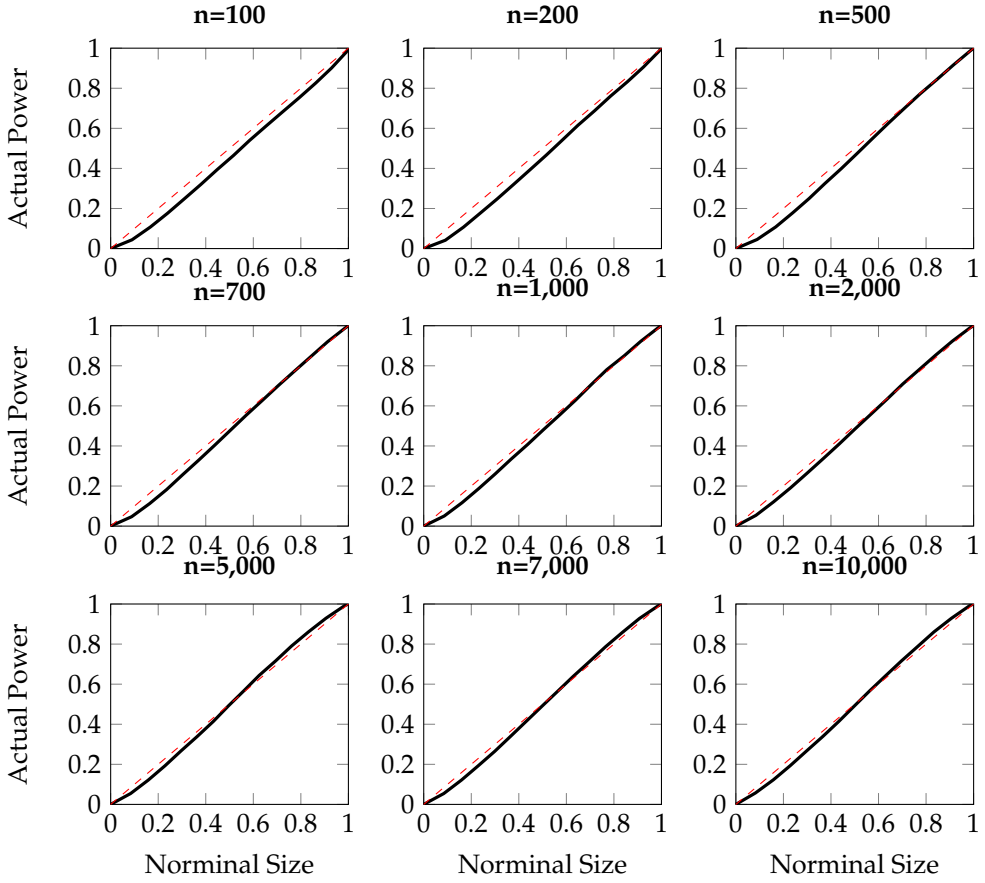


Figure 2: Size-power plots for the one-sided DP test for the artificial process $\{(X_t, Y_t, Z_t)\}$ with $q = 0$, based on $10,000$ independent simulations. Each subplot draws the actual power against the nominal size for different sample sizes, ranging from 100 to $10,000$. The solid curve represents the actual power and the red dash line indicates the diagonal, indicating the nominal size of a test.

sided counterpart, as a result of the absence of equivalence between $q = 0$ and conditional independence. The difference between H_0 and its implication H_0' gives rise to the lack of power of the DP test as the estimated quantity is not positive definite. In the next subsection, a new test statistic, based on the information-theoretical concept transfer entropy, is introduced and the test statistic is shown to be positive definite, which overcomes the inherited drawback of the DP test. In fact, this new test statistic shares many similarities with the DP test statistic, but also has an information-theoretical interpretation for its non-negativity.

2.3 Information-theoretical Interpretation

In a very different context from testing for conditional independence, the problem of information feedback and impact also has drawn much attention since 1950. Information theory, as a branch of applied mathematical theory of probability and statistics, studies the transmission of information over a noisy channel. This entropy, also referred to as Shannon entropy, is a key measure in the field of information theory brought forward by Shannon (1948, 1951). The entropy measures the uncertainty and randomness associated with a random variable. Suppose that S is a random vector with density $f_S(s)$, then the Shannon entropy is defined as

$$H(S) = -\int f_S(s) \log\{f_S(s)\} \, \mathrm{d}s.$$

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

Although inspiring, those results cannot be applied directly to measure conditional dependence. We therefore consider the transfer entropy (TE) introduced by Schreiber (2000) is a suitable measure to serve this purpose. The TE quantifies the amount of information explained in one series k steps ahead from the state of another series, given the information contained in

its own past. We briefly introduce the TE and KLIC before we further discuss its relation with the modified DP test.

Suppose that we have a bivariate process $\{(X_t, Y_t)\}$, and for brevity we put $X = \{X_t\}$, $Y = \{Y_t\}$ and $Z = \{Y_{t+k}\}$. Again we limit ourselves to $k = 1$ lag for simplicity, and consider the three-dimensional vector $W = (X, Y, Z)$ as before. The transfer entropy $\mathrm{TE}_{X\to Y}$ is a nonlinear and nonparametric measure for the amount of information contained in X about Z, in addition to the information about Z that already contained in Y. Although the TE defined by Schreiber (2000) applies to discrete variables, it is easily generalized to continuous variables. Conditional on Y, $\mathrm{TE}_{X\to Y}$ is defined as

$$\begin{aligned}
\mathrm{TE}_{X\to Y} &= E_W\left(\log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y) f_{Z|Y}(Z|Y)}\right) \\
&= \int\int\int f_{X,Y,Z}(x,y,z) \log \frac{f_{X,Z|Y}(x,z|y)}{f_{X|Y}(x|y) f_{Z|Y}(z|y)} \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z \\
&= E_W\left(\log \frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \log \frac{f_{X,Y}(X,Y)}{f_Y(Y)} - \log \frac{f_{Y,Z}(Y,Z)}{f_Y(Y)}\right) \\
&= E_W\left(\log f_{X,Y,Z}(X,Y,Z) + \log f_Y(Y) - \log f_{X,Y}(X,Y) - \log f_{Y,Z}(Y,Z)\right).
\end{aligned} \tag{10}$$

Using the conditional mutual information $I(Z, X|Y = y)$, the TE can be equivalently formulated in terms of four Shannon entropies as

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

In order to construct a test for Granger causality based on the TE, it remains to be shown that the TE is a proper basis for testing the null hypothesis. The following theorem, as a direct result of the properties of the KLIC, lays the quantitative foundation for testing based on the TE.

Theorem 1. *The transfer entropy* $\mathrm{TE}_{X\to Y}$ *is positive definite, that is,* $\mathrm{TE}_{X\to Y} \geq 0$ *with equality if and only if* $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y) f_{Z|Y}(Z|Y)$.

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

It is not difficult to verify that the condition for $\mathrm{TE}_{X\to Y} = 0$ coincides with Eqs. (2) and (3) for Granger non-causality under the null hypothesis. This positive definiteness makes $\mathrm{TE}_{X\to Y}$ a desirable measure 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 $\mathrm{TE}_{X\to Y}$, one may follow the recipe of Kraskov, Stögbauer, and Grassberger (2004) by measuring k-nearest neighbor distances. A more natural method, applied in this paper, is to use the plug-in kernel estimates given in Eq. (6), and replace unknown expectations by sample averages.

However, the direct use of the TE to test Granger non-causality is not easy due to the lack of asymptotic theory for the test statistic. As shown by Granger and Lin (1994), the asymptotic distribution of entropy-based estimators usually depends on strict assumptions regarding the dataset. Over the years several break-throughs have been made with the application of entropy to testing serial independence, e.g. Robinson (1991) obtains an asymptotic $N(0,1)$ distribution for an entropy measure by a sample-splitting technique and Hong and White (2005) derive asymptotic normality under bounded support data and quartic kernel assumptions. However, the limiting distribution of the natural nonparametric TE estimator is still unknown under more general conditions.

One may argue in favor of using simulation techniques to overcome the problem of the lack of asymptotic dtheory. However, as suggested by Su and White (2008), there exist estimation biases of TE statistics for non-parametric dependence measures under the smoothed bootstrap procedure. Even with parametric test statistics, Barnett and Bossomaier (2012) notice that the TE-based estimator is generally biased. Surrogate data are 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 transfer. We therefore consider the direct usage of the TE for non-parametric tests for Granger non-causality difficult, if not impossible.

Below we show that a first order Taylor expansion of the TE provides a way out to construct the asymptotic distribution of this meaningful information measure. In the next section, we show that the first order Taylor expansion of the TE can form the basis of a modified DP test for conditional independence. This not only helps to circumvent the problem of asymptotic distribution for entropy based statistic, but also endows the modified DP test with positive definiteness.

In the remaining part of this section we will introduce the first order Taylor expansion of

the TE, and the positive definiteness of the measure will be given afterwards. Starting with Eq. (10), we perform the first order Taylor expansion locally at $\mathrm{TE}_{X\to Y} = 0$, which is

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

where 'h.o.t' stands for 'higher order terms'. By ignoring higher order terms, we define the first order expansion $t = E_W\left[\frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}-1\right]$ as a measure for conditional dependence. The following theorem states that t inherits the positive definiteness of the TE.

Theorem 2. *The transfer entropy t is positive definite, that is, $t \geq 0$ 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 measure t has the desirable property of positive definiteness, which is absent for the measure q of the DP test. However, direct estimation of Eq. (11) does not lead to a practically useful test statistic without the asymptotic distribution. In the next subsection we show that the nonparametric estimator of t is asymptotically normal. The key to this result is the fact that the DP statistic and the present statistic only differ in terms of the weight function $g(x,y,z)$ in Eq. (4), and that the proof of asymptotic normality of the DP test can be easily adjusted to accommodate this new weight function.

2.4 A Modified DP Test

In comparing Eqs. (3) and (4), it can be seen that the discrepancy between H_0' and H_0 arises from incorporation of the weight function $g(x,y,z) = f_Y^2(y)$ to the null hypothesis. In principle, a different positive function $g(x,y,z)$ will also work, such as those discussed by Diks and Panchenko (2006). As long as the weight function allows for a U-statistic representation of the corresponding estimator, the asymptotic normality is maintained. Particularly, we propose to modify the DP test by dividing all terms in the expectation of Eq. (5) by a function $v(x,y,z)$ given by $v(x,y,z) = (f_{X,Y}(x,y)f_{Y,Z}(y,z))$. Similar to the discussion in Section 2.2, we test an implication of H_0 rather than H_0 itself, but will choose $v(x,y,z)$ so that H_0'' and H_0 coincide.

This new implication of H_0 has the form

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

One can also think of Eq. (12) as the result of plugging in a different weight 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) and hence to H_0 by Thm. 2. To estimate t, we propose to use the following statistic with density estimator defined in Eq. (6):

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

The reason for estimating t in this form is that, with the sample statistic $T_n'(h)$, we can obtain a degree three U-statistic representation of t, similar to that for the DP test statistic, by which asymptotic normality follows. Before introducing the formal theorem, Thm. 3 shows that the modified DP statistic $T_n'(h)$ generically is non-degenerate thus may be used for performing a statistical test based on the usual first-order asymptotics. In other words, although the statistic $T_n'(h)$ may be degenerate in some special cases, these are shown to be non-generic.

Theorem 3. *The limiting distribution of $T_n'(h)$ generically is non-degenerate.*

Proof. See Appendix A.3. □

The asymptotic normality of $T_n'(h)$ is stated in Thm. 4 below, which relies on the following two lemmas concerning the uniform consistency of density estimators.

Lemma 1. *Let $\{W_i\} = \{(X_i,Y_i,Z_i)\}, i \in \mathbb{N}$ be a sequence of k-dimensional random variables with Lebesgue density f. For the estimation of f, based on the first n values W_i, we use the kernel density estimator $f_n = \hat{f}$ with kernel function $\mathbb{K}(w)$, as given in Eq. (6). If $f(w)$ is*

continuous at $w \in \mathbb{R}^k$ *and* $\mathbb{K}(w)$ *is of bounded variation, then*

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

provided that any of the conditions (A1) or (A2) hold, where

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

$$\sum_{i=1}^{\infty} e^{-\gamma i h_n^{2k}} < \infty, \text{ for all } \gamma \in R_+,$$

$$\text{or } \left(\frac{\log\log i}{i}\right)^{1/2k} = o(h_n).$$

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

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

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

Lemma 1 provides the uniform consistency 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 to serve our purpose here we need uniform convergence, which is stronger than pointwise convergence. We refer to Wegman (1972) and Wied and Weißbach (2012) for a detailed discussion between different types of convergence.

We next consider $\tilde{T}'_n$, which differs from T'_n in Eq. (14) only in having $\hat{v}(.)$ in the denominator replaced by the true unknown function $v(.)$. In the next lemma, the short-hand notation $v_i = v(W_i)$ and $\eta_i = \eta(W_i) = f_{X,Y,Z}(X_i, Y_i, Z_i) f_Y(Y_i) - f_{X,Y}(X_i, Y_i) f_{Y,Z}(Y_i, Z_i)$ is used.

Lemma 2. *If condition (A1) and/or (A2) of Lemma 1 hold, and* $\text{Var}\left(\frac{\eta_i}{v_i}\right) < \infty$*, then* $\sqrt{n}(T'_n(h_n) - t)$ *and* $\sqrt{n}(\tilde{T}'_n(h_n) - t)$ *have the same limiting distribution provided that it exist. More formally stated,*

Proof. See Appendix A.4. □

Theorem 4. *For a time series* $\{(X_t, Y_t)\}_{t=1}^n$ *of length* n *observed from a weakly dependent stationary bivariate process, and given that we use a joint Gaussian density estimation kernel*

with bandwidth $h_n = Cn^{-\beta}, C > 0, \beta \in (\frac{1}{4}, \frac{1}{3})$,

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

where S_n^2 *is a HAC estimator of the long-run variance* σ^2 *of* $\sqrt{n}(T'_n(h) - t)$.

Proof. See Appendix A.5. □

We wish to comment on Eq. (14) regarding the treatment of the marginals. Note that t is invariant under invertible smooth transformations of the marginals due to the form of Eq. (13) assuming that X_t and Y_t are continuous (the ratio of densities of the same variables is invariant under marginal transforms). Therefore, the dependence structure between X_t and Y_t remains intact under invertible marginal transforms. Although our testing framework does not depend crucially on the restrictive assumption of a uniform distribution for the time series as in Pompe (1993) and Hong and White (2005), we recommend to use the probability integral transformation (PIT) on each of the marginals, which usually improves the performance of statistical dependence tests, as Diks and Panchenko (2006) suggest. The reason is that, contrary to directly calculating the test statistics on the original data, the bounded support after transforming the marginals to a uniform distribution avoids non-existing moments during the bias and variance evaluation, which helps to stabilize the test statistic. There are alternative ways to transform the marginal variables into a bounded support, for example, by using a logistic function as Hong and White (2005). Here we decided to just apply the PIT, as it doesn't require any user-specified parameters, and always leads to identical (uniform) marginals. 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.

Since the transfer entropy-based measure t is non-negative, tests based on the statistic $T'_n(h)$ are implemented as one-sided tests, rejecting the null hypothesis if $\sqrt{n}T'_n(h)/S_n > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of standard Normal distribution for a given significance level α.

2.5 Bandwidth Selection

In nonparametric settings, there typically is no uniformly most powerful test against all alternatives. Hence it is unlikely that a uniformly optimal bandwidth exists. As long as the

bandwidth tends to zero with as $h = Cn^{-\beta}, C > 0, \beta \in (\frac{1}{4}, \frac{1}{3})$, our test has unit asymptotic power. Yet, we may define the optimal bandwidth in the sense of asymptotically minimal mean squared error (MSE). When balancing the first and forth leading terms in Eq. (A.10) to minimize the squared bias and variance, for a second order kernel, it is easy to show that the optimal bandwidth for the DP test is given by

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

with q_2 and $\mathrm{E}[s(W)]$ the series expansion for the second moment of kernel function and expectation of bias, respectively. Since the convergence rate of the MSE, derived in Appendix A.5, is not affected by the way we construct the new test statistic, the derivation of Eq. (15) remains intact and we calibrated the optimal bandwidth for our new test, finding

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

where the scale factor 0.6 involved is a result of bias and variance adjustment for replacing the square kernel by the Gaussian kernel (the variance of the uniform DP kernel was $1/\sqrt{3} \approx 0.57735$, which we rounded off to 0.6). Intuitively, the q_2 and $\mathrm{E}[s(W)]$ terms are different from those for the 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 a (G)ARCH process, the optimal bandwidth is approximately given by $h_{DP} = Cn^{-2/7}$ where $C \approx 8$. Applying Eq. (16), we proceed with $h^* = 4.8n^{-2/7}$ for (G)ARCH processes. To gain some insights into the bandwidth, we illustrate the test size and power with a 2-variate ARCH process, given by

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

We let $0 < a < 0.4$ and run 5,000 Monte Carlo simulations for time series length varying from 200 to 5,000. The size is assessed based on testing Granger non-causality from $\{X_t\}$ to $\{Y_t\}$, and for the power we use the same process but testing from Granger non-causality from $\{Y_t\}$ to $\{X_t\}$. The results are presented in Table 1, from which it can be seen that the modified DP test is conservative in the sense that its empirical size is lower than the nominal size 0.05

in all cases, while the power increases when a increases and when the sample size increases.

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

	n	200	500	1,000	2,000	5,000
	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: Empirical size and power of the modified DP test for the process given in Eq. (17) for different sample sizes and parameter a. The values represent observed rejection rates over $5,000$ realizations for nominal size 0.05.

3 Size/power 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) for which the DP test fails to detect the impact of X on Y. The modified DP test is performed with $10,000$ replications, with the same bandwidth. The counterpart of the power-size plots for the DP test in Fig. 2 is delivered in Fig. 3. In contrast with the lack of power of the 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.

Next, we use numerical simulations to study the behavior of the modified DP test, while direct comparisons between the modified DP test T'_n and the DP test T_n are also given. Three processes are being considered. In the first experiment, we consider a simple bivariate VAR process, given by

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

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

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

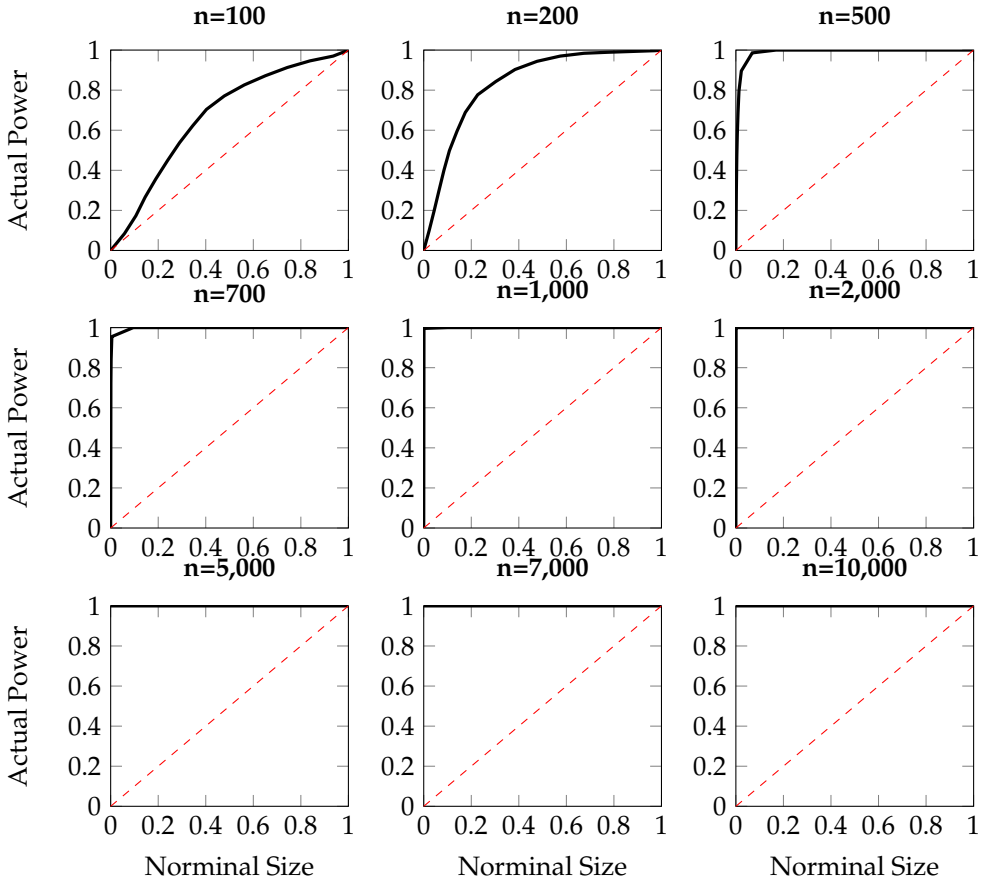


Figure 3: Size-power plots for the one-sided modified DP test for the artificial process $\{(X_t, Y_t, Z_t)\}$ with $q = 0$, based on $10,000$ independent replications. Each subplot shows the observed power against the nominal size for different sample sizes, ranging from 100 to $10,000$. The solid curve represents the observed power and the red dashed line corresponds with the diagonal, indicating the nominal size of the test.

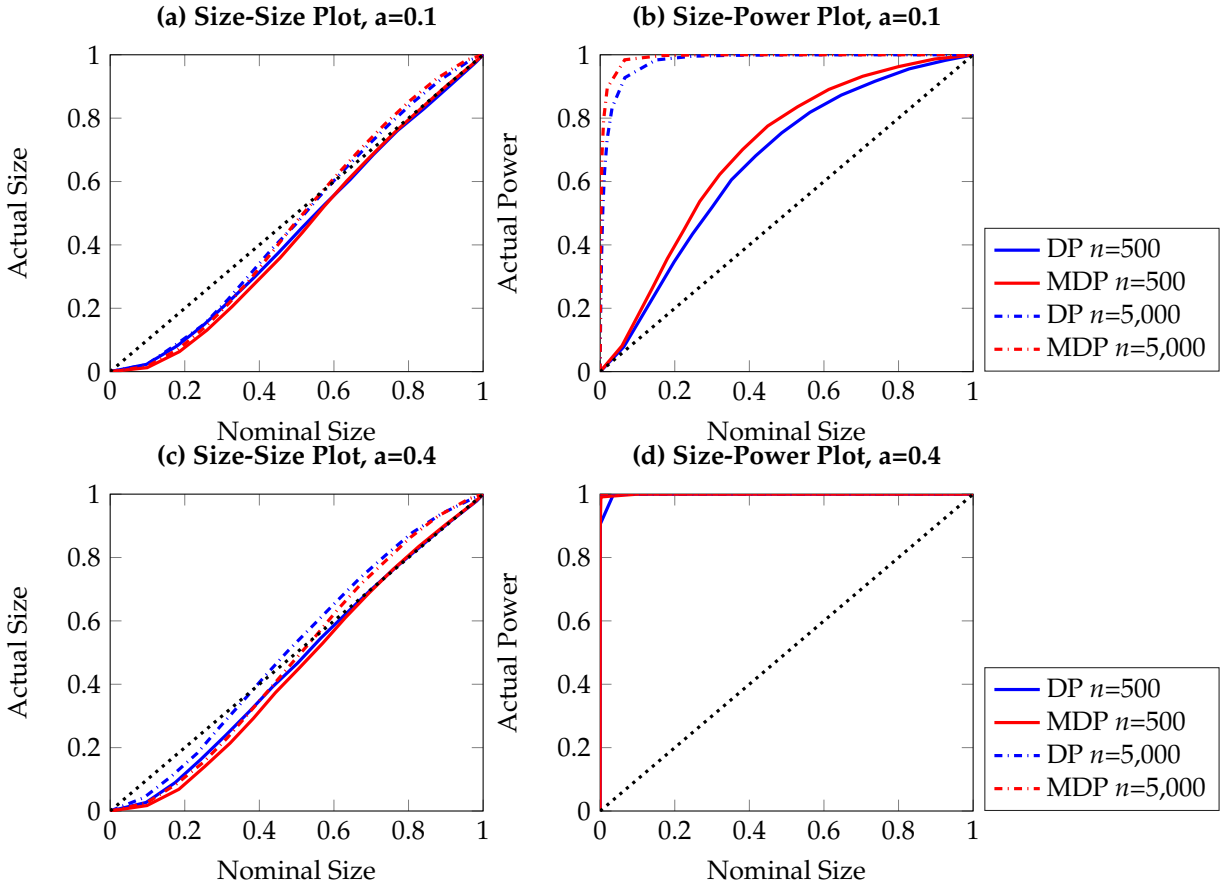


Figure 4: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is the bivariate linear VAR of Eq. (18), with Y affecting X. The left (right) column shows observed rejection rates under the null (alternative), the blue lines stands for the DP test while the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

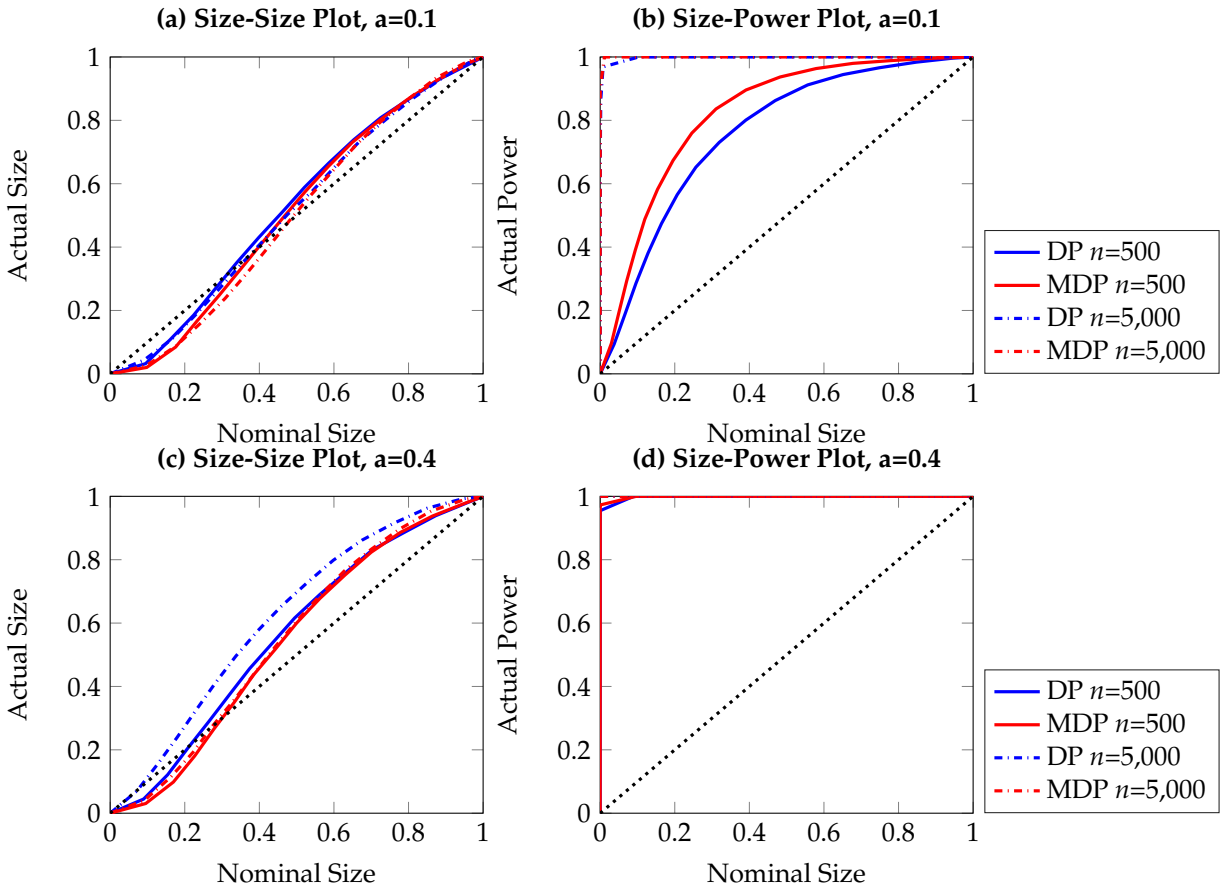


Figure 5: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is bivariate Non-linear VAR as in Eq. (19), with Y affecting X. The left (right) column shows observed rejection rates under the null(alternative), the blue lines stands for DP test while the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

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

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

The results, which are shown in Figs. 4 to 6, are obtained with 5,000 simulations for each process. We present the DP test and the modified DP test with both the empirical size-size and size-power plots for the three processes in Eqs. (18) to (20) for sample size $n = 500$ and $n = 5,000$, respectively. The control parameter a is considered to take the values 0.1 and 0.4. As before, the empirical size is obtained by testing for Granger non-causality from $\{X_t\}$ to $\{Y_t\}$, and the empirical power is the observed rejection rate of testing for Granger non-causality from $\{Y_t\}$ to $\{X_t\}$.

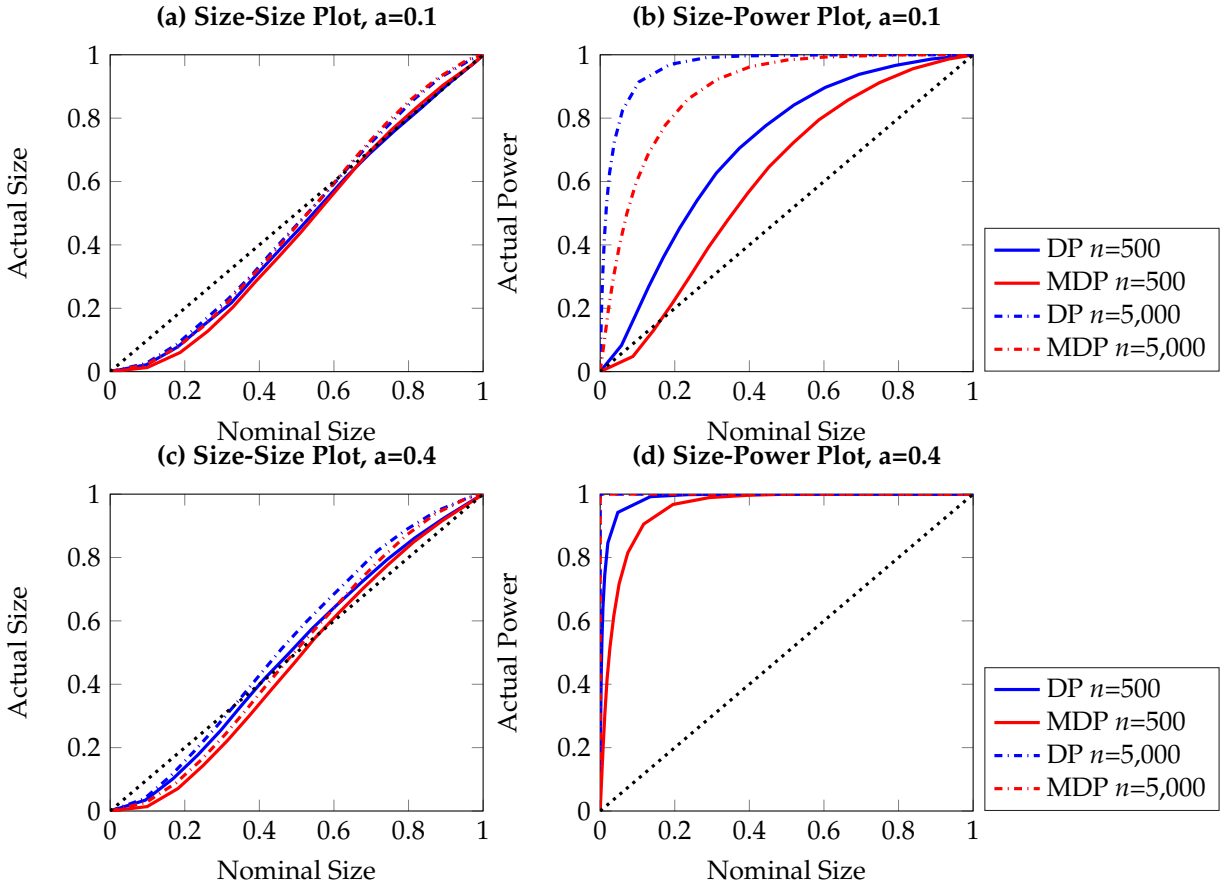


Figure 6: Size-size and size-power plots of Granger non-causality tests, based on 5,000 simulations. The DGP is bivariate ARCH as in Eq. (20), with Y affecting X. The left (right) column shows observed rejection rates under the null(alternative), the blue lines stands for DP test while the red lines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

It can be seen from Figs. 4 to 6 that the modified DP test is slightly more conservative than the DP test under the null hypothesis. However the size distortion reduces when the sample size increases. The modified DP test is more powerful than the DP test in the linear and nonlinear VAR settings given in Eqs. (18) and (19). Overall, we see that the larger the sample size and the stronger the causal effect are, the better the asymptotic performance of the modified DP test is.

4 Empirical Illustration

4.1 Stock Volume–Return Relation

In this section, we first revisit the stock return-volume relation considered by Hiemstra and Jones (1994) and Diks and Panchenko (2006). This topic has a long research history. Early empirical work mainly focused on the positive correlation between volume and stock price change, see Karpoff (1987). The later literature exposed 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 a 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. The daily volume and adjusted daily closing prices were obtained from Yahoo Finance. The time series were converted by taking log returns 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 performed a two-stage adjustment process, similar to the procedure applied in Hiemstra and Jones (1994)[2]. We apply our test not only to the raw data, but also on VAR filtered residuals and EGARCH(1,1,1) filtered residuals.[3] The idea of filtering is to remove linear dependence

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

[3]We have tried different error distributions like Normal, Students't, GED and Hansen (1994)'s skewed t.

and the effect of heteroskedasticity to isolate the nonlinear and higher moment relationships among series, respectively.

Tables 2 to 4 report the resulting t statistics for both the 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, after standardization, where the latter value roughly corresponds to the derived optimal bandwidth ($h = 0.6138$) and the larger bandwidth, also used in Diks and Panchenko (2006), is added as a robustness check.

Table 2: Test Statistics for the S&P500 returns and volume data. 'MPD' stands for 'modified DP'.

	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: Test statistics for Granger causality between S&P500 returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. 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: Test statistics for Granger causality between NASDAQ returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. The asterisks indicate significance at the 5% (*) and 1% (**) levels.

Generally speaking, the results indicate that the effect in the return-volume direction is stronger than vice versa. 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, the linear Granger

The differences caused by different distributional assumptions are small; we only report the results based on the Students't distribution due to space considerations.

Table 4: Test Statistics for the DJIA returns and volume data. 'MDP' stands for 'modified DP'.

	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: Test statistics for the Granger causality between DJIA returns and volume data. Results for bandwidth values 1.5 and 0.6 are reported. The asterisks indicate significance at the 5% (*) and 1% (**) levels.

test finds no evidence of causal impact while the nonparametric tests claim strong causal effect except for the DJIA where only the modified DP test finds a causal link from volume to return. As argued above, the results for the linear test are suspicious since it only examines linear causal effects in the conditional mean; information exchange from higher moments is completely ignored.

A 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 a strong causal effect in two directions for S&P and NASDAQ, but for the DJIA, the t-statistics of the DP test are weaker than those of the modified DP test. The bi-directional causality between return and volume remains unchanged after linear VAR filtration, although the DP test again shows weaker evidence. The result also suggests that the causality is strictly nonlinear. The linear test (F-test) is unable to spot these nonlinear linkages.

Further, in the direction from Volume to Return, these nonlinear causalities tend to vanish after EGARCH filtering. Thus the bi-directional linkage is reduced to a one-directional relation from return to volume. The modified DP statistics, however, are in general larger than the DP t-values, and indicate more causal relations. In contrast with the DP test, our test suggests that the observed nonlinear causality cannot be completely attributed to second moment effects. Heteroskedasticity modeling may reduce this nonlinear feature to some extent, but its impact is not as strong as the DP test would suggest.

4.2 Application to Intraday Exchange Rates

In the second application, we apply the modified DP test to intraday exchange rates. We consider five major currencies: JPY, AUD, GBP, EUR and CHF, all against the 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 26,496 high frequency observations. We use 5-minute data, corresponding to the sampling frequency of 288 quotes per day, which is high enough to avoid measurement errors (see Andersen and Bollerslev (1998)) but also low enough for the micro-structure not to be of major concern.

Although the foreign exchange market is one of the most active financial markets in the world, where trading takes place 24 hours per day, 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 differences between the log prices at consecutive times are obtained. Next, we remove the conditional mean dynamics by fitting an MA(1) model and using the residuals as our return series following Bollerslev and Domowitz (1993). Finally, intraday seasonal effects are filtered out using estimated time-of-day dummies following Diebold, Hahn, and Tay (1999), i.e.

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

where $r_{i,n,t}$ denotes intraday log returns after MA(1) filtering. The subscript $i = 1, \ldots, 5$ indicates five different currency and n, t stands for time t on 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 is 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 $\sum_{t=1}^{T} d_{i,t} = 1$. Figs. 7 to 9 report the first 200 autocorrelations of returns, absolute returns and squared returns, when checking on the raw series, MA(1) residuals and EGARCH residuals, respectively.

We perform pairwise nonparametric Granger causality tests on the MA(1) filtered and seasonally adjusted data, as well as on the standardized residuals after EGARCH(1,1,1) filtering. We use Hansen (1994)'s skewed t distribution to model the innovation terms. We choose a bandwidth of 0.2768, according to Eq. (16).

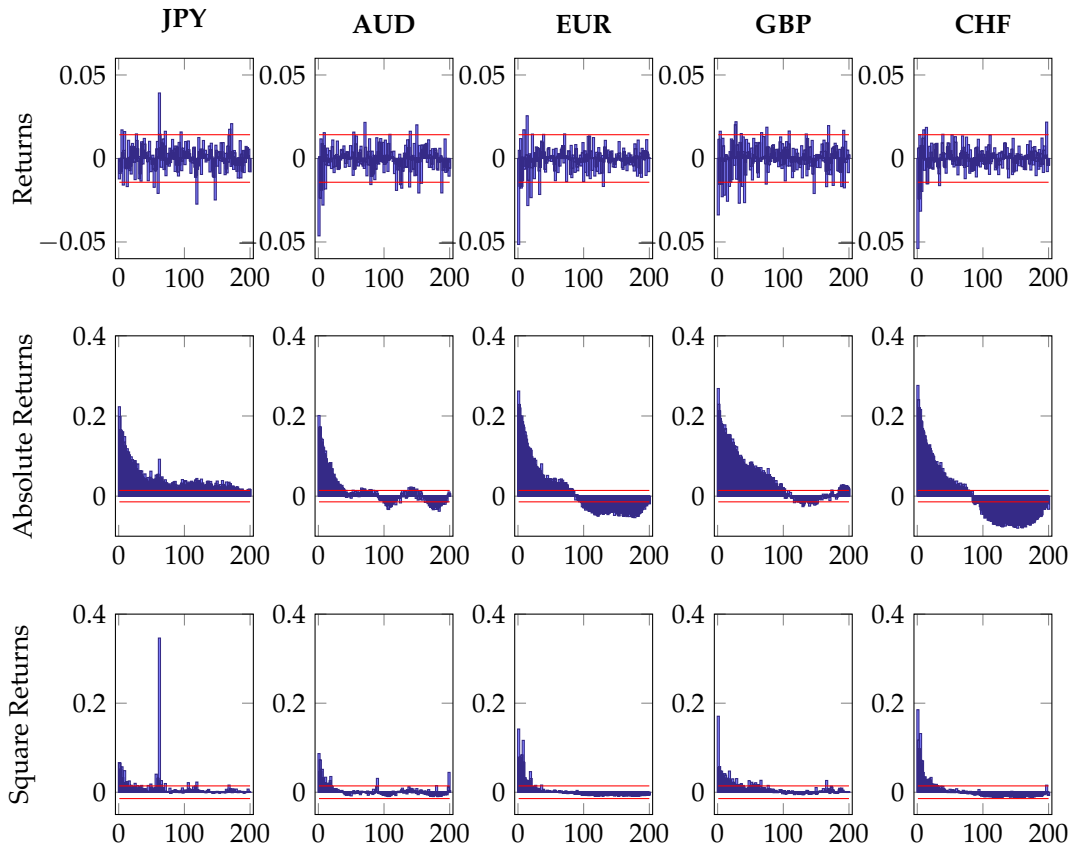


Figure 7: Autocorrelations of Returns, Absolute Returns and Square Returns, up to 200 lags.

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: Statistics for pairwise Granger non-causality tests on high-frequency returns of five major currencies. The data are first cleaned by the MA(1) and seasonal components, and then standardized based on the EGARCH conditional 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.

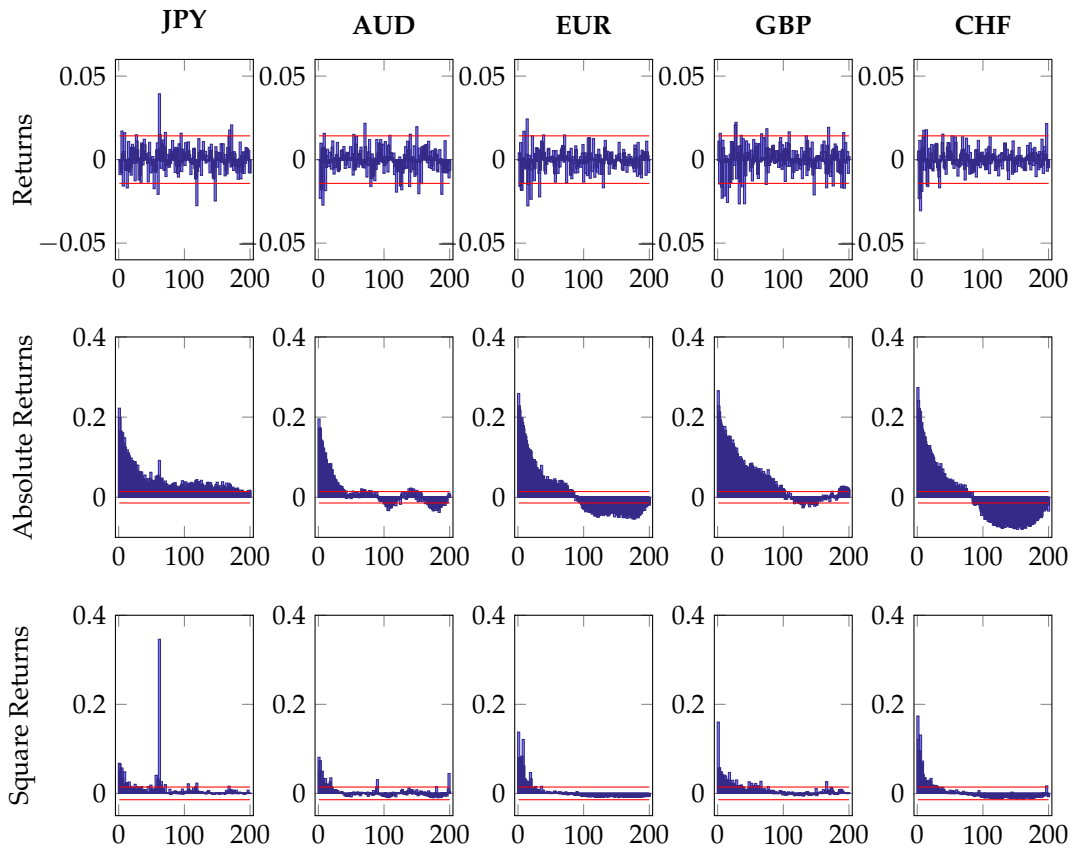


Figure 8: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) Component removed, up to 200 lags.

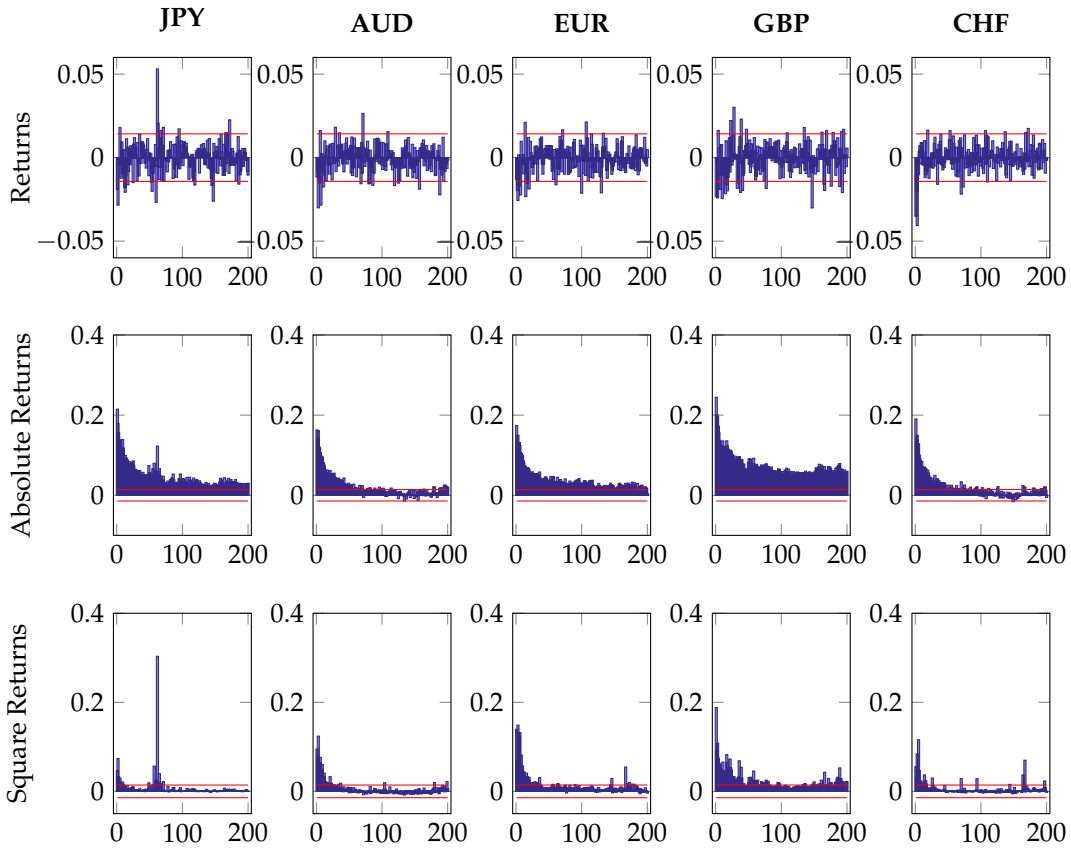


Figure 9: Autocorrelation of Returns, Absolute Returns and Square Returns after MA(1) and GARCH filtering, up to 200 lags.

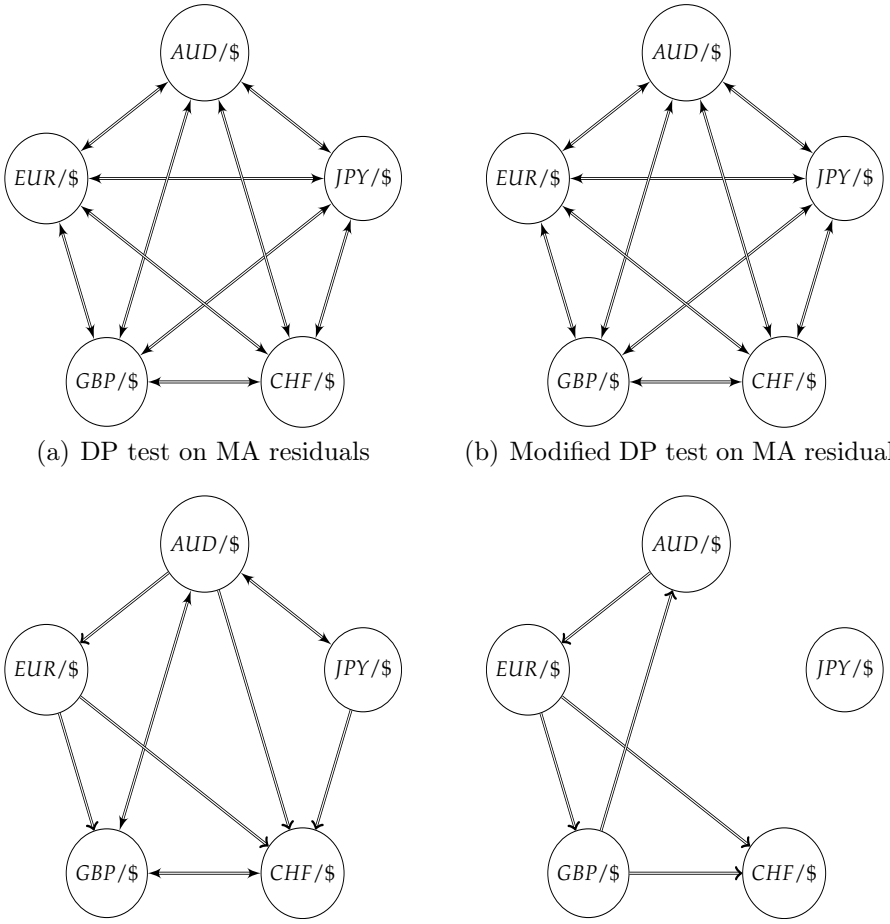
The test results are shown in Table 5 for both MA(1) de-meaned and de-seasoned data, as well as EGARCH filtered data. Although not reported here, there is statistical evidence for strong bi-directional causality among all currency pairs on raw return data at 5-minute lag. These bi-directional causalities do not disappear after removing the MA(1) component and seasonal component. However, the observed information spillover is significantly weaker after the EGARCH filtering. When testing based on the EGARCH standardized residuals, only a few pairs still show signs of a strong causal relation. Especially, the directional relation of EUR $\rightarrow$ CHF is the only one detected by both the DP test and the modified DP test at the 1% level of significance. A graphical representation is provided in Fig. 10, where one can clearly see that most causal links are gone after EGARCH filtering. The modified DP test exposes five uni-directional linkages among the EGARCH filtered returns at the 5% level. The EUR and GBP are the most important driving currencies. While the DP test also admits the importance of JPY and particularly AUD, which shows bi-directional causality between JPY and GBP.

To sum up, we find evidence of strong causal links among exchange returns at an intraday high-frequency timescale. Each currency has predictive power for other currencies, implying high co-movements in the international exchange market. Although those directional linkages are not affected by the de-meaning procedure, we may reduce most of them by taking the volatility dynamics into account. When filtering out heteroskedasticity by EGARCH estimation, there only exist a few pairs containing spillover effects.

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 asymptotic normality of the test statistic is derived by taking advantage of a U-statistic representation, similar to that applied in the DP test. The modified DP statistic, however, improves the DP statistic in at least the two respects: firstly, the positive definiteness of the quantity on which the test statistic is based, paves the way for properly testing for differences between conditional densities; secondly, the weight function in our test is motivated from an information-theoretical point of view, while the weight function in the DP test was selected in an ad hoc manner.

The simulation study confirms that the modified DP test has good size and power properties for a wide range of data generating processes. In the first application, a direct comparison with



(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: Graphical representation of pairwise causalities on MA and seasonally filtered residuals, as well as EGARCH filtered residuals. The arrows in the graph indicate a directional causality at the 5% level of significance.

the DP test confirms that the DP test may suffer from a lack of power for specific processes, while the second application to high frequency exchange return data helps us better understand whether the spillover channel in exchange rate markets arises from conditional mean, conditional variance or higher conditional moments. Some obvious extensions to future work include the incorporation of additional lags of the variables and a generalization to higher-variate settings to allow for conditioning on additional, possibly confounding, variables.

Acknowledgements

The authors thank seminar participants at the University of Amsterdam and the Tinbergen Institute, as well as participants of 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) and the 2017 International Association for Applied Econometrics Conference (Sapporo, Jun 26-30, 2017). The authors wish to extend particular gratitude to Simon Broda and Chen Zhou for their comments on an earlier version of this paper; and to SURFsara for providing the LISA cluster. Hao Fang is grateful to SNDE and IAAE for their conference support.

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, 1–5.

Barrett, A. B., L. Barnett, and A. K. Seth (2010): "Multivariate Granger causality and generalized variance," *Physical Review E*, 81(4), 041907, 1–14.

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.

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

——— (1986): "Rigorous statistical procedures for data from dynamical systems," *Journal of Statistical Physics*, 44, 67–93.

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, Berlin.

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 (2016): "Nonlinear granger causality: Guidelines for multivariate analysis," *Journal of Applied Econometrics*, 31(7), 1333–1351.

Ding, M., Y. Chen, and S. L. Bressler (2006): "Granger causality: Basic theory and application to neuroscience," vol. Handbook of time series analysis: Recent theoretical developments and applications, chap. 17, pp. 437–460. John Wiley & Sons, New York.

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*, 37(3), 424–438.

Granger, C. W. J. (1989): *Forecasting in Business and Economics*. Academic Press, New York.

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

Hansen, B. E. (1994): "Autoregressive conditional density estimation," *International Economic Review*, 35(3), 705–730.

——— (2009): "Lecture notes on nonparametrics," *Lecture Notes, University of Wisconsin, Madison*.

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(1), 109–126.

Kraskov, A., H. Stögbauer, and P. Grassberger (2004): "Estimating mutual information," *Physical review E*, 69(6), 066138, 1–16.

Kullback, S. (1968): *Information Theory and Statistics*. Courier Corporation, New York.

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*, 39(3), 243–250.

Schreiber, T. (2000): "Measuring information transfer," *Physical Review Letters*, 85(2), 461–464.

Schuster, E. F. (1969): "Estimation of a probability density function and its derivatives," *The Annals of Mathematical Statistics*, 40(4), 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, New York.

Skaug, H. J., and D. Tjøstheim (1993): "Nonparametric tests of serial independence," in *Developments in Time Series Analysis*, ed. by T. S. Rao, Developments in time series analysis, 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(4), 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*. Chapman & Hall/CRC, New York.

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 1–19.

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 of Positive-definiteness of the Transfer Entropy

According to the definition of TE in Eq. (14), the expectation over the logarithm of the density ratio is evaluated: $\mathrm{TE}_{X\to Y} = E_W\left(\log \frac{f_{Z,X|Y}(Z,X|Y)}{f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)}\right)$. 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 the TE as $\mathrm{TE}_{X\to Y} = E\left(-\log(R)\right)$. 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

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

On 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 have proved that $\mathrm{TE}_{X\to Y} \equiv -E\left(\log(R)\right) \geq 0$. It is obvious that the equality holds if and only if $R = 1$, which is equivalent to $f_{Z,X|Y}(Z,X|Y) = f_{X|Y}(X|Y)f_{Z|Y}(Z|Y)$. This completes the proof of Thm. 4.

A.2 Proof of Positive-definiteness of the First Order TE Statistic

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

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

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 support of (X, Y, Z). In the fourth step we completed the square by using the fact that the integral over the whole support of the pdf is 1. Finally, $t \geq 0$ follows naturally from the integrand being non-negative.

A.3 Proof of Non-Degeneracy of the Modified DP Test Statistic

To show that the asymptotic normality in Thm. 4 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 non-degenerate.

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

$$\begin{aligned} K(w_1, w_2, w_3) = &\left[(\kappa_{XYZ}(w_1 - w_2)\kappa_Y(w_1 - w_3) - \kappa_{XY}(w_1 - w_2)\kappa_{YZ}(w_1 - w_3))/v(w_1)\right]/6 \\ &+ \text{permutations of } w_1, w_2, w_3, \end{aligned} \tag{A.4}$$

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.4), we obtain

$$
\begin{aligned}
r_1(w_1;h) =& E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(w_1-W_3)\kappa_Y(w_1-W_2)-\kappa_{XY}(w_1-W_3)\kappa_{YZ}(w_1-W_2)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-w_1)\kappa_Y(W_3-W_2)-\kappa_{XY}(W_3-w_1)\kappa_{YZ}(W_3-W_2)\right)/v(w_1)\right]/6 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/6 \\
=& E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]/3 \\
&+E\left[\left(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3)\right)/v(w_1)\right]/3 \\
&+E\left[\left(\kappa_{XYZ}(W_3-W_2)\kappa_Y(W_3-w_1)-\kappa_{XY}(W_3-W_2)\kappa_{YZ}(W_3-w_1)\right)/v(w_1)\right]/3 \\
\equiv& E_1(w_1;h)/3+E_2(w_1;h)/3+E_3(w_1;h)/3,
\end{aligned}
\tag{A.5}
$$

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

$$
\begin{aligned}
E_1(w_1) =& \lim_{h\to 0} E\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right] \\
=& \lim_{h\to 0}\int\int f_W(w_2)f_W(w_3)\left[\left(\kappa_{XYZ}(w_1-W_2)\kappa_Y(w_1-W_3)\right.\right. \\
&\left.\left.-\kappa_{XY}(w_1-W_2)\kappa_{YZ}(w_1-W_3)\right)/v(w_1)\right]\mathrm{d}w_2\mathrm{d}w_3 \\
=& \frac{1}{v(w_1)}\int\int f_{XYZ}(w_2)f_{XYZ}(w_3)\left(\delta_{XYZ}(w_1-w_2)\delta_Y(w_1-w_3)\right. \\
&\left.-\delta_{XY}(w_1-w_2)\delta_{YZ}(w_1-w_3)\right)\mathrm{d}w_2\mathrm{d}w_3 \\
=& \frac{1}{v(w_1)}\left(f_{XYZ}(w_1)f_Y(w_1)-f_{XY}(w_1)f_{YZ}(w_1)\right),
\end{aligned}
\tag{A.6}
$$

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.6) 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,

$$\begin{aligned}
E_2(w_1,h) =& E\left[(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3))/v(w_1)\right]\\
E_2(w_1) =& \lim_{h\to 0} E\left[(\kappa_{XYZ}(W_2-w_1)\kappa_Y(W_2-W_3)-\kappa_{XY}(W_2-w_1)\kappa_{YZ}(W_2-W_3))/v(w_1)\right]\\
=& \lim_{h\to 0} E\left[(\kappa_{XYZ}(W_2-w_1)f_Y(W_2)-\kappa_{XY}(W_2-w_1)f_{YZ}(W_2))/v(w_1)\right]\\
=& \lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XYZ}(w_2-w_1)f_Y(w_2)/v(w_1)\,\mathrm{d}w_2\\
& -\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& \int f_{XYZ}(w_2)\delta_{XYZ}(w_2-w_1)f_Y(w_2)/v(w_1)\,\mathrm{d}w_2\\
& -\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)-\lim_{h\to 0}\int f_{XYZ}(w_2)\kappa_{XY}(w_2-w_1)f_{YZ}(w_2)/v(w_1)\,\mathrm{d}w_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)\\
& -\lim_{h\to 0}\int f_{XYZ}(x_2,y_2,z_2)\kappa_{XY}(x_2-x_1,y_2-y_1)f_{YZ}(y_2,z_2)/v(w_1)\,\mathrm{d}x_2\mathrm{d}y_2\mathrm{d}z_2\\
=& f_{XYZ}(w_1)f_Y(w_1)/v(w_1)\\
& -\frac{1}{v(w_1)}\int f_{XYZ}(x_2,y_2,z_2)\delta_{XY}(x_2-x_1,y_2-y_1)f_{YZ}(y_2,z_2)\,\mathrm{d}x_2\mathrm{d}y_2\mathrm{d}z_2\\
=& \frac{1}{v(w_1)}(f_{XYZ}(w_1)f_Y(w_1)-\int f_{XYZ}(x_1,y_1,z_2)f_{YZ}(y_1,z_2)\,\mathrm{d}z_2).
\end{aligned} \tag{A.7}$$

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:

$$
\begin{aligned}
E_3(w_1, h) =& E\left[(\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1))/v(w_1)\right] \\
E_3(w_1) =& \lim_{h\to 0} E\left[(\kappa_{XYZ}(W_3 - W_2)\kappa_Y(W_3 - w_1) - \kappa_{XY}(W_3 - W_2)\kappa_{YZ}(W_3 - w_1))/v(w_1)\right] \\
=& \lim_{h\to 0} \int f_{XYZ}(w_3)\kappa_Y(w_3 - w_1) f_{XYZ}(w_3)/v(w_1)\,\mathrm{d}w_3 \\
& - \lim_{h\to 0} \int f_{XY}(w_3)\kappa_{YZ}(w_3 - w_1) f_{XYZ}(w_3)/v(w_1)\,\mathrm{d}w_3 \\
=& \lim_{h\to 0} \int f_{XYZ}(x_3, y_3, z_3)\kappa_Y(y_3 - y_1) f_{XYZ}(x_3, y_3, z_3)/v(w_1)\,\mathrm{d}x_3\mathrm{d}y_3\mathrm{d}z_3 \\
& - \lim_{h\to 0} \int f_{XY}(x_3, y_3)\kappa_{YZ}(y_3 - y_1, z_3 - z_1) f_{XYZ}(x_3, y_3, z_3)/v(w_1)\,\mathrm{d}x_3\mathrm{d}y_3\mathrm{d}z_3 \\
=& \frac{1}{v(w_1)}\Big(\int f_{XYZ}(x_3, y_1, z_3) f_{XYZ}(x_3, y_1, z_3)\,\mathrm{d}x_3\mathrm{d}z_3 - \int f_{XY}(x_3, y_1) f_{XYZ}(x_3, y_1, z_1)\,\mathrm{d}x_3\Big).
\end{aligned}
\tag{A.8}
$$

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

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

upon plugging this into Eqs. (A.7) and (A.8) we obtain

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

and

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

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

A.4 Proof of Lemma 2

For a vector $w = (x, y, z)$, let

$$\eta(w) = f_{X,Y,Z}(x,y,z) f_Y(y) - f_{X,Y}(x,y) f_{Y,Z}(y,x),$$

and

$$\hat{\eta}(w) = \hat{f}_{X,Y,Z}(x,y,z) \hat{f}_Y(y) - \hat{f}_{X,Y}(x,y) \hat{f}_{Y,Z}(y,x).$$

By Lemma 1, $\hat{f}(\cdot) \xrightarrow{a.s.} f(\cdot)$ uniformly, as $n \to \infty$, but also $\hat{v}(\cdot) \xrightarrow{a.s.} v(\cdot)$ and $\hat{\eta}(\cdot) \xrightarrow{a.s.} \eta(\cdot)$ uniformly by the continuous mapping theorem.

Write $v_i = v(W_i)$ and $v_{i,n} = \hat{v}(W_i)$ for $v(W_i)$ and its estimator based on $W_1, \ldots, W_n$, and likewise

$$\eta_i = \eta(W_i)$$

and

$$\eta_{i,n} = \hat{f}_{X,Y,Z}(X_i,Y_i,Z_i) \hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i,Y_i) \hat{f}_{Y,Z}(Y_i,Z_i).$$

One may write $T_n'(h_n) = \frac{1}{n}\sum_{i=1}^n \frac{\eta_{i,n}}{v_{i,n}(h_n)}$ and $\tilde{T}_n'(h_n) = \frac{1}{n}\sum_{i=1}^n \frac{\eta_{i,n}}{v_i}$. We then find that, up to

higher-order terms in n^{-1} resulting from the factor $(n-1)/(n-2)$ in Eq. (14),

$$\sqrt{n}\left(T_n'(h_n) - \tilde{T}_n'(h_n)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\eta_{i,n}}{v_i}\left(\frac{v_i}{v_{i,n}} - 1\right).$$

The squared difference satisfies

$$n\left(T_n'(h_n) - \tilde{T}_n'(h_n)\right)^2 = n\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\eta_{i,n}}{v_i}\left(\frac{v_i}{v_{i,n}} - 1\right)\right)^2 \leq \sup_{j\in\{1,\ldots,n\}}\left(\left(\frac{v_j}{v_{j,n}} - 1\right)^2\right)\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left(\frac{\eta_{i,n}}{v_i}\right)^2 \xrightarrow{P} 0 \times E\left(\left(\frac{\eta_i}{v_i}\right)\right)^2.$$

by the uniform convergence of $\hat{v}(\cdot)$. It follows that

$$T_n'(h_n) - \tilde{T}_n'(h_n) = o_P\left(\frac{1}{\sqrt{n}}\right),$$

if $\mathrm{Var}\left(\frac{\eta_i}{v_i}\right) < \infty$.

A.5 Proof of the Asymptotic Distribution of the Modified DP Statistic

Applying Lemma 2, it remains to be proven that

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

According to the definition of $\tilde{T}_n'(h)$ in Eq. (14) 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 can 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 this point estimator. For the moment we consider the case where the vectors W_i are assumed to be independent and identically distributed. We later allow for weak dependence in the time series context as described at the end of this section.

The test statistic $\tilde{T}_n'(h)$ can be expressed by a degree three U-statistic $\tilde{K}(W_i, W_j, W_k, h)$ by symmetrization with respect to the indices i, j, k. Further, defining two kernel functions as $\tilde{K}_1(w_i) = E[\tilde{K}(w_i, W_j, W_k, h)]$ and $\tilde{K}_2(w_i, w_j, h) = E[\tilde{K}(w_i, w_j, W_k, h)]$, 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, being

$$\begin{aligned}\tilde{K}_1(w_i,h)-\lim_{h\to 0}\tilde{K}_1(w_i,h)&=s(w_i)h^{\alpha}+s^*(w_i,h),\ \alpha>0,\\ E[(\tilde{K}_2(W_i,W_j,h))^2]&=q_2h^{-\gamma}+q_2^*(h),\ \gamma>0,\\ E[(\tilde{K}(W_i,W_j,W_k,h))^2]&=q_3h^{-\delta}+q_3^*(h),\ \delta>0,\end{aligned}\tag{A.9}$$

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. The conditions in Eq. (A.9) are satisfied if α is set as the order of kernel function $\mathbb{K}(.)$, which is 2 for the Gaussian kernel, and γ, δ depend on the dimensions of the variables under consideration via $\gamma=d_X+d_Y+d_Z$ and $\delta=d_X+2d_Y+d_Z$. Define $C_0=2\text{Cov}\left(\lim_{h\to 0}\tilde{K}_1(W_i,h),s(W_i)\right)$, we can show that the mean squared error of DP statistic as a function of sample size dependent bandwidth is given by

$$\text{MSE}[T_n(h)]=\left(E[s(W_i)]\right)^2h^{2\alpha}+\frac{9}{n}C_0h^{\alpha}+\frac{9}{n}\text{Var}\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right]+\frac{18}{n^2}q_2h^{-\gamma}+\frac{6}{n^3}q_3h^{-\delta}+\text{h.o.t.}\tag{A.10}$$

The scaling factor $(1/v(\cdot))$ in the modified test statistic $\tilde{T}_n'(h)$ enters the MSE in Eq. (A.10) by mainly changing the bandwidth-independent variance term. For the other bandwidth-dependent terms, $(1/v(\cdot))$ just re-scales the coefficients without affecting the convergence rates. Thus, as in Diks and Panchenko (2006) we may still allow for all the h-dependent terms to be $o_P(n^{-1})$ to ensure that $\frac{9}{n}\text{Var}\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right]$-term asymptotically dominates (in which case asymptotic normality of the test statistic obtains). Therefore, adopting a sample size-dependent bandwidth $h_n=Cn^{-\beta}$, with C, $\beta>0$, one finds

$$\sqrt{n}\frac{\tilde{T}_n'(h)-t}{S_n}\xrightarrow{d}N(0,1)\quad\text{if}\quad\frac{1}{2\alpha}<\beta<\frac{1}{d_X+d_Y+d_Z},\tag{A.11}$$

where S_n^2 is a consistent estimator of the asymptotic variance $9\text{Var}\left[\lim_{h\to 0}\tilde{K}_1(W_i,h)\right]$. The bivariate case, for $\alpha=2$ and $d_X+d_Y+d_Z=3$, requires $\beta\in(1/4,1/3)$. In the time series setting, under the assumption that the processes are stationary and weakly dependent, the long-run variance of $\sqrt{n}(T_n'(h)-t)$ is given by $\sigma^2=9\left(\text{Var}\left(U_t\right)+2\sum_{\ell=1}^{\infty}\text{Cov}\left[U_t,U_{t+\ell}\right]\right)$, where $U_t=\lim_{h\to 0}(\tilde{K}_1(W_t,h))$. The variance σ^2 can then be estimated using a HAC estimator for

the long-run variance of U_t (Denker and Keller 1983; 1986).

A.6 Optimal Bandwidth for the Modified DP Test

The optimal bandwidth should balance the squared bias and variance of the test statistic, given in Eq. (A.10). 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 the 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),

$$\begin{aligned} \text{Bias}(\hat{f}(x)) &= \frac{\mu_\nu(\kappa)}{\nu!} \sum_{j=1}^{k} \frac{\partial^\nu}{\partial x_j^\nu} f(x) h_j^\nu + o_P(h_1^\nu + \ldots + h_k^\nu), \\ \text{Var}(\hat{f}(x)) &= \frac{f(x) R(\kappa)^k}{(n-1) h_1 h_2 \ldots h_k} + o_P((n-1) h_1 h_2 \ldots h_k), \end{aligned} \tag{A.12}$$

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 $R(\kappa) = \int_{-\infty}^{\infty} \kappa(t)^2 dt$ 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 $\mathbf{H} = (h_1, \ldots, h_k)'$. It is not difficult to see that $\text{E}[s(W)]$ and q_2 defined in Eq. (A.9) depend on the kernel function used trough the functions $\mu_\nu(\kappa)$ and $R(\kappa)$.

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

$$\begin{aligned} \mu_\nu^{SQ}(\kappa) &= 1/3, \quad R^{SQ}(\kappa) = 1/2, \\ \mu_\nu^{G}(\kappa) &= 1, \quad R^{G}(\kappa) = 1/2\sqrt{\pi}. \end{aligned} \tag{A.13}$$

In our research, when we substitute the square kernel by the Gaussian kernel, the squared bias-related $\text{E}[s(W)]$ and the variance-related q_2 will change correspondingly. Directly applying Eq. (A.13), we have $q^G(\kappa)_2 = 3q^{SQ}(\kappa)_2$, $R^G(\kappa) = R^{SQ}(\kappa)/\sqrt{\pi}$. After plugging this into Eq. (15)

and performing some calculations, one finds

$$h^* \approx 0.6 h_{DP}. \tag{A.14}$$