JOURNAL OF APPLIED ECONOMETRICS
*J. Appl. Econ.* **31**: 1333–1351 (2016)
Published online 27 November 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2495

# NONLINEAR GRANGER CAUSALITY: GUIDELINES FOR MULTIVARIATE ANALYSIS

CEES DIKS[a,b] AND MARCIN WOLSKI[c*]
[a] *Center for Nonlinear Dynamics in Economics and Finance (CeNDEF), University of Amsterdam, The Netherlands*
[b] *Tinbergen Institute, Amsterdam, The Netherlands*
[c] *European Investment Bank, Luxembourg, Luxembourg*

SUMMARY

We propose an extension of the bivariate nonparametric Diks–Panchenko Granger non-causality test to multivariate settings. We first show that the asymptotic theory for the bivariate test fails to apply to the multivariate case, because the kernel density estimator bias and variance cannot both tend to zero at a sufficiently fast rate. To overcome this difficulty we propose to reduce the order of the bias by applying data sharpening prior to calculating the test statistic. We derive the asymptotic properties of the 'sharpened' test statistic and investigate its performance numerically. We conclude with an empirical application to the US grain market, using the price of futures on heating degree days as an additional conditioning variable. Copyright © 2015 John Wiley & Sons, Ltd.

*Received 18 December 2013; Revised 21 September 2015*

*Supporting information may be found in the online version of this article.*

## 1. INTRODUCTION

Since the introduction of Granger causality over four decades ago (Granger, 1969), the body of literature on this topic has grown substantially, and Granger causality has become a standard methodological concept not only among economists and econometricians but also physicists, biologists and others (Guo *et al.*, 2010). Granger causality has induced an ongoing discussion on the nature and validity of the concept, in which its methodological limitations have been pointed out repeatedly. Although the spectrum of arguments against the notion of Granger causality is broad, the main line of criticism concerns the oversimplified nature of the dependence relations in the economy that parametric Granger causality tests often rely on (Cartwright, 2007). The scope of this paper is to alleviate this limitation by allowing for a more flexible nonparametric setting for Granger causality testing in multivariate contexts.

To the best of our knowledge, the only fully nonparametric test that has proven correct asymptotic size properties under the null hypothesis that there is no Granger causality (to be defined below) from a given variable, $X$, say, to $Y$, is the test proposed by Diks and Panchenko (2006; hereafter DP) and, even then, this is only the case in the narrow setting of a bivariate time series process $\{(X_t, Y_t)\}$. This limitation prevents the test from taking into account possible confounding effects of other variables fully nonparametrically. The only way in which additional variables can currently be controlled for is by pre-filtering the multivariate data using a parametric model, such as a vector autoregressive (VAR) model or a multivariate (generalized) autoregressive conditional heteroskedasticity ((G)ARCH) model, and analyzing the residuals of the fitted model pair-wise. The main aim of this paper is to open up the possibility to condition on additional variables in a fully nonparametric way.

Consider a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $t \in \mathbb{Z}$. We say that $\{X_t\}$ is a (general) Granger cause of $\{Y_t\}$ (or simply $X$ is a Granger cause of $Y$) if past and current values of $X$ contain

* Correspondence to: Marcin Wolski, European Investment Bank, Luxembourg, Luxembourg. E-mail: m.wolski@eib.org

additional information on future values of $Y$ that is not contained in past and current $Y$-values alone. If we denote the information contained in past observations, $X_s$ and $Y_s$ for $s \leq t$, by $\mathcal{F}_{X,t}$ and $\mathcal{F}_{Y,t}$, respectively, and let '$\sim$' denote equivalence in distribution, the formal definition is as follows.

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

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

The absence of Granger causality, i.e. $(Y_{t+1}, \ldots, Y_{t+k})|(\mathcal{F}_{X,t}, \mathcal{F}_{Y,t}) \sim (Y_{t+1}, \ldots, Y_{t+k})|\mathcal{F}_{Y,t}$ for all $k \geq 1$ is referred to as Granger non-causality.

The notion of Granger causality given in Definition 1 is very general, not only depending on the conditional mean of future $Y$-values given current and past $X$- and $Y$-values, but also on their entire conditional distribution. An advantage of such a general notion of Granger causality is that it does not depend on any parametric assumption regarding the data-generating process, besides it being a strictly stationary bivariate time series process. In testing the hypothesis of the absence of this form of Granger causality, it is clear that one is not restricted to the class of linear VAR models usually considered in the literature. In fact, the fully nonparametric tests for Granger non-causality considered here turn out to be consistent against a wide range of types of alternatives, including nonlinear Granger causality in mean (as shown analytically below) and variance (shown numerically).

Hiemstra and Jones (1994) were the first to propose a nonparametric test for bivariate Granger non-causality, which was later extended to the multivariate case by Bai *et al.* (2010). However, Diks and Panchenko (2005) showed that the Hiemstra–Jones test, strictly speaking, does not test the null hypothesis of Granger non-causality or even an implication of it, and, as a result, the quantity on which the test statistic is based, and which Hiemstra and Jones presumed to be zero under the null hypothesis, may in fact be nonzero under the null hypothesis. In a setting that could be representative for financial and economic applications (bivariate ARCH) this was shown to lead to severe spurious rejection rates under the null hypothesis, that even tend to one asymptotically as the sample size increases.

To overcome this problem, DP proposed a new bivariate test statistic that *does* test an implication of the null hypothesis of Granger non-causality. However, as shown below, in higher-variate settings there exists no sequence of bandwidth values as a function of the sample size for which both the bias and variance of the test statistic converge to zero fast enough, and the standardized test statistic is no longer asymptotically standard normal. As a consequence, the standard DP test statistic, like the HJ test, also suffers from size distortions, albeit only in multivariate settings and potentially less severe asymptotically. The main aim of this paper is to address this problem by generalizing the bivariate DP test to an appropriate multivariate counterpart.

Definition 1 of Granger non-causality for bivariate time series is too restrictive to apply meaningfully in a more complex setting where there are more than two interacting variables. For instance, the bivariate definition does not take into account the possible effects of a third 'confounding' variable, $\{Q_t\}$, say, which may lead to the detection of spurious Granger causal relations. Ideally, one would like to control for *every possible source of variation of every kind* ..., as noted by Cartwright (2007). However, it is clear that this is fundamentally infeasible without further assumptions, since there may always be additional unobserved variables that should be controlled for. We therefore focus on the more modest goal of controlling for an observed, possibly multivariate, additional variable $Q_t$ using the following generalization of Definition 1.

**Definition 2. Granger causality (multivariate)** For a strictly stationary multivariate time series process $\{(X_t, Y_t, Q_t)\}$, $t \in \mathbb{Z}$, where $X_t$ and $Y_t$ are univariate and $Q_t$ is univariate or multivariate, $\{X_t\}$

is a Granger cause of $\{Y_t\}$ if, for some $k \geq 1$,

$$(Y_{t+1}, \ldots, Y_{t+k})|(\mathcal{F}_{X,t}, \mathcal{F}_{Y,t}, \mathcal{F}_{Q_t}) \nsim (Y_{t+1}, \ldots, Y_{t+k})|\mathcal{F}_{Y,t}\mathcal{F}_{Q,t}$$

We say that there is Granger non-causality (from $\{X_t\}$ to $\{Y_t\}$) if the converse holds, that is, if for all $k \geq 1$, $(Y_{t+1}, \ldots, Y_{t+k})|(\mathcal{F}_{X,t}, \mathcal{F}_{Y,t}, \mathcal{F}_{Q,t}) \sim (Y_{t+1}, \ldots, Y_{t+k})|\mathcal{F}_{Y,t}\mathcal{F}_{Q,t}$. Although we may allow for multivariate $\{X_t\}$ and $\{Y_t\}$, for clarity of exposition we will assume both $\{X_t\}$ and $\{Y_t\}$ to be univariate. This corresponds to the practically most relevant case where one is interested in determining which of the individual variables drive which of the other variables.

We extend the bivariate DP test to the multivariate setting by reducing the kernel estimator bias using the data-sharpening (DS) method proposed by Hall and Minnotte (2002). The performance of this approach is then investigated both numerically and empirically on the US grain market. We chose this specific market due to the clear presence of a potential confounding variable, since it can be expected that the price of each grain is not only influenced by the prices of other grains through market mechanisms, but that the whole market to a large extent is driven by weather expectations. Therefore, this market serves as a natural context to apply the multivariate test in practice.

This paper is organized as follows. Section 2 discusses the asymptotic properties of the DP test and explains in detail why the standardized test statistic fails to be asymptotically normal in multivariate settings. In Section 3 the standard kernel density estimator is replaced by its sharpened form and it is shown that the new test statistic *is* asymptotically normally distributed. To assess the accuracy of the asymptotic approximation as a function of the sample size, and to obtain an impression of the finite sample power of the test, we complement our theoretical results by computer simulations. In Section 4 we apply the new test to the US grain market. Section 5 summarizes and concludes.

## 2. ASYMPTOTIC PROPERTIES OF THE BIVARIATE DP TEST

Consider a strictly stationary bivariate time series process $(X_t, Y_t)$, $t \in \mathbb{Z}$. Given an observed time series $\{(X_t, Y_t)\}_{t=1}^n$ generated by this process, a test for Granger non-causality, from $X$ to $Y$, say, aims to find statistical evidence against the null hypothesis

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

with Granger causality defined according to Definition 1. We limit ourselves to $k = 1$ (one-step-ahead effects of $X$ on $Y$), which is the case considered most often in practice. Under the null hypothesis $Y_{t+1}$ is conditionally independent of $X_t, X_{t-1}, \ldots$, given $Y_t, Y_{t-1}, \ldots$. In a nonparametric setting, conditioning on the infinite past is impossible without some model restriction, such as the assumption that the process is of finite Markov order. Under this implicit assumption, in practice conditional independence is usually tested up to a finite number of lags $l_X$ and $l_Y$, i.e.

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

where $X_t^{l_X} = (X_{t-l_X+1}, \ldots, X_t)$ and $Y_t^{l_Y} = (Y_{t-l_Y+1}, \ldots, Y_t)$. For a strictly stationary bivariate time series $\{(X_t, Y_t)\}$ this is a statement about the invariant distribution of the $l_X + l_Y + 1$-dimensional vector $W_t = (X_t^{l_X}, Y_t^{l_Y}, Z_t)$, where $Z_t = Y_{t+1}$. To keep the notation simple, and to bring about the fact that the null hypothesis is a statement about the invariant distribution of $W_t$, we often drop the time index and introduce $W = (X, Y, Z)$, where the latter is a random vector with the invariant distribution of $(X_t^{l_X}, Y_t^{l_Y}, Y_{t+1})$.

Consider the simplest setting, where $l_X = l_Y = 1$, so that $W = (X, Y, Z)$ denotes a three-variate random variable, distributed as $W_t = (X_t, Y_t, Y_{t+1})$. Throughout we assume that $W$, which in the



DOI: 10.1002/jae

bivariate setting is three-dimensional, is a continuous random variable. The higher-dimensional case will be discussed below.

For $k = l_X = l_Y = 1$, the null hypothesis can be phrased as

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

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

$$H_0: \quad \frac{f_{X,Y,Z}(x,y,z)}{f_Y(y)} - \frac{f_{X,Y}(x,y)}{f_Y(y)}\frac{f_{Y,Z}(y,z)}{f_Y(y)} = 0 \tag{2}$$

for $(x, y, z)$ in the support of $W$. This implies that for any weight function $g(X, Y, Z)$

$$E\left(\left[\frac{f_{X,Y,Z}(X,Y,Z)}{f_Y(Y)} - \frac{f_{X,Y}(X,Y)}{f_Y(Y)}\frac{f_{Y,Z}(Y,Z)}{f_Y(Y)}\right]g(X,Y,Z)\right) = 0 \tag{3}$$

whenever the expectation exists. This may be viewed as an infinite number of moment restrictions, indexed by $g(x, y, z)$.

When designing a nonparametric test for a hypothesis of this type, one possible approach is to base the test on a quantity that is sensitive to violation of any of these moment conditions. In the case that this succeeds it will provide an omnibus test with power against any alternative. Although it is usually not hard to find quantities with this property, such as the integrated square of the left-hand-side of equation (2), not many of these, if any, will admit analytically tractable asymptotic theory or even provably valid bootstrap critical values. And even if asymptotic theory is available, the tests obtained in this way need not have optimal power against specific alternatives which the researcher is primarily interested in. In practice, therefore, one often decides to test not all but just a single, or a few, moment restrictions, arrived at by choosing one or more specific functions $g(x, y, z)$ based on practical considerations. For instance, when developing a test for conditional independence via empirical likelihood, Su and White (2014) express the null hypothesis in terms of an infinite number of moment restrictions, but eventually focus on testing a single, data-weighted average of those conditions.

Likewise, based on size/power simulation results, DP decided to focus on the single weight function $g(x, y, z) = f_Y^2(y)$. This results in testing the implication of $H_0$, as stated in equation (2), that

$$H_0': \qquad q \equiv E\left[f_{X,Y,Z}(X,Y,Z)f_Y(Y) - f_{X,Y}(X,Y)f_{Y,Z}(Y,Z)\right] = 0 \tag{4}$$

Although $q$ as an operator on $f_{X,Y,Z}(x,y,z)f_Y(y) - f_{X,Y}(x,y)f_{Y,Z}(y,z)$ is not positive definite, it turns out that for empirical time series rejections in the vast majority of cases are caused by an estimate of $q$ being significantly larger than zero, while significantly negative estimates of $q$ are rarely observed. This is confirmed by numerical simulations under various alternatives of interest. This is why a test based on $q$ is often implemented as a one-sided test of $H_0'$ against the alternative $H_a'$: $q > 0$, rejecting when an estimate of $q$ is significantly larger than zero.

Although testing the implication $H_0'$ instead of $H_0$ will lead to loss of power against certain specific alternatives, there are several arguments motivating a test based on $q$. Firstly, Skaug and Tjøstheim (1993), in the closely related context of testing for independence rather than conditional independence, found that although their analogue of $q$, $E\left[f_{X,Z}(X,Z) - f_X(X)f_Z(Z)\right]$, is not positive definite, their test based on it consistently outperformed other, positive definite, functionals they considered in their study. Their heuristic argument for this is that $\int\int f_{X,Z}(x,z)(f_{X,Z}(x,z) - f_X(x)f_Z(z))\,\mathrm{d}x\,\mathrm{d}z$ can be considered a weighted average of $f_{X,Z}(x,z) - f_X(x)f_Z(z)$ for which the weight $f_{X,Z}(x,z)$ is large if $f_{X,Z}(x,z) > f_X(x)f_Z(z)$ and small if $f_{X,Z}(x,z) < f_X(x)f_Z(z)$. They actually prove that $E\left[f_{X,Z}(X,Z) - f_X(X)f_Z(Z)\right] \geq 0$ for the bivariate normal case, with equality if and only if $f_{X,Z}(x,z) = f_X(x)f_Z(z)$. Conversely, they offer a (discrete) counter-example, where $f_{X,Z}(x,z) \not\equiv f_X(x)f_Z(z)$, and nevertheless $E\left[f_{X,Z}(X,Z) - f_X(X)f_Z(Z)\right]$ is negative. Secondly, purely positive definite functionals, such as the integrated square or absolute value of the left-hand-side of equation (2), do typically not offer analytic handles such as a corresponding U-statistic representation, which enable us to trace the asymptotic distribution of the test statistic.

Testing an implication of the null hypothesis rather than the null hypothesis itself is common practice and this need not be problematic as long as the test has nontrivial power against a sufficiently wide variety of alternatives, and in particular against alternatives that are of practical interest. Well-known linear Granger causality tests based on testing a restriction in a parametric VAR model can be considered an example, as these are tailored to detecting linear Granger causality in the conditional mean, and as such are consistent against some but not all other types of Granger causality. Linear Granger causality tests, for instance, do not have unit asymptotic power against Granger causality in second conditional moments.

### 2.1. Consistency within Certain Important Model Classes

Besides heuristic and numerical evidence that $q$ is a suitable basis for a test for Granger non-causality, we at least also wish to analytically identify several alternatives against which such a test will have unit asymptotic power. This can be done by focusing on specific classes of processes and/or distributions that give rise to strictly positive values of $q$ if $H_0$ does not hold. Within such classes, a test based on a consistent estimator of $q$ will have unit asymptotic power. Three such classes are covered by the following theorems, the proofs of which are provided in the online Appendix to this paper, available as supporting information.

The first result concerns Granger causality through dependence in the (conditional) mean.

**Theorem 1. Granger causality in mean** Suppose we have a strictly stationary process for which $Z_t = Y_{t+1}$ given $(x_t, y_t)$ can be represented as $Z_t = g(x_t, y_t) + \varepsilon_{t+1}$, where $\varepsilon_{t+1}$ is a zero-mean noise term which is independent of $(X_t, Y_t)$, and $g(x,y)$ the regression function (conditional mean) of $Z_t$ given $(x,y)$. Then $q \geq 0$ with equality if and only if $X_t$ and $Z_t$ are conditionally independent given $Y_t$.

As argued in the proof in the online Appendix, under the assumptions of this theorem Granger causality necessarily is equivalent to (strict) Granger causality in mean. Here 'strict' should be understood in the sense that $X_t$ affects the conditional mean of $E(Y_{t+1}|x_t, y_t)$ by shifting the whole conditional distribution of $Y_{t+1}$ given $(x_t, y_t)$ up or down, i.e. without affecting the conditional distribution of the innovation $\varepsilon_{t+1} = Y_{t+1} - E(Y_{t+1}|x_t, y_t)$ at the same time. We therefore refer to the result stated in Theorem 1 by saying, with slight abuse of language, that the DP test has unit asymptotic power against Granger causality in mean.

Note that the class of alternatives with Granger causality in mean is much larger than the class of alternatives with only linear Granger causality in mean, which de facto is the class of alternatives on which nearly all research focuses, being mainly based on classical linear VAR models.

To provide an example of a data-generating process with Granger causality in mean, consider the (possibly nonlinear) general bivariate VAR(1) model given by

$$
\begin{aligned}
X_{t+1} &= g_X(X_t, Y_t) + \varepsilon_{X,t+1} \\
Y_{t+1} &= g_Y(X_t, Y_t) + \varepsilon_{Y,t+1}
\end{aligned}
\tag{5}
$$

where $\varepsilon_{X,t+1}$ and $\varepsilon_{Y,t+1}$ represent possibly correlated zero-mean innovations and $\{X_t\}$ is Granger causing $\{Y_t\}$ in mean if $g_Y(x,y)$ depends on $x$. The term 'general' refers to the fact that the conditional means $g.(x,y)$ are not restricted to be linear in $x$ and $y$, as would be the case in a standard



DOI: 10.1002/jae

VAR(1) model. Being able to detect general (i.e. linear and/or nonlinear) Granger causality in mean is important, first because it is a natural generalization of standard linear Granger causality in mean, and secondly because it can provide crucial information on which variables to include in a nonlinear or nonparametric model for the conditional mean, bearing direct relevance to model identification and forecasting.

A second class for which the DP test is consistent is that for which $X$ and $Z$ are both uniformly distributed given $Y$, as stated in the following theorem, which is phrased in terms of the distribution of $W = (X, Y, Z)$, i.e. without explicit reference to the time series structure of $W_t$.

**Theorem 2. Uniform conditional marginals** Suppose that the density of $W = (X, Y, Z)$ is a member of the class of densities for which $X$ and $Z$ each have a uniform conditional distribution given $Y$. Then $q \geq 0$ with $q = 0$ if and only if $X$ and $Z$ are conditionally independent given $Y = y$.

A third class of densities for which the DP test is consistent is that for which the conditional distribution of $X$ and $Z$ given $Y$ is bivariate normal.

**Theorem 3. Conditional bivariate normal distribution of $(X, Z)$ given $Y$** Suppose that the multivariate density $W = (X, Y, Z)$ is a member of the class of densities for which $(X, Z)$ has a bivariate normal conditional distribution given $Y$. Then $q \geq 0$ with $q = 0$ if and only if $X$ and $Z$ are conditionally independent given $Y = y$.

Theorem 3 includes the important class of bivariate normal distributions as a special case. It therefore also includes all linear VAR(1) processes with bivariate normal innovations.

Besides the special cases summarized in the above theorems, numerical simulations show that there are many other types of Granger causality against which the DP test has non-trivial power, including $X$ affecting the variance of future $Y$-variables (Granger causality through the conditional second moment), referred to as (G)ARCH effects or volatility spillover. One can, of course, test for (G)ARCH using parametric models such as VAR models with innovations following a multivariate (G)ARCH process, but in such a parametric approach misspecification of, for instance, the conditional mean as a linear function while it is actually nonlinear may lead to spurious detection of (G)ARCH spillover effects. This is avoided in a nonparametric approach without specific model assumptions.

## 2.2. Asymptotic Normality

Given local kernel density estimators around each of the observations $W_i$, $i = 1, \ldots, n$,

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

based on a density estimation kernel $K$ and bandwidth $\varepsilon$, where $d_W$ is the dimension of $W$, a natural estimator of $q$, is given by

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

where the normalization factor $(n-1)/(n(n-2))$ is inherited from the $U$-statistic representation of $T_n(\varepsilon)$. Since we wish to test whether $q$ is significantly larger than zero, we take $T_n(\varepsilon)$ as the basic (non-standardized) test statistic and develop asymptotic theory for $T_n(\varepsilon)$ along the lines of DP. In the bivariate case it turns out that for appropriate choices of $\varepsilon = \varepsilon_n$, $T_n(\varepsilon_n)$ is asymptotically normally distributed, and can be standardized to a $z$-score under $H_0$.

The asymptotic behavior of $T_n(\varepsilon)$ can be obtained from considerations originally developed for mean squared error (MSE) optimal bandwidth selection under shrinking bandwidth conditions (Powell and Stoker, 1996). The test statistic $T_n(\varepsilon)$ has a corresponding third order $U$-statistic representation with a $U$-statistic kernel given by $\tilde{K}(W_i, W_j, W_k)$, which is symmetrized with respect to interchanging any two of the three arguments (see the online Appendix). Let us furthermore denote $\tilde{K}_2(w_1, w_2) = E[\tilde{K}(w_1, w_2, W_3)]$ and $\tilde{K}_1(w_1) = E[\tilde{K}(w_1, W_2, W_3)]$. Under mild regularity conditions on the joint density of $W$, the rates of convergence of the point-wise bias as well as the second moment kernel expansions depend on the bandwidth as

$$\tilde{K}_1(w_i, \varepsilon) - \lim_{\varepsilon \to 0} \tilde{K}_1(w_i, \varepsilon) = s(w_i)\varepsilon^{\alpha} + s^*(w_i, \varepsilon), \quad \alpha > 0 \tag{8}$$

$$E\left[(\tilde{K}_2(W_1, W_2))^2\right] = q_2 \varepsilon^{-\gamma} + q_2^*(\varepsilon), \quad \gamma > 0 \tag{9}$$

$$E\left[(\tilde{K}(W_1, W_2, W_3))^2\right] = q_3 \varepsilon^{-\delta} + q_3^*(\varepsilon), \quad \delta > 0 \tag{10}$$

where the remainder terms are negligible, i.e. $E\|s^*(W_i, \varepsilon)\|^2 = \mathrm{o}(\varepsilon^{2\alpha})$, $q_2^*(\varepsilon) = \mathrm{o}(\varepsilon^{-\gamma})$ and $q_3^*(\varepsilon) = \mathrm{o}(\varepsilon^{-\delta})$. It turns out that $\alpha$ is equal to the (kernel) order of the density estimation kernel function $K$ (and hence the local kernel estimation bias is $O(\varepsilon^{\alpha})$) and $\gamma$ and $\delta$ depend on the dimensions of the variables under consideration as $\gamma = d_X + d_Y + d_Z = d_W$ and $\delta = d_X + 2d_Y + d_Z$ (DP).

DP show that if $\varepsilon$ tends to zero as $n$ increases, the MSE of the test statistic can be expressed as

$$\mathrm{MSE}[T_n(\varepsilon)] = (E[s(W_i)])^2 \varepsilon^{2\alpha} + \frac{9}{n} C_0 \varepsilon^{\alpha} + \frac{9}{n} \mathrm{var}\left[\lim_{\varepsilon \to 0} \tilde{K}_1(W_i, \varepsilon)\right] + \frac{18}{n^2} q_2 \varepsilon^{-\gamma} + \frac{6}{n^3} q_3 \varepsilon^{-\delta} + \text{h.o.t.} \tag{11}$$

where $C_0 = 2\mathrm{cov}\left[\lim_{\varepsilon \to 0} \tilde{K}_1(W_i, \varepsilon), s(W_i)\right]$ and 'h.o.t.' stands for higher-order terms, i.e. terms converging to zero faster than those explicitly given on the right-hand side.

Asymptotic normality of $T_n(\varepsilon)$ is obtained if and only if the MSE is asymptotically dominated by the third term on the right-hand side of equation (11), which is the usual expression for the variance (and hence MSE) of a $U$-statistic with an $\varepsilon$-independent $U$-statistic kernel. This is the case if all other terms are of smaller order, i.e. $o(n^{-1})$. The second term is always $o(n^{-1})$, and the fifth is dominated by the fourth, so that only the first and the fourth terms need to be kept under control asymptotically, leading to two conditions on the possible rates at which $\varepsilon$ is allowed to tend to zero as $n$ increases. The first term is associated with the bias of $\tilde{K}_1(W_i, \varepsilon)$ as a local estimator of $\lim_{\varepsilon \to 0} \tilde{K}_1(W_i, \varepsilon))$ at $W_i = w_i$, and the fourth with its variance (averaged across $W$).

Adopting a bandwidth that tends to zero with $n$ as $\varepsilon = \varepsilon_n = Cn^{-\beta}$, $C, \beta > 0$, one finds (Theorem 1 of DP)

$$\sqrt{n}\frac{T_n(\varepsilon_n) - q}{S_n} \xrightarrow{d} \mathrm{N}(0, 1) \quad \text{iff} \quad \frac{1}{2\alpha} < \beta < \frac{1}{d_X + d_Y + d_Z} \tag{12}$$

where $S_n^2$ is a consistent estimator of the asymptotic variance $\sigma^2 = 9\mathrm{var}[\lim_{\varepsilon \to 0} \tilde{K}_1(W_i, \varepsilon))]$.

For a standard second-order kernel ($\alpha = 2$), in the baseline bivariate case with $d_X = d_Y = d_Z = 1$, the test statistic is asymptotically normally distributed for any positive constant $C$ and $\beta \in (1/4, 1/3)$. Additionally, if there is serial dependence between the vectors $W_i$, under suitable mixing conditions (see Denker and Keller, 1983) the asymptotic normality result can be extended to weakly dependent time series by taking into account covariances between the contributions being summed over in the test statistic in its $U$-statistic representation (see the online Appendix). Since, as argued in Section 2.1,

$q$ is typically larger than zero under $H_a$, tests based on $q$ are usually implemented as one-sided tests, rejecting when $q$ is statistically significantly larger than 0. That is, based on asymptotic critical values, for a given significance level $\alpha$ the null hypothesis is rejected if $\sqrt{n}(T_n(\varepsilon_n) - q)/S_n \geq z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $(1-\alpha)$th quantile of the standard normal distribution.

### 2.3. The Dimensionality Problem

Let us now consider what happens if we increase the dimensionality. For clarity, imagine that we wish to test for Granger non-causality from $\{X_t\}$ to $\{Y_t\}$ taking into account observations from an additional time series[1] $\{Q_t\}$, so that the analogue of equation (1) becomes

$$Y_{t+1}|\left(X_t^{l_X}, Y_t^{l_Y}, Q_t^{l_Q}\right) \sim Y_{t+1}|\left(Y_t^{l_Y}, Q_t^{l_Q}\right)$$

Let us again consider the case where all lags are equal to one: $l_X = l_Y = l_Q = 1$. From the considerations in the previous section, it follows that asymptotic normality can only be obtained for $\beta$ in the range between $1/(2\alpha)$ and $1/\left(d_X + d_Y + d_Z + d_Q\right) = 1/d_W$. For a standard second-order $(\alpha = 2)$ density estimation kernel, one observes that if the dimensionality $d_W$ of $W$ is increased from 3 to any number larger than or equal to 4, there is no longer a feasible interval of $\beta$-values that render $T_n(\varepsilon)$ asymptotically normal.

The problem results from the decrease in the maximum rate $1/d_W$ at which we can let the bandwidth go to zero while letting the variance term (the fourth term on the right-hand side of equation (11)) decrease sufficiently fast, in higher dimensions. Making the point-wise kernel estimator bias $E[s(W_i)]$, where now $W_i = (X_i, \boldsymbol{Y}_i \equiv (Y_i, Q_i), Z_i)$ (first term on the right-hand side of equation (11)) vanish sufficiently fast asymptotically requires the bandwidth to go to zero faster than $1/(2\alpha) = 1/4$, independent of the dimension of $W$. By increasing the dimension, the bias term is dominating the MSE when we let the bandwidth tend to zero sufficiently slowly to ensure that the variance converges to zero fast enough, or, conversely, the variance is not converging to zero sufficiently fast if we let the bandwidth tend to zero fast enough for the bias to disappear asymptotically.

This is related to the so-called *curse of dimensionality*. As suggested by Scott (1992), in statistics the problem is a consequence of sparsity of data in larger dimensions. Imagine, for instance, a uniform sample over the $[-1, 1]^d$ hypercube, where $d$ is the total number of dimensions. Given an arbitrarily small region of radius $\mu < 1$, as $d \to \infty$ the number of points within $[-\mu, \mu]^d$ tends to 0 fast. This suggests that in higher-dimensional spaces the smoothing parameter should be larger in order to capture a comparable number of points. However, by increasing the bandwidth we increase the bias of the estimator, violating the consistency of the test statistic in this case.

There are several methods by which this dimensionality problem can be addressed. Scott (1992) suggests using principal components, projection pursuit or informative components analysis. These solutions, however, require additional assumptions on the underlying structure of the data. For instance, they might be of great advantage when dealing with 100-dimensional spaces where one could assume that the data structure falls into a 20-dimensional manifold. The minimum number of components may even be smaller, but may often still be larger than 3, so that the dimension reduction might still be insufficient to provide consistency of the test.

Another solution can be obtained by improving the precision of the density estimator by reducing the estimator bias. Since this does not require any particular underlying structure in the data, this is a natural choice in a nonparametric multivariate setting.

[1] In fact the additional variable $\{Q_t\}$ need not necessarily be an extra observed time series. It may also correspond to an additional lagged variable, such as $Q_t = Y_{t-2}$.

## 3. DATA SHARPENING AS A BIAS REDUCTION METHOD

The intuition behind data sharpening (DS) is to slightly perturb the original dataset by applying a sharpening map $\psi_p(\cdot)$ in order to reduce the bias of the estimator (here $p$ is the order of bias reduction, defined such that after bias reduction the bias is of order $O(\varepsilon^p)$). The idea of the perturbation is to concentrate points more where they are already dense and thin them out more where they are already sparse. The explicit form of $\psi_p(\cdot)$ depends on the order of the bias reduction one would like to obtain.

There are several advantages of DS over standard bias reduction techniques. First, it allows for arbitrarily high orders of bias reduction (Hall and Minnotte, 2002). Since testing for Granger causality is widely recognized for its practical purposes, the flexibility of this method is a great advantage. Secondly, as we confirm in our study, it does not affect the kernel function directly, but only reduces the contribution of the bias in expansion (11) of the MSE of the test statistic. Thirdly, it is easy and straightforward to implement, even in a general nonparametric multivariate setting.

The sharpening map $\psi_p(\cdot)$ is data driven, and involves a nonparametric kernel-based estimator of the local derivative (gradient) of the density, depending on a DS bandwidth, $\varepsilon_{\text{DS}}$, which should be chosen by the user. For consistency of the gradient estimator the data-sharpening bandwidth, $\varepsilon_{\text{DS}}$, should tend to zero strictly slower than $n^{-1/(\gamma+2)}$, where $\gamma = \dim W = d_W$ (Fukunaga and Hostetler, 1975). Since this condition can be imposed by the user, throughout we assume that it is satisfied.

The sharpened form of the plug-in density estimator given in equation (6) is

$$\hat{f}^s_W(W_i) = ((n-1)\varepsilon)^{-d_W} \sum_{j,j\neq i} K\left(\frac{W_i - \psi_p(W_j)}{\varepsilon}\right) \tag{13}$$

The sharpened form of the test statistic, $T_n^s(\varepsilon)$, is obtained by plugging the sharpened density estimators into the test statistic (7), i.e.

$$T_n^s(\varepsilon) = \frac{(n-1)}{n(n-2)} \sum_i \left(\hat{f}^s_{X,Y,Z}(X_i, Y_i, Z_i)\hat{f}^s_Y(Y_i) - \hat{f}^s_{X,Y}(X_i, Y_i)\hat{f}^s_{Y,Z}(Y_i, Z_i)\right) \tag{14}$$

As shown in the online Appendix, this reduces the point-wise bias to order $o(\varepsilon^p)$, while the other properties of the kernel $\tilde{K}$ remain the same. For $T_n^s(\varepsilon)$ the local bias is of the form given in equation (8) with $\alpha = p$, while the parameters $\gamma$ and $\delta$ from equations (9) and (10) remain unchanged. (The practical implementation of the DS bias reduction to order $p = 4$ in the 3-variate setting is described in the online Appendix.) This reasoning is summarized in the following theorem, which extends the asymptotic normality of the bivariate DP test statistic to the multivariate case.

**Theorem 4.** For $W \in \mathbb{R}^d$ having a sufficiently smooth density, there exists a sharpening function $\psi_p(.)$ reducing the local bias to order $p$, with $p$ sufficiently large so that there is a sequence of bandwidths $\varepsilon_n = Cn^{-\beta}$ with $C > 0$ and $\beta \in (1/(2p), 1/d)$, which guarantees that for a weakly dependent, strictly stationary process the sharpened test statistic $T_n^s$ satisfies

$$\sqrt{n}\frac{\left(T_n^s(\varepsilon_n) - q\right)}{S_n} \xrightarrow{d} \mathrm{N}(0,1)$$

where $S_n^2$ is a consistent estimator of the asymptotic variance of $\sqrt{n}\left(T_n^s(\varepsilon_n) - q\right)$.

***Proof*.** Given the necessary modifications to the leading terms of the MSE of the test statistic, provided in the online Appendix, the proof of asymptotic normality is analogous to that of Theorem 1 of DP. □



DOI: 10.1002/jae

In order to illustrate this in a practical application we return to the dimensionality problem described in Section 2.3. The original kernel estimator bias of order $O(\varepsilon^2)$, which was effectively blocking the consistency of the test, is reduced to $O(\varepsilon^4)$ by applying a sharpening function of the form

$$\psi_4(W) = I + \varepsilon^2 \frac{\kappa_2}{2} \frac{\hat{f}'(W)}{\hat{f}(W)} \tag{15}$$

where $I$ is the identity function, $\kappa_2$ is the second moment of the kernel and $\hat{f}'$ the estimator of the gradient of $f$. The lower order of the bias enables us to find a range of feasible $\beta$-values again, in this case $\beta \in (1/8, 1/4)$, for which the test statistic $T_n^s(\varepsilon_n)$ is asymptotically normal.

Following Choi and Hall (1999) we employ a Nadaraya–Watson estimator as a plug-in estimator for the ratio $\hat{f}'/\hat{f}$ in the DS function (15), giving rise to

$$\psi_4(W) = I + \frac{\varepsilon^2}{2\varepsilon_{\mathrm{DS}}^2} \frac{\sum_{i=1}^{n}(W_i - W) K\left(\frac{W - W_i}{\varepsilon_{\mathrm{DS}}}\right)}{\sum_{i=1}^{n} K\left(\frac{W - W_i}{\varepsilon_{\mathrm{DS}}}\right)} \tag{16}$$

where for simplicity we have taken the kernel function used in the data-sharpening step to be the same as that of the final density estimation step. In the context of density estimation, Choi and Hall (1999) consider constant ratios of $\varepsilon^2/(2\varepsilon_{\mathrm{DS}}^2)$. Here we need more flexibility, since often the DS bandwidth $\varepsilon_{\mathrm{DS}}$, to maintain consistency of the gradient estimator in the DS step, should converge to zero at a slower rate than $n^{-1/(\gamma+2)}$, which may be outside the range of rates $n^{-\beta}$ at which $\varepsilon$ is allowed to tend to zero as $n$ tends to infinity.

Besides DS there are several other methods for bias reduction in kernel density estimation. The literature distinguishes, among others, higher-order kernels (Granovsky and Muller, 1991), variable bandwidth estimators (Abramson, 1982), variable location estimators (Samiuddin and El-Sayyad, 1990) and parametric transformation methods (Abramson, 1984). Under sufficient smoothness of the underlying density, these all reduce the bias from $O(\varepsilon^2)$ to $O(\varepsilon^4)$. Although it is likely that these methods can also be successfully adapted to our setting, additional asymptotic theory for the adjusted test statistic would have to be developed. We therefore leave this for future consideration.

### 3.1. Consistency of the Multivariate Version of the DP Test

Theorems 1–3 directly extend to the multivariate case, provided that $\boldsymbol{Y}$ in the proofs is taken to represent not just $Y_t$ but $(Y_t, Q_t)$, where $Q_t$ is the (possibly vector-valued) extra variable to condition upon.

An important class of models that is covered by the multivariate version of Theorem 1 is the class of first-order general VAR models, defined as follows.

**Definition 3.** A first-order general VAR (general VAR(1)) model has the form $\boldsymbol{X}_t = \boldsymbol{G}(\boldsymbol{X}_{t-1}) + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{G}(\boldsymbol{x})$ is a vector-valued function of $\boldsymbol{x}$, and $\boldsymbol{\varepsilon}_t$ a white noise error term, possibly with nonzero covariance between its elements.

As for the bivariate VAR(1) model described above, the term 'general' refers to the fact that the conditional mean $\boldsymbol{G}(\boldsymbol{x})$ can be general and need not be a linear function of $\boldsymbol{x}$, as is the case for VAR models.

If the time series $\{(X_t, Y_t, Q_t)\}$ is generated by a general VAR(1) model, $Z_t = Y_{t+1}$ given $(x_t, y_t, q_t)$ can be represented as $Z_t = g(x_t, y_t, q_t) + \varepsilon_{t+1}$, where $\varepsilon_{t+1}$ is a zero-mean noise term which is independent of $(X_t, Y_t, Q_t)$, and $g(x, y, q)$ the regression function (conditional mean) of $Z$

given $(x, y, q)$. This corresponds to the situation covered by Theorem 1. Of course, since the labels of the variables are arbitrary, the same holds for each of the possible pairs of variables selected to be the $X$ and $Z$ variables.

### 3.2. Bandwidth Selection

An (asymptotically) MSE optimal bandwidth $\varepsilon^*$ is obtained by minimizing the contributions from the first and fourth terms on the right-hand side of the expansion (11) of the MSE, since if the conditions for asymptotic normality hold, these are the two most dominant contributions to the MSE after the third term (which is independent of the bandwidth). It can be easily verified that this leads to

$$\varepsilon^* = C^* n^{\frac{-2}{2\alpha+\gamma}} \tag{17}$$

with

$$C^* = \left( \frac{18\gamma q_2}{2\alpha (E[s(W)])^2} \right)^{\frac{1}{2\alpha+\gamma}} \tag{18}$$

Recall that $\gamma = d_X + d_Y + d_Z$ and $\alpha$ is the order of the local bias, which for a second-order kernel is 2 if no DS is applied, and $\alpha = p > 2$ when DS is applied. Throughout the simulations and applications we restrict ourselves to the cases $\alpha = 2$ without DS and $\alpha = p = 4$ with DS.

This expression for the optimal bandwidth is analogous to that derived in DP. DS changes the point-wise bias of the density estimator, intuitively affecting both the rate of convergence, i.e. the parameter $\alpha$, and the coefficient of the local bias term, i.e. $s(w_i)$.

**Numerical estimation of $C^*$ for a multivariate ARCH process.** In order to obtain some insights into the effects of DS on the optimal bandwidth selection, we consider a 3-variate ARCH process, representing the dimensionality problem from the previous section. The process considered is

$$\begin{aligned} Q_t &\sim \mathrm{N}\left(0, c + aQ_{t-1}^2\right) \\ X_t &\sim \mathrm{N}\left(0, c + aY_{t-1}^2\right) \\ Y_t &\sim \mathrm{N}\left(0, c + aQ_{t-1}^2\right) \end{aligned} \tag{19}$$

The process is set up such that $\{Q_t\}$ is driving both itself and $\{Y_t\}$, and $\{Y_t\}$ drives $\{X_t\}$. Although other designs are possible, this guarantees that $\{X_t\}$ is not Granger causing $\{Y_t\}$, corrected for the presence of $\{Q_t\}$. Due to the nature of the Granger causal links the process is stationary and mixing if the main driving variable $\{Q_t\}$ is, which is the case if $c > 0$ and $0 < a < 1$.

Because of the analytic intractability of $C^*$ we rely on Monte Carlo simulations to obtain an estimate for it. We simulate 100 independent realizations of the process defined by equations (19) with $a = 0.4$ and $c = 1$ for time series of length $n = 5,000$. Throughout the simulations and empirical applications we use the second-order kernel function $K(v) = (2\pi)^{-d_W/2} \exp\left(-v^T v/2\right)$, the standard multivariate Gaussian kernel.

We extract values for $E[s(W)]$ and $q_2$ based on the leading terms in the expansions (8) and (9) using standard kernel methods for density and derivative estimation, described by Wand and Jones (1995) and Silverman (1998). By substituting these estimates into equation (18) we estimate the optimal constant $C^*$ to be 2.4. This in turn allows us to calculate the optimal bandwidths $\varepsilon_n^*$ for various sample sizes using the optimal bandwidth formula (17). These bandwidths are reported in the row labeled '$\varepsilon^*$' in Table I.

The optimal bandwidths converge to zero slower with $n$ than those of DP. This is a direct result of the application of the DS method; given that the sharpened estimator has lower bias order, the density

Table I. Observed rejection rates of the $T_n^s(\varepsilon)$ test for process (19) for different sample sizes. The values represent observed rejection rates over 10,000 realizations, for nominal size 0.05

| sample size ($n$) | 100 | 200 | 500 | 1000 | 2000 | 5000 |
|---|---|---|---|---|---|---|
| $\varepsilon_{\mathrm{DS}}$ | 2.07 | 1.88 | 1.65 | 1.49 | 1.35 | 1.18 |
| $\varepsilon^*$ | 1.11 | 0.99 | 0.85 | 0.76 | 0.68 | 0.58 |
| Size | 0.015 | 0.02 | 0.027 | 0.032 | 0.036 | 0.043 |
| Power | 0.533 | 0.884 | 0.999 | 1.000 | 1.000 | 1.000 |

estimation kernel can be expanded to include a wider range of points, allowing also the variance to become smaller.

**Data-sharpening bandwidth.** As noted above, to obtain consistent estimates of the gradient of the density the DS bandwidth, $\varepsilon_{\mathrm{DS}}$, should tend to zero strictly slower than $n^{-1/(\gamma+2)}$, where $\gamma = \dim W$ (Fukunaga and Hostetler, 1975). In our simulations and applications we use a DS bandwidth equal to

$$\varepsilon_{\mathrm{DS}} = C_{\mathrm{DS}} n^{-1/(\gamma+3)} \tag{20}$$

where $C_{\mathrm{DS}}$ is calibrated at the value 4.0, as simulations suggest that this leads to a test with fairly good size and power properties for small samples.

### 3.3. Performance of the Multivariate Test

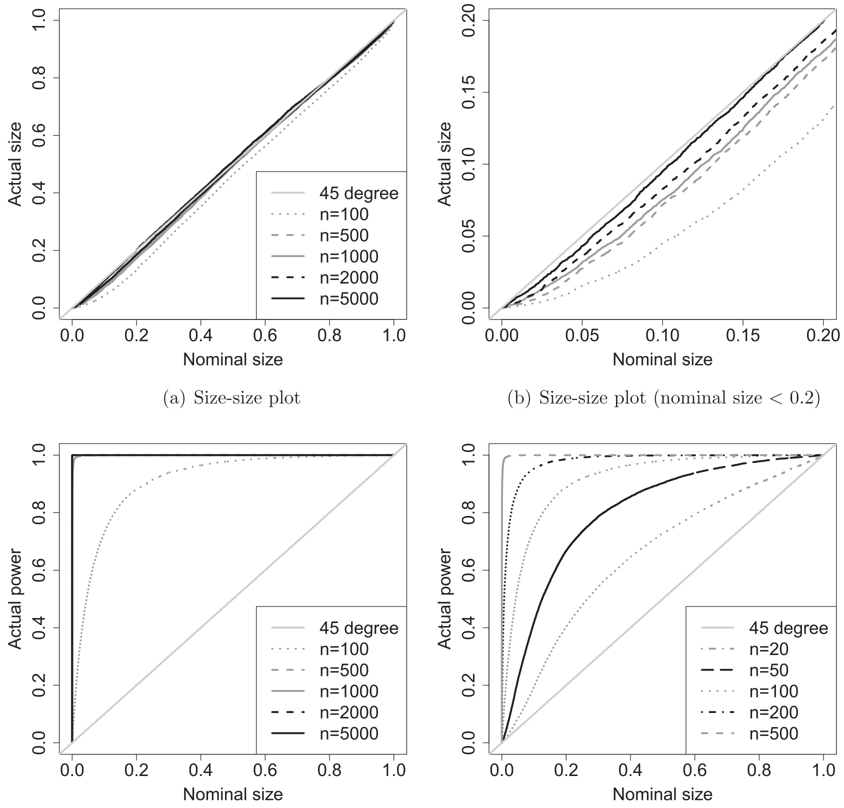
Given the above estimates for the optimal bandwidth values, we next turn to the assessment of the performance of the DS-augmented DP test. Here we rely on Monte Carlo methods again. Since the process (19) matches one of the best-known stylized facts of empirical financial returns time series (conditional heteroskedasticity), we use it (with $a = 0.4$ and $c = 1$) as the underlying process for the size assessment of the test, by testing for Granger non-causality from $\{X_t\}$ to $\{Y_t\}$. For the power assessment we use the same process, but now test Granger non-causality from $\{Y_t\}$ to $\{X_t\}$ so that, conditional on $\{Q_t\}$, the null hypothesis of Granger non-causality does not hold.

The results obtained from 10,000 simulations for various time series lengths are presented in the size–size and size–power plots in Figure 1. Table I gives the size and power for nominal size 0.05.

One may readily observe from Table I that the test demonstrates already very high power for time series length 500 and larger. For a nominal size of 0.05, the power ranges from 0.533 for $n = 100$ via 0.999 for $n = 500$, to 1.000 for longer time series. The results suggest that the test yields satisfactory size and power for time series of length 500 and larger.

It can be observed from Figure 1(a, b) that under the null hypothesis the size of the test converges to the nominal size as the sample size increases. However, for relatively short samples the test is somewhat conservative. Nevertheless, one can clearly observe the benefits of DS compared to the original DP setting; while the bivariate test was found by DP to be persistently conservative in multivariate environments, the size of the DS-augmented test converges to the ideal, nominal, rejection probability.

To observe the direct effects of the DS step on the test performance, we carried out two additional control experiments. First, we compared the size and power of the test under the same optimal multivariate bandwidth but without the DS step (so that $\psi_4(W)$ is the identity function). Secondly, we checked the performance of the test without DS, combined with the bandwidths suggested by DP, albeit here in a Gaussian setting. The size–size and size–power plots for the control experiments are given in the online Appendix. The control experiments show that the DS step mainly has a beneficial effect on the size of the test, making the test less conservative. The effect on the power is small.




(a) Size-size plot (b) Size-size plot (nominal size < 0.2)

(c) Size-power plot (d) Size-power plot (shorter samples)

Figure 1. Size–size and size–power plots of the $T_n^s(\varepsilon)$ test for the 3-variate process (equation (19)) for different sample sizes under a shrinking bandwidth condition, based on 10,000 independent realizations: (a) size–size plot; (b) size–size plot (nominal size < 0.2); (c) size–power plot; (d) size–power plot (shorter samples).

Table II. Size/power of DS-augmented DP test (columns DS), DP test with DS-implied bandwidths but without the DS-correction (columns DS*) and original DP test (columns DP) for process (19) with reversed direction of causality for sample size 500. We report observed rejection rates over 10,000 realizations for nominal size 0.01, 0.05 and 0.1 and for causality strengths (parameter $a$ from equation (19)) of 0, 0.01, 0.05, 0.1, 0.2 and 0.3

| | Nominal size = 0.01 | | | Nominal size = 0.05 | | | Nominal size = 0.1 | | |
|---|---|---|---|---|---|---|---|---|---|
| | DS | DS* | DP | DS | DS* | DP | DS | DS* | DP |
| $\boldsymbol{a = 0}$ | 0.002 | 0.002 | 0.001 | 0.021 | 0.025 | 0.021 | 0.059 | 0.063 | 0.058 |
| $\boldsymbol{a = 0.01}$ | 0.002 | 0.003 | 0.002 | 0.032 | 0.033 | 0.030 | 0.083 | 0.090 | 0.080 |
| $\boldsymbol{a = 0.05}$ | 0.015 | 0.018 | 0.014 | 0.115 | 0.118 | 0.104 | 0.238 | 0.232 | 0.211 |
| $\boldsymbol{a = 0.1}$ | 0.082 | 0.084 | 0.067 | 0.336 | 0.323 | 0.287 | 0.522 | 0.496 | 0.460 |
| $\boldsymbol{a = 0.2}$ | 0.498 | 0.478 | 0.421 | 0.818 | 0.798 | 0.763 | 0.915 | 0.900 | 0.877 |
| $\boldsymbol{a = 0.3}$ | 0.882 | 0.872 | 0.836 | 0.981 | 0.977 | 0.970 | 0.994 | 0.872 | 0.990 |

To obtain an assessment of the influence of the DS on the power of the test for a moderate sample size, we carry out an additional simulation study for sample size $n = 500$ for various strengths of the Granger causal links in process (19) with reversed direction of dependence. The results, based on 10,000 realizations, are presented in Table II. One can verify that the DS-augmented version of



DOI: 10.1002/jae

the DP test (columns DS) always yields larger power in a 3-variate setting than the DP test (columns labeled DP). Power gains are particularly visible for a small nominal size of 0.01 and higher values of the parameter $a$. For instance, for nominal size 0.01 and $a = 0.2$, the DS-augmented test has a power that is larger by 0.077. The test with the DS-implied bandwidth values but without the DS correction (columns labeled DS*) has slightly larger power than the DS specification for small values of $a$. For those the power gains in the DS* setting are marginal and range from 0.001 to 0.007. This effect, however, vanishes in favor of the DS test as the nominal size and/or strength of Granger causality increase.

**Computational time.** One might be concerned about the increase in computational time due to the addition of the DS step, which leads to an additional operation of order $O(n)$ for each of the $n$ sample points. However, since the bivariate DP test already is of order $O(n^2)$, this only gives rise to a constant factor (of about 2) in computational time compared to the bivariate test.

## 4. NONLINEAR GRANGER CAUSALITY IN THE US GRAIN MARKET

In order to show the practical application of the sharpened DP test, we choose the US grain market as it offers an intuitive and straightforward context for testing for Granger non-causality. There is a common agreement among professionals that any causal relation between prices of different crops has to be corrected, at any given moment, for the expectations that are maintained regarding future weather conditions that may influence crop production in either quality or quantity (see, for instance, Popp *et al.*, 2003, Carreck and Christian, 1997). A conditioning variable quantifying weather expectations would serve as a perfect example of an additional $Q$ variable, in line with the 3-variate example in the previous sections.

We investigate corn, beans and wheat as they are the most representative grains traded in the US grain market. To represent weather expectations we use prices of 1-month-ahead future contracts, traded in USD at the Chicago Board of Trade (CBoT). The weather variable is obtained by using rolling monthly futures on heating degree days (HDD), averaged over Philadelphia, New York, Portland, Chicago and Cincinnati. Daily time series comprise the period from 1 September 2010 until 6 March 2013 giving a total of 633 observations. The data are obtained from Bloomberg.

We take logarithms of the observations of all time series to stabilize their variance and subsequently test whether they are stationary using standard unit root tests; the augmented Dickey–Fuller (ADF) test and the Phillips–Perron (PP) test, as described by Fuller (1995) and Phillips and Perron (1988), respectively. The results for the two tests, provided in the online Appendix, are essentially identical as long as the number of lags selected by the ADF test (based on Bayes' information criterion) is zero, which is the case for all grains. For HDD the number of lags selected by the ADF was nonzero, explaining the modest differences between the ADF test and the PP test statistics, although both tests still lead to identical conclusions: for the log prices time series no evidence was found that the time series were stationary, except for the log price time series of corn, which was found to be stationary at the 5% significance level. Johansen cointegration tests (both trace and maximum eigenvalue) applied to the three non-stationary log price variables jointly did not indicate the presence of cointegration at the 5% significance level. Since, moreover, the first differences of all variables (log returns) were found to be stationary, we decided to base the Granger non-causality tests in the market on the log returns time series of all variables.

In the analysis we consider pair-wise relations and relations in the complete system setting separately. In the former we take into account the direct relations between two grains only and in the latter we look at the analysis with all grains included, since in the system setting the $Q$ variable is two dimensional, consisting of the remaining grain variable, $Q_1$, say, and the weather variable, $Q_2$, say.

In order to benchmark the results, we compare them with the standard linear Granger causality setting, as proposed by Granger (1969). We also investigate the causal relations in the VAR-filtered residuals, to see whether any of the Granger causal links detected are indeed nonlinear. We study the explicit role of the weather variable by comparing our results with the original DP test, i.e. without any conditioning variable. To make sure that the kernel functions used across the methodologies correspond, we use a version of the standard DP test with a Gaussian kernel.

In the analysis we set the lag of each conditioning variable equal to 1, as suggested by the Bayesian information criterion for the VAR specification. In the presence of DS we use the bandwidths suggested by the optimal bandwidth formula (17) with the optimal value $C^* = 2.4$ suggested by our numerical simulations in the 3-variate setting with $n = 5,000$. To calculate the bandwidths for the DS estimators we use the DS bandwidth formula (20) with $C_{\mathrm{DS}} = 4.0$ as also suggested by our simulation results. The bandwidths for the standard DP test were taken from DP, but scaled by $1/\sqrt{3}$ to adjust the variance of the Gaussian kernel to that of the uniform kernel used in DP. This leads to a bandwidth determined also by the optimal bandwidth formula (17), but now with $\alpha = 2$ (rather than 4) and $C^* = 4.6$ (instead of 2.4). For the sample size (633) of our time series the resulting bandwidths are 0.73 for the standard DP test, 0.82 for the sharpened test in the 3-variate settings and 0.86 for the sharpened test in the 4-variate system setting. Before running the tests, we standardize the data by transforming the marginals to either normal or uniform.

The results for the pair-wise relations are presented in panels (a) and (b) in Table III and for the complete system in panels (a) and (b) in Table IV.

One may readily observe from Table III(a) that the US grain market does not show much linear Granger causality, the only exception being weak evidence of a linear impact of beans on future corn prices in all settings. After VAR filtering this relation disappears, as expected.

Interestingly, our results suggest that the dynamical relations between US grain prices exhibit many nonlinearities. Looking at the basic pair-wise setting (Table IIIa), there is evidence for strong causal linkages between corn and wheat. Also, we find weak significant effects from beans to corn prices. If, however, we condition on weather forecasts (Table IIIb), some of the linkages vanish or become weaker, in particular in the uniform transformation setting. This suggests that weather forecasts to a large extent drive many of the apparent Granger causal relations in the US grain market. From our pair-wise general results it seems that conditioning on weather conditions mostly affects the wheat–corn Granger causal relations found but to a lesser extent also the corn–beans relations.

In the basic system setting the corn–wheat causal relation is preserved (Table IVa), being significant also after linear filtering. Also, the linear beans–corn relation, present in the basic pair-wise setting, is present, but the nonlinear beans–corn relation has disappeared relative to the basic pair-wise setting. As the results in Table IV(b) show, after conditioning on weather forecasts we observe that the majority of the relations are preserved (only the VAR-filtered wheat–corn relation appears to gain in significance). Interestingly, the system setting exhibits many regularities from the pair-wise study, although the influence of wheat on corn is less significant.

Our analysis suggests that the causality is directed from corn to the wheat prices, with a weak feedback. Those causalities are still present after VAR filtering, evidencing their nonlinear nature. The corn–beans relations show a limited nonlinear Granger causal relationship from beans to corn, which is no longer present after conditioning on weather forecasts and in the system settings.

A straightforward explanation of our results might be that the nonlinear causal relation found in the corn–wheat market is directed primarily from larger to smaller markets, with a feedback effect echoing the initial disturbance. Corn is the most heavily traded grain on the CBoT. Intuitively, larger markets should affect those of smaller size as they are deeper and more liquid (Sari *et al.*, 2012). This reasoning is in fact in line with our previous finding on the role of the weather forecasts in the grain market. Since the majority of shocks in the grain market are weather-related, they serve as a common factor affecting each of the grain prices and mitigating the effects of the grain-specific shocks. Controlling for weather expectations therefore reveals causal relations between grain-specific shocks, which seem to be propagating from larger to smaller markets.

Table III. Causality results for the pair-wise relations of the log returns on the US grain market. Panel (a) represents the specification without conditioning on weather (HDD) whereas panel (b) reflects the specification with conditioning on weather

| | | Linear Granger causality | | | | Nonlinear Granger causality (N) | | | | Nonlinear Granger causality (U) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variables | | Raw data | | VAR residuals | | Raw data | | VAR residuals | | Raw data | | VAR residuals | |
| X | Y | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X |
| (a) *Without conditioning on weather* | | | | | | | | | | | | | |
| Corn | Wheat | | | | | *** | *** | *** | *** | ** | ** | ** | ** |
| Corn | Beans | | * | | | | | | | | * | | * |
| Beans | Wheat | | | | | | | | | | | | |
| (b) *With conditioning on weather* | | | | | | | | | | | | | |
| Corn | Wheat | | | | | *** | ** | *** | ** | | ** | * | * |
| Corn | Beans | | * | | | | | | | | | | |
| Beans | Wheat | | | | | | | | | | | | |

*Note:* Asterisks denote statistical significance at *10%, **5% and ***1%. Period: 1 September 2010 - 6 March 2013. Nonlinear tests are performed on standardized data, transformed to (N)ormal or (U)niform marginals. The number of lags is $l_X = l_Y = 1$ from the Bayesian information criterion.

Table IV. Causality results for the system setting of the log returns on the US grain market. Panel (a) represents the specification without conditioning on weather (HDD), whereas panel (b) reflects the specification with conditioning on weather

| Variables | | Linear Granger causality | | | | Nonlinear Granger causality (N) | | | | Nonlinear Granger causality (U) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Raw data | | VAR residuals | | Raw data | | VAR residuals | | Raw data | | VAR residuals | |
| X | Y | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X |
| (a) *Without conditioning on weather* | | | | | | | | | | | | | |
| Corn | Wheat | | | | | *** | ** | *** | * | ** | * | ** | * |
| Corn | Beans | | * | | | | | | | | | | |
| Beans | Wheat | | | | | | | | | | | | |
| (b) *With conditioning on weather* | | | | | | | | | | | | | |
| Corn | Wheat | | | | | *** | ** | *** | ** | ** | * | ** | * |
| Corn | Beans | | * | | | | | | | | | | |
| Beans | Wheat | | | | | | | | | | | | |

*Note:* Asterisks denote statistical significance at *10%, **5% and ***1%. Period: 1 September 2010 - 6 March 2013. Nonlinear tests are performed on standardized data, transformed to (N)ormal or (U)niform marginals. The number of lags is $l_X = l_Y = l_{Q_1} = 1$ from the Bayesian information criterion.



DOI: 10.1002/jae

## 5. SUMMARY AND CONCLUDING REMARKS

In this paper we extend the bivariate DP test for Granger non-causality to the multivariate case. We first show why the asymptotic theory of the nonparametric test proposed by DP does not hold in a multivariate setting. The reason is that the local bias and the variance cannot both tend to zero sufficiently fast with the sample size in the multivariate case. In order to resolve this we propose using a sharpened form of the test statistic, which under mild regularity conditions again leads to asymptotic normality of the standardized test statistic. We find that the sharpening function reduces the bias of the original estimator without affecting the order of the variance, as originally suggested by Hall and Minnotte (2002). We assess the size and power of the sharpened test numerically, and find that it has correct size and larger power than a naive multivariate implementation of the DP test.

Additionally, we identify a number of classes of processes within which the bivariate test and its multivariate generalization are consistent against any fixed alternative. These include first-order multivariate linear/nonlinear AR processes, as well as multivariate normal processes.

To show the practical side of our study, we apply our test to the US grain market since, because of its weather-dependent structure, this serves as an ideal context to assess our methodology. We consider Granger causality between corn, beans and wheat, with and without conditioning on weather expectations, represented by future contracts on HDD. Our results suggest that the US grain market exhibits many nonlinear relations, particularly between corn and wheat prices. We discover that weather conditions seem to drive the causal relation from wheat to corn and also affect the corn–beans relation by acting as a common factor. Controlling for the common factor, we reveal the true nonlinear Granger causal relations between the corn and wheat markets, suggesting that the causality spreads from larger, i.e. deeper and more liquid, to smaller markets.

Our results might have important further implications for food market analysis. As suggested by Gilbert (2010), future contracts are the major transition channel through which macro variables affect food prices. Understanding the structure of the feedback loops in the dynamics governing these markets is of great societal relevance, as it may prevent possible bubbles and instant food price rises, such as those observed during 2007 and 2008.

### ACKNOWLEDGEMENTS

The authors would like to thank Peter Hall, Dimitris Kugiumtzis, Kosma Matysiak, Valentyn Panchenko and the participants of the 21st Symposium of the Society for Nonlinear Dynamics and Econometrics for their useful comments. Cees Diks acknowledges the support of the Netherlands Organisation for Scientific Research (NWO) under the Complexity Programme. Marcin Wolski acknowledges the financial support from the European Doctorate in Economics – Erasmus Mundus. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. Any views expressed are those of the authors and do not necessarily represent the views of the European Investment Bank.

### REFERENCES

Abramson IS. 1982. On bandwidth variation in kernel estimates-a square root law. *Annals of Statistics* **10**(4): 1217–1223.

Abramson IS. 1984. Adaptive density flattening-a metric distortion principle for combating bias in nearest neighbor methods. *Annals of Statistics* **12**(3): 880–886.

Bai Z, Wong WK, Zhang B. 2010. Multivariate linear and nonlinear causality tests. *Mathematics and Computers in Simulation* **81**(1): 5–17.

Carreck N, Christian D. 1997. A study of grain dormancy and viability in spring barley. *European Journal of Agronomy* **6**(3–4): 155–161.

Cartwright N. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press: Cambridge, UK.

Choi E, Hall P. 1999. Data sharpening as a prelude to density estimation. *Biometrika* **86**(4): 941–947.

Denker M, Keller G. 1983. On U-statistics and v. Mises' statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **64**(4): 505–522.

Diks C, Panchenko V. 2005. A note on the Hiemstra–Jones test for Granger non-causality. *Studies in Nonlinear Dynamics and Econometrics* **9**(2): 1–7.

Diks C, Panchenko V. 2006. A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics and Control* **30**(9–10): 1647–1669.

Fukunaga K, Hostetler LD. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory* **21**(1): 32–40.

Fuller WA. 1995. *Introduction to Statistical Time Series.* Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience: New York.

Gilbert CL. 2010. How to understand high food prices. *Journal of Agricultural Economics* **61**(2): 398–425.

Granger CWJ. 1969. Investigating causal relations by econometric models and cross- spectral methods. *Econometrica* **37**(3): 424–438.

Granovsky BL, Muller H G. 1991. Optimizing kernel methods: A unifying variational principle. *International Statistical Review/Revue Internationale de Statistique* **59**(3): 373–388.

Guo S, Ladroue C, Feng J. 2010. Granger causality: theory and applications. In *Frontiers in Computational and Systems Biology*, Feng J, Fu W, Sun F (eds), Vol. 15 of Computational Biology. Springer: New York; 83–111.

Hall P, Minnotte MC. 2002. High order data sharpening for density estimation. *Journal of the Royal Statistical Society, Series B* **64**(1): 141–157.

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

Phillips PCB, Perron P. 1988. Testing for a unit root in time series regression. *Biometrika* **75**(2): 335–346.

Popp MP, Dillon CR, Keisling TC. 2003. Economic and weather influences on soybean planting strategies on heavy soils. *Agricultural Systems* **76**(3): 969–984.

Powell JL, Stoker TM. 1996. Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics* **75**(2): 219–316.

Samiuddin M, El-Sayyad GM. 1990. On nonparametric kernel density estimates. *Biometrika* **77**(4): 865–874.

Sari R, Hammoudeh S, Chang CL, McAleer M. 2012. Causality between market liquidity and depth for energy and grains. *Energy Economics* **34**(5): 1683–1692.

Scott DW. 1992. *Multivariate Density Estimation. Wiley Series in Probability and Mathematical Statistics*. Wiley-Interscience: New York.

Silverman BW. 1998. *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC: London.

Skaug H, Tjøstheim D. 1993. Nonparametric tests of serial independence. In *Developments in Time Series Analysis*, Subba Rao T (ed). Chapman & Hall: London; 207–229.

Su L, White H. 2014. Testing conditional independence via empirical likelihood. *Journal of Econometrics* **182**(1): 27–44.

Wand MP, Jones MC. 1995. *Kernel Smoothing.* Monographs on Statistics and Applied Probability. Chapman & Hall: New York.




DOI: 10.1002/jae