

# Nonlinear Granger Causality: Guidelines for Multivariate Analysis

Cees Diks<sup>1</sup> and Marcin Wolski<sup>1,2</sup>

<sup>1</sup>Center for Nonlinear Dynamics in Economics and Finance (CeNDEF), University of Amsterdam, Valckenierstraat 65-67, 1018 XE Amsterdam, The Netherlands

<sup>2</sup>Bielefeld Graduate School of Economics and Management (BiGSEM), Bielefeld University, Universitätsstraße 25, D-33615 Bielefeld, Germany

## Abstract

In this paper we propose an extension of the nonparametric Granger causality test, originally introduced by Diks and Panchenko [2006]. A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics & Control* 30, 1647-1669]. We show that the basic test statistics lacks consistency in the multivariate setting. The problem is the result of the kernel density estimator bias, which does not converge to zero at a sufficiently fast rate when the number of conditioning variables is larger than one. In order to overcome this difficulty we apply the data-sharpening method for bias reduction. We then derive the asymptotic properties of the ‘sharpened’ test statistics and we investigate its performance numerically. We conclude with an empirical application to the US grain market.

**JEL Codes:** C12, C51, E3

**Keywords:** Financial time series, Granger causality, Nonparametric methods

## 1 Introduction

Since the introduction of Granger causality over four decades ago (Granger, 1969), the body of literature on this topic has grown substantially, becoming a standard methodology not only

---

The authors would like to thank Peter Hall, Kosma Matysiak, Valentyn Panchenko and the participants of the 21st Symposium of the Society for Nonlinear Dynamics and Econometrics for their useful comments. Marcin Wolski acknowledges also the financial support from the European Doctorate in Economics - Erasmus Mundus. The corresponding e-mail addresses are: C.G.H.Diks@uva.nl and M.Wolski@uva.nl.

among economists and econometricians, but also finding the followers in physics or even biology (Guo et al., 2010). Not surprisingly, it alleviated an ongoing discussion on the nature and validity of the concept, pointing out its methodological limitations. Although, the spectrum of arguments against the idea of Granger causality is very broad, the main line of criticism follows from the very simple nature of the dependence relations in the economy, which Granger causality originally assumes (Cartwright, 2007). The scope of this paper is to contribute to the discussion allowing for a more complex structural setting in the nonparametric Granger causality testing.

Imagine a strictly stationary bivariate process  $\{(X_t, Y_t)\}$ . We say that  $\{X_t\}$  is a Granger cause of  $\{Y_t\}$  if past and current values of  $X$  contain additional information on future values of  $Y$  that is not contained in past and current  $Y$ -values alone. If we denote the information contained in past observations  $X_s$  and  $Y_s$ ,  $s \leq t$ , by  $\mathcal{F}_{X,t}$  and  $\mathcal{F}_{Y,t}$ , respectively, and let ‘ $\sim$ ’ denote equivalence in distribution, the formal definition is:

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

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

Clearly, such a definition is very simplistic and seems to be inappropriate to apply in the complex environments. An obvious shortcoming is the fact that the vectors of interests are assumed to be univariate, making the whole problem detached from reality. In other words, this methodology does not allow to control for *every possible source of variation of every kind [...]* as argued by (Cartwright, 2007). An advantage of such a general definition is, however, that it does not assume any parametric relations between the time series and instead it focusses on the conditional dependence of distributions only<sup>1</sup>.

The most commonly used nonparametric test for the above hypothesis testing (Def. 1) is the one proposed by Hiemstra and Jones (1994). Its main advantage lies in a very clear and intuitive reasoning together with a strong asymptotic theory, derived even for a multivariate setting<sup>2</sup> (Bai et al., 2010). Nevertheless, the test can severely over-reject if the null is satisfied (Diks and Panchenko, 2005). Therefore, Diks and Panchenko (2006) (hereafter DP) proposed a new test statistics which corrects for this shortcoming but, as it turns out, because of the large kernel estimator bias the DP test lacks consistency in the multivariate setting.

The goal of this paper is therefore twofold. First, in order to reduce the kernel estimator bias we apply the data-sharpening method (Hall and Minnotte, 2002) and we derive the asymptotic properties for the sharpened DP test in a multivariate setting. Second, we investigate its

---

<sup>1</sup>This brings additional modeling flexibility and does not bind us to the linear autoregressive model as originally proposed by Granger (1969).

<sup>2</sup>Throughout the paper, we will refer to a multivariate setting by a situation where vector  $\{X_t\}$  is allowed to be multidimensional, i.e.  $\{X_t\} = \{X_{1,t}, X_{2,t}, \dots, X_{m,t}\}$ ,  $m \geq 1$ . In principle, the dimensions of  $\{X_t\}$  vector might describe its corresponding lags, i.e.  $X_{1,t} = X_{t-1}$ ,  $X_{2,t} = X_{t-2}$  etc, so that by the bivariate case we refer to the situation where  $\{X_t\}$  is univariate.

performance both numerically and empirically on the US grain market. We chose this specific market due to its straightforward causal relationship, where the price of each grain is influenced not only by the other grains, but to a large extent the whole market is driven by weather forecasts. Therefore, it serves as an almost ideal environment to test our new methodology in practice.

This paper is organized as follows. Section 2 discusses the asymptotic properties of the original DP test and shows why it lacks consistency in the multivariate setting. In Section 3 we replace the standard kernel density estimator by its sharpened form and we show that the new test statistics is asymptotically normally distributed. We confirm the 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 DP test

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

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

where  $X_t^{l_X} = (X_{t-l_X+1}, \dots, X_t)$  and  $Y_t^{l_Y} = (Y_{t-l_Y+1}, \dots, Y_t)$ . 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 just write  $W = (X, Y, Z)$ , where the latter is a random vector with the invariant distribution of  $(X_t^{l_X}, Y_t^{l_Y}, Y_{t+1})$ .

Let us consider the simplest setting, where  $l_X = l_Y = 1$  so that  $W = (X, Y, Z)$  denotes a three-variate random variable, distributed as  $W_t = (X_t, Y_t, Y_{t+1})$ . (We investigate the problems associated with increased dimensionality in the next section. Throughout we will assume that  $W$  is a continuous random variable.) The DP test restates the null hypothesis in terms of the joint probability distribution  $f_{X,Y,Z}(X, Y, Z)$  and its marginals, i.e.

$$q \equiv E [f_{X,Y,Z}(X, Y, Z)f_Y(Y) - f_{X,Y}(X, Y)f_{Y,Z}(Y, Z)] = 0. \quad (1)$$

Given the standard correlation integral density estimator

$$\hat{f}_W(W_i) = \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} I\left(\frac{|W_i - W_j|}{\varepsilon}\right), \quad (2)$$

where  $\varepsilon$  is a bandwidth and  $I(\cdot)$  is the indicator function taking values 1 for any argument within the unit circle, one may readily find a natural estimator of  $q$  being given as

$$T_n(\varepsilon) = \frac{(n-1)}{n(n-2)} \sum_i (\hat{f}_{X,Y,Z}(X_i, Y_i, Z_i) \hat{f}_Y(Y_i) - \hat{f}_{X,Y}(X_i, Y_i) \hat{f}_{Y,Z}(Y_i, Z_i)). \quad (3)$$

The asymptotic behavior of  $T_n(\varepsilon)$  follows directly from the reasoning originally designed for the MSE (mean squared error) optimal bandwidth selection under the shrinking conditions, developed by Powell and Stoker (1996). The test statistics has a corresponding third order U-statistics representation with a kernel given by  $\tilde{K}(W_i, W_j, W_k)$ . Let us denote  $\tilde{K}_1(w_1) = E[\tilde{K}(w_1, W_2, W_3)]$  and  $\tilde{K}_2(w_1, w_2) = E[\tilde{K}(w_1, w_2, W_3)]$ , and assume that the rates of convergence of the pointwise bias as well as the second moment kernel expansions depend on the bandwidth size in the following way (in fact these are the conditions imposed by Powell and Stoker (1996))

$$\tilde{K}_1(w_i, \varepsilon) - \lim_{\varepsilon \rightarrow 0} \tilde{K}_1(w_i, \varepsilon) = s(w_i)\varepsilon^\alpha + s^*(w_i, \varepsilon), \quad \alpha > 0, \quad (4)$$

$$E[(\tilde{K}_2(W_1, W_2))^2] = q_2\varepsilon^{-\gamma} + q_2^*(\varepsilon), \quad \gamma > 0, \quad (5)$$

$$E[(\tilde{K}(W_1, W_2, W_3))^2] = q_3\varepsilon^{-\delta} + q_3^*(\varepsilon), \quad \delta > 0, \quad (6)$$

where the remainder terms are negligible, i.e.  $E\|s^*(W_i, h)\|^2 = o(h^{2\alpha})$ ,  $(q_2^*(\varepsilon))^2 = o(\varepsilon^{-\gamma})$  and  $(q_3^*(\varepsilon))^2 = o(\varepsilon^{-\delta})$ . Parameters  $\alpha$ ,  $\gamma$  and  $\delta$  follow directly from the specification of the kernel function  $\tilde{K}$  and might be derived analitically. In fact, it might be verified that  $\alpha$  is of the same magnitude as the local kernel estimator bias and Diks and Panchenko (2006) show that two remaining parameters depend on the dimensionality of the system as  $\gamma = d_x + d_y + d_z$  and  $\delta = d_x + 2d_y + d_z$ .

Having pointed that out, the MSE of the test statistics might be expressed as

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

where  $C_0 = 2\text{Cov}\left[\lim_{\varepsilon \rightarrow 0} \tilde{K}_1(W_i, \varepsilon), s(W_i)\right]$ . In order to guarantee asymptotic normality of  $T_n(\varepsilon)$  all the  $\varepsilon$ -dependent terms in Eq. (7) have to be  $o(n^{-1})$ . Given the bandwidth shrinking condition, i.e.  $\varepsilon = Cn^{-\beta}$ , one may find that this implies

$$\sqrt{n} \frac{T_n(\varepsilon_n) - q}{\sigma} \xrightarrow{d} N(0, 1) \quad \text{iff} \quad \frac{1}{2\alpha} < \beta < \frac{1}{d_x + d_y + d_z}, \quad (8)$$

with  $\sigma^2 = 9\text{Var}[\lim_{\varepsilon \rightarrow 0} \tilde{K}_1(W_i, \varepsilon)]$ .

Clearly, given the standard kernel density estimator with bias of order 2 and the basic model specification 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, given that there is no dependence between vectors  $W_i$  and assuming suitable mixing conditions (see Denker and Keller (1983a)), one may extend the proof by taking into account covariances between local density estimators.

## 2.1 The dimensionality problem

Let us now consider what happens if we increase dimensionality by one. For clarity purposes, imagine that we would like to condition the causal relationship on one more variable, denote it by  $Q$ , so that the hypothesis of conditional independence becomes<sup>3</sup>

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

Let us keep  $l_X = l_Y = l_Q = 1$ . Following the reasoning from the previous section, one may find that the asymptotic normality condition requires  $\beta$  to be in range between  $1/(2\alpha)$  and  $1/(d_x + d_y + d_z + d_q)$ . Given the same standard kernel density estimator with the local bias of order 2, one observes that if we increase the dimensionality of the original problem by any number  $v \geq 1$ , there is no feasible  $\beta$ -region which would endow  $T_n(\varepsilon)$  with asymptotic normality.

The associated problem results from a too large expected pointwise kernel estimator bias, i.e.  $E[s(W_i)]$ . By increasing the vector space, we decrease the estimator precision, which seems to play a crucial role in the MSE of the test statistics.

One may relate this problem 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 arbitrary small region of radius  $\mu < 1$ , it might be shown that as  $d \rightarrow \infty$  the number of points within  $[-\mu, \mu]^d$  tends to 0. Straightforward implication suggests that in finite higher dimensional spaces, the smoothing parameter should be larger in order to capture similar number of points. Nevertheless, by increasing the bandwidth window we decrease the precision of the estimator, violating the consistency of the test statistics in this case.

There are several methods which could decrease the dimensionality problem. Scott (1992) suggests principal components method, projection pursuit or informative components analysis. These solutions, however, put additional boundaries on the underlying structure of the data. For instance, they might be of a great advantage when dealing with 100-dimensional spaces where one could assume that the data structure falls into the 20-dimensional manifold. In our example it is very likely, however, that the minimum number of independent manifolds

---

<sup>3</sup>In practice, it is difficult to find an explicit representation of  $Q$  variable. However, one may think of increased dimensionality problem as conditioning on more than one lag, for instance  $Q_t = X_{t-1}$ .

is larger than 3 so that the dimension reduction does not necessarily have to improve the test performance. Moreover, as argued by Cartwright (2007), we do not want to decrease the complexity of the environment.

Another solution is precision improvement, or in other words, reduction of the estimator bias. Since it does not assume any particular underlying data structure, it is of greater advantage in our setting.

### 3 Data sharpening as a bias reduction method

The intuition behind Data Sharpening (DS) is to slightly perturb the original dataset by a sharpening function  $\psi_p(\cdot)$  in order to obtain the desirable properties of the estimator (here  $p$  is the order of bias reduction). The idea of the perturbation is to concentrate points where they were already dense and thin them where they were originally sparse. The explicit form of  $\psi_p(\cdot)$  depends then on the order of the bias reduction we would like to get but the technique might be in principle applied to obtain arbitrary low levels of bias reduction (Hall and Minnotte, 2002).

There are several advantages of DS over the standard bias reduction techniques. First, as mentioned before, it allows for very high levels of bias reduction. Since testing for Granger causality is widely recognized for its practical purposes, the universality of a method is of a great importance. Second, as we confirm in our study, it does not affect the kernel function directly, which leaves the properties of the MSE of the test statistics untouched. Third, it is easy and straightforward to implement, even in a multivariate setting.

With respect to Eq. (2), let us consider a sharpened form of the estimator

$$\hat{f}_W^s(W_i) = \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} K_{multi} \left( \frac{W_i - \psi_p(W_j)}{\varepsilon} \right), \quad (9)$$

where  $K_{multi}(W) = (2\pi)^{-d_W/2} \exp(-1/2W^T W)$  is the standard multivariate Gaussian kernel, as described in Wand and Jones (1995) and Silverman (1998).<sup>4</sup>

We obtain the sharpened form of the test statistics,  $T_n^s(\varepsilon)$ , by substituting the sharpened estimators into Eq. (3). As we show in the Appendix, the pointwise bias is of order  $o(\varepsilon^p)$  with other properties of the kernel  $\hat{K}$  being the same. This in fact makes the bias of  $T_n^s(\varepsilon)$  (from Eq. (4)) being  $\alpha = p$  with parameters  $\gamma$  and  $\delta$  from Eq. (5) and Eq. (6) unchanged. This reasoning might be summarized in the following corollary, which is a generalization of the theorem in Diks and Panchenko (2006) and proposition in Hall and Minnotte (2002)

**Corollary 1.** *For any sufficiently smooth, continuous and infinitely differentiable density, there exist a sharpening function  $\psi_p(\cdot)$ , where  $p$  is the order of bias reduction, for which one may*

---

<sup>4</sup>In principle, our reasoning holds for any sufficiently smooth, symmetric and multiplicative probability density as a kernel function. Correlation integral kernel, as originally applied by Diks and Panchenko (2006), proves not to be smooth enough which led us to the standard Gaussian kernel.

find a sequence of bandwidths  $\varepsilon_n = Cn^{-\beta}$  with  $C > 0$  and  $\beta \in (1/(2p), 1/D)$ , where  $D < \infty$  is the total dimensionality of the problem, which guarantees that for a weakly-dependent process the sharpened test statistic  $T_n^s$  satisfies:

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

where  $S_n$  is the autocorrelation consistent estimator for  $\sigma^2$ .

The proof of Corollary 1 might be found in the Appendix A.1.

In order to illustrate its practical application, let us consider the same dimensionality problem as described in Section 2.1. The original kernel estimator bias of order  $o(\varepsilon^2)$ , which was effectively blocking the consistency of the test, might be reduced to  $o(\varepsilon^4)$  by applying the sharpening function of the form

$$\psi_4(W) = I + \varepsilon^2 \frac{\kappa_2}{2} \frac{\hat{f}''(W)}{\hat{f}(W)}, \quad (10)$$

where  $I$  is the identity function and  $\kappa_2$  is the second moment of the kernel.<sup>5</sup> For the sake of clarity, the detailed derivations and expressions might be found in the Appendix A.2. Clearly, it is possible now to find a range for  $\beta$ -values which would guarantee asymptotic normality; in this case it is  $\beta \in (1/8, 1/4)$ .

There are several other methods of kernel bias reduction. The literature distinguishes *inter alia* among higher order kernels (Granovsky and Mller, 1991), variable bandwidth estimators (Abramson, 1982), variable location estimators (Samiuddin and El-Sayyad, 1990) or parametric transformation methods (Abramson, 1984). Under sufficient smoothness of the underlying density, they all reduce the bias from  $o(\varepsilon^2)$  to  $o(\varepsilon^4)$  as the sample size increases. Although it is likely that they might be also successfully applied in our setting, their properties do not guarantee a clear-cut asymptotic theory for the test statistics. Thereof, we leave this exercise for future consideration.

### 3.1 Bandwidth selection

The optimal bandwidth, denote it by  $\varepsilon^*$ , comes at the smallest MSE of the test statistics,  $T_n^s(\varepsilon)$ . Following the Diks and Panchenko (2006) methodology, this implies that the sum of dominating squared terms in Eq. (7) is minimized, so that under the bandwidth shrinking condition

$$\varepsilon^* = C^* n^{\frac{-2}{2\alpha+\gamma}}. \quad (11)$$

---

<sup>5</sup>We employ the Nadaraya-Watson estimator as a plug-in estimator for sharpening function as suggested by Choi and Hall (1999).

with

$$C^* = \left( \frac{18\gamma q_2}{2\alpha E[s(W)]^2} \right)^{\frac{1}{2\alpha+\gamma}}. \quad (12)$$

One may readily observe that the general formula for optimal bandwidth is parallel to the one derived in the Diks and Panchenko (2006). DS changes the pointwise bias of the estimator density estimator, intuitively affecting both the rate of convergence, i.e.  $\alpha$  parameter, and the leading bias term, i.e.  $s(w_i)$ .

In order to get more insight on the effects of DS on the optimal bandwidth selection in the DP setting, it is worthwhile to test it in a similar environment to the one proposed by Diks and Panchenko (2006). Thereof, we consider here the interdependent ARCH process, however for the sake of presentational purposes, extended to the 3-variate setting, representing the dimensionality problem from the previous section. Imagine the ARCH process without instantaneous dependence

$$\begin{aligned} Q_t &\sim N(0, c + aQ_{t-1}^2) \\ X_t &\sim N(0, c + aY_{t-1}^2) \\ Y_t &\sim N(0, c + aQ_{t-1}^2). \end{aligned} \quad (13)$$

It is clear that the process satisfies the null that  $\{X_t\}$  is not Granger causing  $\{Y_t\}$ , corrected for the presence of  $\{Q_t\}$ . Parameters  $c$  and  $a$  are chosen in order to guarantee stationarity and ergodicity, i.e.  $c > 0$  and  $0 < a < 1$ .

Because of the complexity of the problem, in order to get more insight on the magnitude of the optimal constant  $C^*$ , and optimal bandwidth value  $\varepsilon^*$  thereof, we rely on the Monte Carlo methods. We perform 1000 simulations of process (13) with  $a = 0.4$  and  $c = 1$  for different sample sizes. We extract values for  $q_2$  and  $E[s(W)]$  using standard kernel methods for density and derivative estimation, described in Wand and Jones (1995) and Silverman (1998). The results are presented in Table 1.

Table 1: Optimal constants and bandwidth values for the  $T_n^s(\varepsilon)$  test of the 3-variate process (13) for different sample sizes under the bandwidth shrinking condition. The values represent the mean over 1000 simulations.

sample size (n)	50	100	200	500	1000
$C^*$	0.83	0.89	0.94	0.97	0.98
$\varepsilon^*$	0.43	0.41	0.39	0.34	0.31

Intuitively, reported optimal bandwidths are smaller than those from Diks and Panchenko (2006). This is a straightforward result of the DS method. Given that the sharpened estimate has lower bias, the test does not have to include such a wide range of points in order to yield similar properties. This in fact guarantees asymptotic normality of the sharpened test statistics under smaller bandwidth values.

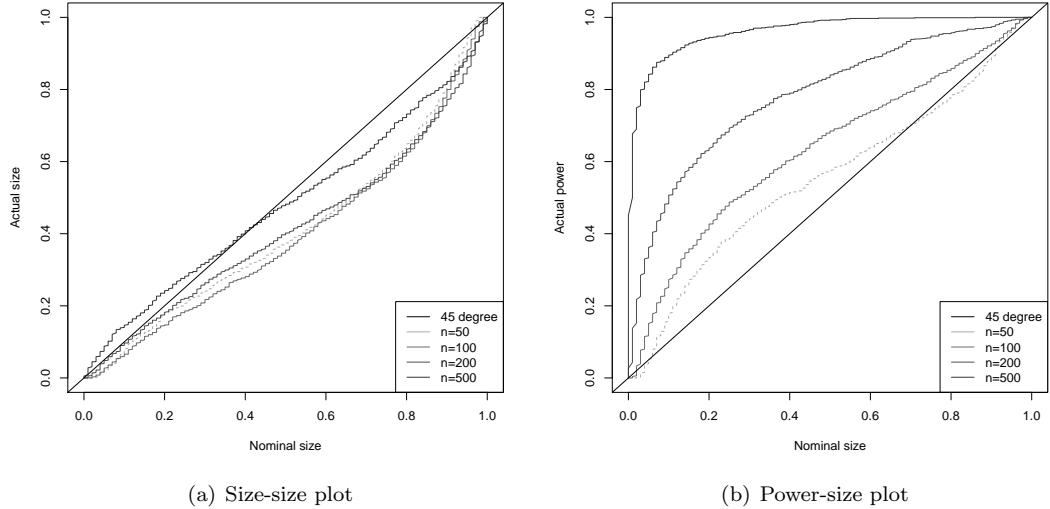


Figure 1: Size-size and power-size plots of the  $T_n^s(\varepsilon)$  test of 3-variate process (13) for different sample sizes under the bandwidth shrinking condition aggregated over 1000 simulations.

### 3.2 Performance of the DS in Granger causality setting

Given the optimal bandwidth values, we may turn to the assessment of the performance of the DS-augmented DP test. Again we rely here on the Monte Carlo methods. Since process (13) matches the basic properties of the observed financial time series (like conditional heteroscedasticity), we use it as an underlying behavior for the simulations for our test size assessment. For the test power assessment we use the same process, however, we switch the causality between  $\{X_t\}$  and  $\{Y_t\}$  so that, even conditioning on  $\{Q_t\}$ , the null hypothesis of no Granger causality is violated.

The results from 1000 simulations for various time series lengths are summarized by the size-size and power-size plots in Fig. 1. One may readily observe that the test demonstrates larger power on larger samples. For 5% significance level, it ranges from 0.05 for  $n = 50$  (no power) to 0.82 for  $n = 500$  (high power). A simple rule of thumb may suggest that the test yields satisfactory results for samples of length 500 and larger. Interestingly, for the same significance levels and sample sizes, the sharpened DP test offers better power than its original counterpart. In fact, the standard DP test yields power of 0.8 for samples of 1000-2000 length.

At the same time test tends to be rather conservative for larger nominal p-values, i.e. it under-rejects when the null is satisfied. However, for relatively smaller significance levels the size-size plot suggests that the larger the sample size, the closer it is to the ideal rejection probability.

One may view DS as an almost ideal tool for bias reduction. We observe, however, a price

for increased precision of the pointwise estimators. For each point in the distribution the algorithm calculates its sharpened form. This in fact shows up as an additional loop in the procedure, increasing the computational time from  $\mathcal{O}(n^2)$  to  $\mathcal{O}(n^3)$ . For relatively short time series it may not seem as a problem but for  $n$  larger than a couple of thousands, computational time might be a bottle neck of the analysis. Therefore, for larger datasets, we recommend using DS together with multicore or cloud computing.

## 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 environment for our hypothesis testing. There is a common agreement among professionals that any causal relation between prices of different crops has to be corrected for the weather conditions at that particular moment (see for instance Popp et al. (2003) and Carreck and Christian (1997)). This conditioning variable suits as a perfect representation of  $Q$  variable, from the 3-variate example in previous sections.

We consider corn, beans and wheat as being most representative of the US grain market. We consider prices of the 1-month ahead rolling future contracts, traded in USD at the Chicago Board of Trade (CBoT). The weather variable is approximated by the rolling monthly futures on Heating Degree Days (HDD), averaged over Philadelphia, New York, Portland, Chicago and Cincinnati. Daily time series comprise the period from 09/01/2010 till 03/06/2013 building together 633 observations. All the data come from Bloomberg.

We take all variables in logs and evaluate their statistical properties to make sure that the time series are stationary. The results are presented in Table 2.

Looking at the raw data, only prices of corn proves to be stationary at the 5% significance level. Therefore, in order to assess Granger causality in the market we consider first differences of all the variables, i.e. log returns.

In the analysis we consider pairwise relations and complete system separately. In the former we take into account the direct relations between two grains only and in the latter we look at the model with all grains included. Since in the system setting,  $Q$  variable is two dimensional, by  $Q_1$  we refer to the conditioning on grain and  $Q_2$  to conditioning on weather.

In order to underpin the results, we relate them with the standard linear Granger causality setting, as proposed by Granger (1969). We also investigate the causality on the VAR-filtered residuals, making sure that discovered causality effects are the results of nonlinearities. We study the explicit role of the weather variable by comparing our results with the original DP test, i.e. without conditioning variable. In the analysis we assume the lag of each conditioning variable to be 1, as suggested by the Bayesian Information Criterion from the VAR specification. Optimal bandwidth values for the original DP test is 1.27 and for the sharpened test 0.33. Before running the tests, we standardize the data by either normal or uniform transformation. The results for the pairwise relations are presented in Tables 3 and 4 and for the complete

Table 2: Unit root tests of the log prices on US grain market in period 09/01/2010 till 03/06/2013 for raw data and for first differences. Test types comprise the Augmented Dickey-Fuller test (ADF) and Phillips-Perron test (PP) as described in Fuller (1995) and Phillips and Perron (1988), respectively. In both tests the null assumes non-stationarity. CV denotes the critical value for a given test specification.

Variable	Test type	Raw data				First diff.	
		Trend	5% CV	Test stat.	Unit root	Test stat.	Unit root
Corn	ADF	no	-2.86	-3.585	no	-24.574	no
	PP	no	-2.86	-3.591	no	-24.568	no
	ADF	yes	-3.41	-3.469	no	-24.611	no
	PP	yes	-3.41	-3.493	no	-24.605	no
Bean	ADF	no	-2.86	-2.666	yes	-24.504	no
	PP	no	-2.86	-2.668	yes	-24.496	no
	ADF	yes	-3.41	-2.564	yes	-24.523	no
	PP	yes	-3.41	-2.575	yes	-24.516	no
Wheat	ADF	no	-2.86	-2.299	yes	-24.905	no
	PP	no	-2.86	-2.288	yes	-24.913	no
	ADF	yes	-3.41	-2.272	yes	-24.890	no
	PP	yes	-3.41	-2.261	yes	-24.898	no
HDD	ADF	no	-2.86	-1.247	yes	-21.707	no
	PP	no	-2.86	-1.547	yes	-22.048	no
	ADF	yes	-3.41	-1.409	yes	-21.690	no
	PP	yes	-3.41	-1.739	yes	-22.032	no

system in Tables 5 and 6.

Table 3: Causality results for the pairwise relations of the log returns on the US grain market, without conditioning on Weather (HDD).  
 $(\ast), (\ast\ast), (\ast\ast\ast)$  denotes p-value statistical significance at 10%, 5% and 1%. Period: 09/01/2010-03/06/2013. Nonlinear tests are performed on standardized data, assuming (N)ormal or (U)iform transformation. Number of lags is  $l_X = l_Y = 1$  from the Bayesian Information Criterion.

Variables		Linear Granger Causality				Nonlinear Granger Causality (N)				Nonlinear Granger Causality (U)				
		X	Y	Raw data	VAR residuals	X→Y	Y→X	X→Y	Y→X	X→Y	Y→X	X→Y	Y→X	Raw data
Corn	Wheat							***	**	***	**	**	***	**
Corn	Beans		*											*
Beans	Wheat													

Table 4: Causality results for the pairwise relations of the log returns on the US grain market, with conditioning on Weather (HDD).  
 $(\ast), (\ast\ast), (\ast\ast\ast)$  denotes p-value statistical significance at 10%, 5% and 1%. Period: 09/01/2010-03/06/2013. Nonlinear tests are performed on standardized data, assuming (N)ormal or (U)iform transformation. Number of lags is  $l_X = l_Y = 1$  from the Bayesian Information Criterion.

Variables		Linear Granger Causality				Nonlinear Granger Causality (N)				Nonlinear Granger Causality (U)				
		X	Y	Raw data	VAR residuals	X→Y	Y→X	X→Y	Y→X	X→Y	Y→X	X→Y	Y→X	Raw data
Corn	Wheat							**	**	***	**	**	***	**
Corn	Beans		*											**
Beans	Wheat													**

Table 5: Causality results for the system setting of the log returns on the US grain market, without conditioning on Weather (HDD).  
 $(\ast), (\ast\ast), (\ast\ast\ast)$  denotes p-value statistical significance at 10%, 5% and 1%. Period: 09/01/2010-03/06/2013. Nonlinear tests are performed on standardized data, assuming (N)ormal or (U)niform transformation. Number of lags is  $l_X = l_Y = l_{Q_1} = 1$  from the Bayesian Information Criterion.

		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
Corn	Wheat					***				**		***	
Corn	Beans	*								**		**	
Beans	Wheat												

Table 6: Causality results for the system setting of the log returns on the US grain market, with conditioning on Weather (HDD).  
 $(\ast), (\ast\ast), (\ast\ast\ast)$  denotes p-value statistical significance at 10%, 5% and 1%. Period: 09/01/2010-03/06/2013. Nonlinear tests are performed on standardized data, assuming (N)ormal or (U)niform transformation. Number of lags is  $l_X = l_Y = l_{Q_1} = 1$  from the Bayesian Information Criterion.

		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
Corn	Wheat					***	*			**	*	**	
Corn	Beans	*				***				**		**	
Beans	Wheat					***				**		**	

One may readily observe that the US grain market does not show much of the linear Granger causality. The only exception is the possible impact of beans on corn prices in all the settings. After VAR filtering this relation disappears, however, as expected.

Interestingly, our results suggest that the relations between US grain prices exhibit lot of nonlinearities. Looking at the basic pairwise setting, there are strong causal linkages between corn and wheat. If we, however, condition on weather forecasts, some of the relations vanish, in particular in the uniform transformation setting. Moreover, we observe that after conditioning, some new causal relations emerge between corn and beans, which additionally are purely nonlinear in nature. It suggests that weather forecasts have a dual role in the grain market. They do not only drive many of the causal relations themselves but they also mask some of the others. From our pairwise results it is clear that weather is masking the corn-beans whereas is driving wheat-corn Granger causality.

In the basic system setting the corn-wheat causal relation is preserved, being significant also after linear filtering. The corn-beans influence is also visible in the uniform transformation setting, confirming the pairwise results. After conditioning on weather forecasts, however, we observe the emergence of the Granger causal relation in the beans-wheat market. Interestingly, the system setting exhibits a lot of regularities from the pairwise study, nevertheless, because of the complexity of the environment, we see new nonlinear relations between all the grains. Prices of corn are Granger causing those of wheat and beans, whereas prices of beans are also influencing those of wheat, conditioning on the weather forecasts.

A straightforward explanation of our results suggests that the nonlinear causal relation emerge from bigger to smaller markets. Corn is most heavily traded grain on the CBoT, followed by beans and wheat. Intuitively, bigger markets should affect those of smaller size as they are deeper and more liquid (Sari et al., 2012). This reasoning is fact in line with our previous finding on the dual role of the weather forecasts in grain market. Since the majority of shocks in the grain market are weather-related, they serve as a common factor and are displayed in all the markets, mitigating the effects of the grain-specific shocks. Correcting for the weather stance allows, therefore, to reveal causal relations between grain-specific shocks, which spread from bigger to smaller markets.

## 5 Conclusions and discussion

This study contributes to the ongoing discussion on the validity of the Granger causality concept, allowing it to be applied in more complex environments. We show that the Granger causality test proposed by Diks and Panchenko (2006) lacks consistency in a multivariate setting. The problems arise as a consequence of a too large pointwise estimator bias, which decreases the precision of the tests statistics and affects its asymptotic properties thereafter. In order to bring back its desirable properties we propose a sharpened form of the test statistics, which under mild regularity conditions is again asymptotically normal. In fact, we confirm that

the sharpening function reduces the original bias of the estimator without any consequences for its further properties, as originally suggested by Hall and Minnotte (2002). We assess properties of the sharpened test numerically, demonstrating that its power is larger than in the basic DP test.

In order to show practical side of our study, we apply the test to the US grain market as, because of its weather-dependent structure, it serves as an ideal environment to assess our methodology. We consider Granger causality between corn, beans and wheat, conditioning on the weather forecasts, approximated by the future contracts on Heating Degree Days. Our results suggests that the US grain market exhibits lot of nonlinear relations. We discover a dual role of the weather forecasts. Firstly, they seem to drive the causal relation from wheat to corn, in the pairwise setting as they serve as a common factor. Secondly, they are masking the causal relations from corn to beans and from beans to wheat in the complete system setting. Correcting for the common factor, we reveal the true nonlinear Granger causal relations in the US grain market, suggesting that the causality spreads from bigger, i.e. deeper and more liquid, to smaller markets.

Our results might have important further implications for the food market analysis. As suggested by Gilbert (2010), future contracts are the major transition channel through which macro variables affect food prices. Understanding possible nonlinear economic dynamics in these markets is therefore of a great significance, as it may prevent possible bubbles and instant food price rises, as the ones observed between 2007 and 2008.

## A Appendix

### A.1 Asymptotic properties of sharpened test statistics (Corollary 1)

We closely follow here the reasoning developed in Diks and Panchenko (2006), however, for the Gaussian kernel and sharpened estimator, as proposed in Hall and Minnotte (2002). We analyze here the case of a random sample as the dependency results follow from the reasoning in Denker and Keller (1983b) and Diks and Panchenko (2006).

By symmetrization with respect to three different indices  $i, j, k$ , for given  $\varepsilon$  the sharpened test statistics (Eq. 3 with sharpened estimators) might be rewritten in the form of the third order U-statistic as

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

with  $W_i = (X_i^{lx}, Y_i^{ly}, Z_i)$ ,  $i = 1, \dots, n$  and sharpened form of the kernel being

$$\begin{aligned}\tilde{K}^s(W_i, W_j, W_k) &= \frac{\varepsilon^{-d_X-2d_Y-d_Z}}{6} \left[ (G_{ik}^{XYZ} G_{ij}^Y - G_{ik}^{XY} G_{ij}^{YZ}) \right. \\ &\quad + (G_{ij}^{XYZ} G_{ik}^Y - G_{ij}^{XY} G_{ik}^{YZ}) + (G_{jk}^{XYZ} G_{ji}^Y - G_{jk}^{XY} G_{ji}^{YZ}) \\ &\quad + (G_{ji}^{XYZ} G_{jk}^Y - G_{ji}^{XY} G_{jk}^{YZ}) + (G_{ki}^{XYZ} G_{kj}^Y - G_{ki}^{XY} G_{kj}^{YZ}) \\ &\quad \left. + (G_{kj}^{XYZ} G_{ki}^Y - G_{kj}^{XY} G_{ki}^{YZ}) \right].\end{aligned}$$

where  $G_{i,j}^W$  is the sharpened form of the multivariate kernel density, i.e.

$$G_{i,j}^W = K_{multi} \left( \frac{W_i - \psi_p(W_j)}{\varepsilon} \right).$$

We assume that the density is smooth enough and infinitely differentiable so that it is possible to find any sharpening function which would guarantee bias reduction of order  $p$ , i.e.

$$E_{W_j}[G_i^W] - G_i^W = \varepsilon^p R_p(W_i) + o(\varepsilon^p),$$

where  $R_p(W_i)$  is the leading bias term associated with  $\varepsilon^p$  evaluated at point  $W_i$ .

Let us define  $\tilde{K}_1^s$  and  $\tilde{K}_2^s$  as in Conditions in Eq. 4-6. The bias of products of estimated densities, i.e.  $s(w_i)$ , follows from the properties of the local estimator bias (see the previous section) and identities such as  $E[\hat{f}_V \hat{f}_W] = E[(f_V + (\hat{f}_V - f_V))(f_W + (\hat{f}_W - f_W))] = f_V f_W + f_V E[\hat{f}_W - f_W] + f_W E[\hat{f}_V - f_V] + o(\varepsilon^p)$ . Therefore, the local bias of the  $T_n^s(\varepsilon)$  might be rewritten as proportional to

$$\begin{aligned}\tilde{K}_1^s(w_i, \varepsilon) - \lim_{\varepsilon \rightarrow 0} \tilde{K}_1^s(w_i, \varepsilon) &= \varepsilon^p (f_Y(y_i) R_p(x_i, y_i, z_i) - f_{X,Y}(x_i, y_i) R_p(y_i, z_i) \\ &\quad + f_{X,Y,Z}(x_i, y_i, z_i) R_p(y_i) - f_{Y,Z}(y_i, z_i) R_p(x_i, y_i)) + o(\varepsilon^p).\end{aligned}$$

Taking into account Condition in Eq. 4, one may find that it holds with  $\alpha = p$  and  $s(w_i)$  being equal the term in the brackets.

Looking at the Condition in Eq. 5, taking the expectations over  $W_k$  for each of the contributions to the kernel function  $\tilde{K}^s$ , one finds that the dominant terms are proportional to  $\varepsilon^{-d_X-2d_Y-d_Z} G_{ij}^{XYZ} G_{ik}^Y$  and  $\varepsilon^{-d_X-2d_Y-d_Z} G_{ji}^{XYZ} G_{jk}^Y$ , for which we have

$$E_{W_k} [\varepsilon^{-d_X-2d_Y-d_Z} G_{ij}^{XYZ} G_{ik}^Y] = \varepsilon^{-d_X-d_Y-d_Z} G_{ij}^{XYZ} f_Y(Y_i) + o(\varepsilon^{-d_X-d_Y-d_Z}),$$

and

$$E_{W_k} [\varepsilon^{-d_X-2d_Y-d_Z} G_{ji}^{XYZ} G_{jk}^Y] = \varepsilon^{-d_X-d_Y-d_Z} G_{ji}^{XYZ} f_Y(Y_j) + o(\varepsilon^{-d_X-d_Y-d_Z}).$$

Since all the terms are vanishing  $\varepsilon$  slower, we can rewrite that  $E \left[ \left( \tilde{K}_2^s(w_i, w_j) \right)^2 \right]$  is equal

$$\begin{aligned} & \frac{1}{36} E \left[ (2\varepsilon^{-d_X-d_Y-d_Z} E_{W_k} [G_{ij}^{XYZ} G_{ik}^Y])^2 \right] + o(\varepsilon^{-d_X-d_Y-d_Z}) \\ &= \frac{4\varepsilon^{-2d_X-2d_Y-2d_Z}}{36} E \left[ (G_{ij}^{XYZ} f_Y(Y_i))^2 \right] + o(\varepsilon^{-d_X-d_Y-d_Z}) \\ &= \frac{4\varepsilon^{-d_X-d_Y-d_Z}}{36} E \left[ (f_{X,Y,Z}(X_i, Y_i, Z_i)) f_Y(Y_i)^2 \right] + o(\varepsilon^{-d_X-d_Y-d_Z}), \end{aligned}$$

where we exploited the fact that  $G_{ij}^{XYZ} G_{ik}^Y G_{ji}^{XYZ} G_{jk}^Y$  are asymptotically perfectly correlated as  $\varepsilon$  tends to 0 sufficiently slowly as  $n \rightarrow \infty$ . This confirms that the  $\gamma$  parameter from the original Diks and Panchenko (2006) methodology is unaffected by the DS, being equal  $d_X + d_Y + d_Z$  with

$$q_2 = \frac{4}{36} E \left[ (f_{X,Y,Z}(X_i, Y_i, Z_i)) f_Y(Y_i)^2 \right].$$

Since the variance of  $\tilde{K}^s$  is limited by  $\varepsilon^{d_X+2d_Y+d_Z}$  as the sample size increases, condition from Eq. 6 holds for  $\delta = d_X + 2d_Y + d_Z$ , again being the same as in Diks and Panchenko (2006). This brings us to the conclusion that DS decreased the local bias of  $T_n^s(\varepsilon)$  only, leaving the further properties of the test statistics unchanged.

## A.2 Application of bias reduction

For practical purposes, let us assume that the  $H = \text{diag}(\varepsilon, \varepsilon, \dots, \varepsilon)$  is a  $d_W \times d_W$  bandwidth matrix so that the local density estimator of  $d_W$ -variate random vector from Eq. (2) becomes

$$\hat{f}_W(W_i) = \frac{\varepsilon^{-d_W}}{n-1} \sum_{j:j \neq i} K_H(W_i - W_j), \quad (14)$$

where  $K_H(W_i - W_j) = K(H^{-1}(W_i - W_j)) = K_{\text{multi}}((W_i - W_j)/\varepsilon)$ .

Assume also that the density function is infinitely differentiable and let  $f'$  be the vector of first-order partial derivatives of  $f$ ,  $f''$  be the matrix of second-order partial derivatives of  $f$ ,  $f^{(3)}$  be the cube of third-order partial derivatives of  $f$ ,  $f^{(4)}$  be the 4 dimensional matrix of fourth-order partial derivatives of  $f$  etc, with all the entries being piecewise continuous and square integrable. For presentational purposes, let us also use  $\int$  as a shorthand for  $\int \cdots \int_{\mathbb{R}^{d_W}}$  and  $dW$  as a shorthand for  $dW_1 \cdots dW_{d_W}$ . By  $I_{d_W}$  we denote also the  $d_W \times d_W$  identity matrix.

Let us consider the case study example from the paper, where we extend the basic analysis to the 3-variate causality testing, i.e.  $Y_{t+1}|(X_t^{l_X}; Y_t^{l_Y}, Q_t^{l_Q}) \sim Y_{t+1}|(Y_t^{l_Y}, Q_t^{l_Q})$ . As it is shown in the text, the standard DP test lacks consistency because of the too large pointwise estimator bias. The original bias of the standard kernel density estimator at point  $W_i$  might be computed

from the second order Taylor expansion around the estimation point (Wand and Jones, 1995)

$$\begin{aligned}
E \left[ \hat{f}_W(W_i) \right] &= \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} E [K_H(W_i - W_j)] = \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} \int_{-\infty}^{\infty} K_H(W_i - W) f(W) dW \\
&= \varepsilon^{-d_W} \int_{-\infty}^{\infty} K_H(W_i - W) f(W) dW = \int_{-\infty}^{\infty} K_{multi}(s) f(W - \varepsilon s) ds \\
&= \int_{-\infty}^{\infty} K_{multi}(s) \left\{ f(W_i) - \varepsilon s^T f'(W_i) + \frac{\varepsilon^2}{2} s^T f''(W_i) s + o(\varepsilon^4) \right\} ds \\
&= f(W_i) + \frac{\varepsilon^2}{2} \kappa_2 \text{tr} \{ f''(W_i) \} + o(\varepsilon^4) = f(W_i) + o(\varepsilon^2),
\end{aligned} \tag{15}$$

where we exploited the fact that  $\int s K_{multi}(s) ds = 0_{d_W}$  and  $\int s s^T K_{multi}(s) ds = \kappa_2 I_{d_W}$ .

The dominant term in the local estimator bias ( $R_2$ ) is driven by  $1/2\kappa_2 \text{tr} \{ f''(W_i) \}$ , which is of order  $o(\varepsilon^2)$ . The idea of DS is to eliminate this term by applying appropriate sharpening function. It can be best illustrated by calculating the expected value of the sharpened estimator with the DS function given by Eq. 10

$$\begin{aligned}
E \left[ \hat{f}_W^s(W_i) \right] &= \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} E [K_H(W_i - \psi_4(W_j))] = \frac{\varepsilon^{-d_W}}{n-1} \sum_{j,j \neq i} \int_{-\infty}^{\infty} K_H(W_i - \psi_4(W_j)) dF(W) \\
&= \varepsilon_n^{-d_W} \int_{-\infty}^{\infty} K_H(W_i - \psi_4(W)) dF(W) = \varepsilon_n^{-d_W} \int_{-\infty}^{\infty} K_H(W_i - V) dF(\psi_4^{-1}(V)) \\
&= \varepsilon_n^{-d_W} \int_{-\infty}^{\infty} K_H(W_i - V) f(\psi_4^{-1}(V)) \left| \frac{\partial \psi_4^{-1}(V)}{\partial V} \right| dV \\
&= \varepsilon_n^{-d_W} \int_{-\infty}^{\infty} K_H(W_i - V) \left\{ f(V) - \frac{\varepsilon^2}{2} \kappa_2 \text{tr} \{ f''(V) \} + \frac{\varepsilon^4}{4} k_2^2 U(V) + o(\varepsilon^6) \right\} dV \\
&= \int_{-\infty}^{\infty} K_{multi}(s) \left\{ f(W_i) - \varepsilon s^T f'(W_i) + \frac{\varepsilon^2}{2} s^T f''(W_i) s - \frac{\varepsilon^2}{2} \kappa_2 \text{tr} \{ f''(W_i) \} \right. \\
&\quad \left. + \frac{\varepsilon^3}{2} \kappa_2 \text{tr} \{ s^T f^{(3)}(W_i) \} - \frac{\varepsilon^4}{4} \kappa_2 \text{tr} \{ s^T f^{(4)}(W_i) s \} + \frac{\varepsilon^4}{4} k_2^2 U(W_i) \right\} ds + o(\varepsilon^6) \\
&= f(W_i) + \varepsilon^4 R_4(W_i) + o(\varepsilon^6),
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
U(V) &= \frac{f'(V)^T f'(V) f'(V)^T f'(V)}{f(V)^3} - \frac{5 f'(V)^T f''(V) f'(V)}{2 f(V)^2} \\
&\quad - \frac{2 f'(V)^T (B_1(f''(V)) - f''(V)) f'(V)}{2 f(V)^2} + \frac{\text{tr} \{ f''(V)^T f''(V) \} - \sum |B_2(f''(V))|}{f(V)} \\
&\quad + \frac{\text{tr} \{ f'(V)^T f^{(3)}(V) \}}{f(V)},
\end{aligned} \tag{17}$$

and

$$R_4(W_i) = \frac{1}{4} \left( \kappa_2^2 U(W_i) - \kappa_4 \text{tr}\{f^{(4)}(W_i)\} \right). \quad (18)$$

Matrix transformation  $B_1(\cdot)$  puts the trace of the argument on each of the diagonal entries and  $B_2(\cdot)$  takes 2x2 submatrix around the diagonal of the argument.

Clearly, the original bias of order  $o(\varepsilon^2)$  has decreased to the order  $o(\varepsilon^4)$  without any effect on the kernel function  $K_{multi}$ , leaving the further properties of Eq. (3) the same as in the original reasoning from Diks and Panchenko (2006). Therefore one may calculate optimal bandwidth values, which endow the test statistics with asymptotic normality, for the 3-variate setting from Eq. 11 by plugging in the estimates for  $s(w_i)$  and  $q_2$ ,

$$\begin{aligned} s(w_i) &= f_{Y,Q}(y_i, Q_i) R_4(x_i, y_i, z_i, q_i) - f_{X,Y,Q}(x_i, y_i, q_i) R_4(y_i, z_i, q_i) \\ &\quad + f_{X,Y,Z,Q}(x_i, y_i, z_i, q_i) R_4(y_i, q_i) - f_{Y,Z,Q}(y_i, z_i, q_i) R_4(x_i, y_i, q_i), \end{aligned}$$

and

$$q_2 = \frac{4}{36} E \left[ (f_{X,Y,Z,Q}(X_i, Y_i, Z_i, Q_i)) f_{Y,Q}(Y_i, Q_i)^2 \right].$$

In fact, bias reduction from  $o(\varepsilon^2)$  to  $o(\varepsilon^4)$  allows to include up to 4 additional variables. Any additional conditioning variable would again violate the consistency of the test, requiring more appropriate sharpening function.

## References

- I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, 10(4):pp. 1217–1223, 1982.
- I. S. Abramson. Adaptive density flattening—a metric distortion principle for combating bias in nearest neighbor methods. *The Annals of Statistics*, 12(3):pp. 880–886, 1984.
- Z. Bai, W.-K. Wong, and B. Zhang. Multivariate linear and nonlinear causality tests. *Mathematics and Computers in Simulation*, 81(1):5 – 17, 2010.
- N. Carreck and D. Christian. A study of grain dormancy and viability in spring barley. *European Journal of Agronomy*, 6(34):155 – 161, 1997.
- N. Cartwright. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, Cambridge, 2007.
- E. Choi and P. Hall. Data sharpening as a prelude to density estimation. *Biometrika*, 86(4):941–947, 1999.
- M. Denker and G. Keller. On u-statistics and v. Mises statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64(4):505–522, 1983a.

- M. Denker and G. Keller. On  $U$ -statistics and v. Mises' statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64:505–522, 1983b.
- C. Diks and V. Panchenko. A note on the Hiemstra-Jones test for Granger non-causality. *Studies in Nonlinear Dynamics and Econometrics*, 9, 2005. art. 4.
- C. Diks and V. Panchenko. A new statistic and practical guidelines for nonparametric granger causality testing. *Journal of Economic Dynamics and Control*, 30:1647–1669, 2006.
- W. A. Fuller. *Introduction to Statistical Time Series (Wiley Series in Probability and Mathematical Statistics)*. Wiley-Interscience, 1995.
- C. L. Gilbert. How to understand high food prices. *Journal of Agricultural Economics*, 61(2):398–425, 2010.
- C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.
- B. L. Granovsky and H.-G. Müller. Optimizing kernel methods: A unifying variational principle. *International Statistical Review / Revue Internationale de Statistique*, 59(3):pp. 373–388, 1991.
- S. Guo, C. Ladroue, and J. Feng. Granger causality: Theory and applications. In J. Feng, W. Fu, and F. Sun, editors, *Frontiers in Computational and Systems Biology*, volume 15 of *Computational Biology*, pages 83–111. Springer London, 2010.
- P. Hall and M. C. Minnotte. High order data sharpening for density estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(1):141–157, 2002.
- C. Hiemstra and J. D. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *Journal of Finance*, 49(5):1639–1664, 1994.
- P. C. B. Phillips and P. Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2):pp. 335–346, 1988.
- M. P. Popp, C. R. Dillon, and T. C. Keisling. Economic and weather influences on soybean planting strategies on heavy soils. *Agricultural Systems*, 76(3):969 – 984, 2003.
- J. L. Powell and T. M. Stoker. Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics*, 75:219–316, 1996.
- M. Samiuddin and G. M. El-Sayyad. On nonparametric kernel density estimates. *Biometrika*, 77(4):pp. 865–874, 1990.
- R. Sari, S. Hammoudeh, C.-L. Chang, and M. McAleer. Causality between market liquidity and depth for energy and grains. *Energy Economics*, 34(5):1683 – 1692, 2012.

- D. W. Scott. *Multivariate Density Estimation (Wiley Series in Probability and Mathematical Statistics)*. Wiley-Interscience, 1992.
- B. W. Silverman. *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1998.
- M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1995.