Chapter 6
Can Graphical Causal Inference Be Extended to Nonlinear Settings?

An Assessment of Conditional Independence Tests

Nadine Chlaß and Alessio Moneta

6.1 Introduction

Graphical models are a powerful tool for causal model specification. Besides allowing for a hierarchical representation of variable interactions, they do not require any a priori specification of the functional dependence between variables. The construction of such graphs hence often relies on the mere testing of whether or not model variables are marginally or conditionally independent. The identification of causal relationships then solely requires some general assumptions on the relation between stochastic and causal independence, such as the Causal Markov Condition and the Faithfulness Condition (Spirtes et al. 2000; Pearl 2000). However, a procedure would require further assumptions to hold. Namely those the independence tests themselves are based on.

In continuous settings, Spirtes et al. (2000) suggest causal inference based on a very restrictive formulation of independence, that is, vanishing partial correlations. Such a measure does, however, limit the applicability of causal inference to linear systems. This constitutes a serious drawback especially for the social sciences where an a priori specification of the functional form proves difficult or at odds with linearity. In short: graphical models theoretically reduce specification uncertainty regarding functional dependence, but their implementation in practice deprives them of this virtue.

In this paper we investigate how causal structures in continuous settings can be identified when both functional forms and probability distributions of the variables remain unspecified. We focus on tests exploiting the fact that if X and Y are conditionally independent given a set of variables Z, the two conditional densities $f(X|Y, Z)$ and $f(X|Z)$ must coincide. We start by estimating the conditional densities $f(X|Y, Z)$ and $f(X|Z)$ via nonparametric techniques (kernel methods). We proceed by testing if some metric expressing the distance between these very conditional densities is sufficiently close to zero. Out of several metrics available in the

N. Chlaß and A. Moneta (✉)
Max Planck Institute of Economics, Jena, Germany
e-mail: chlass@econ.mpg.de; moneta@econ.mpg.de

M. Suárez et al. (eds.), *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, DOI 10.1007/978-90-481-3263-8_6,

literature to express such distance we choose two, the Euclidean, and the Hellinger distance. We investigate in a Monte Carlo study how different tests involving either measure are able to detect statistical independence, conditioned on a small set of variables.

One limitation may result from nonparametric density estimation being subject to the curse of dimensionality. As the number of variables increases, the estimated empirical density converges at a slower rate to its population value. To compensate this drawback we use a local bootstrap procedure which consists of resampling the data for each test. While local bootstrap strongly increases the computational time of the test, it succeeds in counterbalancing the curse of dimensionality. Section 6.2 presents the statistical methods used in detail. Section 6.3 describes the simulation design and our results. Section 6.4 concludes.

6.2 Nonparametric Tests for Conditional Independence

We want to test the following null hypothesis: *X is independent of Y given Z*, that is

$$X \perp\!\!\!\perp Y \mid Z, \tag{6.1}$$

where X and Y are continuous random variables, and Z is a (possibly empty) vector of d continuous random variables $(Z_1, \ldots, Z_d)$. We observe n random realizations $(X_t, Y_t, Z_t),\ t = 1, \ldots, n$.

Note that Fisher's z statistic proposed by Spirtes et al. (2000: 94) to test conditional independence relations in continuous settings, and also incorporated in Tetrad (Scheines et al. 1996), requires normality of the joint probability distribution $f(X, Y, Z)$. The latter is guaranteed by the linearity assumption if the error terms are also normal.

We propose a class of tests based on the estimation and comparison of the following two multivariate distribution $h_1(\cdot)$ and $h_2(\cdot)$:

$$\begin{aligned} h_1(X, Y, Z) &\equiv f(X, Y, Z) f(Z) \\ h_2(X, Y, Z) &\equiv f(X, Z) f(Y, Z). \end{aligned} \tag{6.2}$$

This type of tests exploits the fact that under the null hypothesis:

$$f(X|Y, Z) = f(X|Z),$$

whenever $f(Y, Z)$ and $f(Z) > 0$. Hence, by definition of a conditional density function:

$$\frac{f(X, Y, Z)}{f(Y, Z)} = \frac{f(X, Z)}{f(Z)}.$$

It follows that under the null hypothesis:

$$h_1(\cdot) = h_2(\cdot). \tag{6.3}$$

We estimate h_1 and h_2 using a kernel smoothing approach (see Wand and Jones 1995: Chapter 4). Both h_1 and h_2 are of length $m = d + 2$. In particular, we use the so-called *product kernel* estimators:

$$\begin{aligned}
&\hat{h}_1(x, y, z; b) \\
&\quad = \frac{1}{N^2 b^{m+d}} \left\{ \sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) K\left(\frac{Y_i - y}{b}\right) K\left(\frac{Z_i - z}{b}\right) \right\} \left\{ \sum_{i=1}^{n} K_p\left(\frac{Z_i - z}{b}\right) \right\} \\
&\hat{h}_2(x, y, z; b) \\
&\quad = \frac{1}{N^2 b^{m+d}} \left\{ \sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right) K_Z\left(\frac{Z_i - z}{b}\right) \right\} \left\{ \sum_{i=1}^{n} K\left(\frac{Y_i - y}{b}\right) K_p\left(\frac{Z_i - z}{b}\right) \right\},
\end{aligned} \tag{6.4}$$

where K denotes the *kernel* function, b indicates a scalar bandwidth parameter, and K_p represents a product kernel, i.e., $K_p((Z_i - z)/b) = \prod_{j=1}^{d} K((Z_{j_i} - z_j)/b)$. For our simulations (see next section) we choose the kernel: $K(u) = (3 - u^2)\phi(u)/2$, with $\phi(u)$ the standard normal probability density function. We use a "rule-of-thumb" bandwidth: $b = n^{-1/8.5}$.

Having obtained h_1 and h_2, we test the null hypothesis (6.1) by verifying whether $\hat{h}_1(\cdot)$ and $\hat{h}_2(\cdot)$ are sufficiently similar. There are several ways to measure distance between two products of estimated density functions (see Su and White 2008). Here, we focus on the following ones:

(i) Weighted Hellinger distance proposed by Su and White (2008). In this case the distance is:

$$d_H = \frac{1}{n} \sum_{t=1}^{n} \left\{ 1 - \sqrt{\frac{h_2(X_t, Y_t, Z_t)}{h_1(X_t, Y_t, Z_t)}} \right\}^2 a(X_t, Y_t, Z_t), \tag{6.5}$$

where $a(\cdot)$ is a nonnegative weighting function. The weighting function $a(\cdot)$, as well as the resulting test statistics are specified in Su and White (2008).

(ii) Euclidean distance as proposed by Szekely and Rizzo (2004) in their "energy test." In this case, we have:

$$d_E = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{1_i} - h_{2_j}|| - \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{1_i} - h_{1_j}|| - \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} ||h_{2_i} - h_{2_j}||, \tag{6.6}$$

where $h_{1_i} = h_1(X_i, Y_i, Z_i)$, $h_{2_i} = h_2(X_i, Y_i, Z_i)$, and $|| \cdot ||$ is the Euclidean norm.

(iii) Euclidean distance as proposed by Baringhaus and Franz (2004) in their "Cramer test." There is no substantial difference with (ii) in the distance proposed, which is $d_E/2$. There is only some difference in the method to obtain the critical values (see Baringhaus and Franz 2004).

When Z is empty we obtain p values using (ii) and (iii) as implemented in the R packages `energy` and `cramer` respectively. The Hellinger distance test cannot be used in this case, since it has been designed for Z non-empty.

When Z is non-empty we obtain p-values for (i), (ii), and (iii) using a local bootstrap procedure, as described in Su and White (2008: 840, 841) and Paparoditis and Politis (2000: 144, 145). Local bootstrap imposes the null hypothesis in the resampling scheme and counts how many times the bootstrap statistic is larger than the statistic calculated on the basis of the real data. More specifically, local bootstrap proceeds as follows: (1) Draw a bootstrap sampling Z_t^* (for $t = 1, \ldots, n$) from the estimated kernel density $\hat{f}(z) = n^{-1} b^{-d} \sum_{t=1}^{n} K_p((Z_t - z)/b)$. (2) For $t = 1, \ldots, n$, given Z_t^*, draw X_t^* and Y_t^* *independently* from the estimated kernel density $\hat{f}(x|Z_t^*)$ and $\hat{f}(y|Z_t^*)$ respectively. These functions are defined as follows:

$$\hat{f}(x|Z_t^*) = \frac{\sum_{s=1}^{n} K\left(\frac{X_s - x}{b}\right) K_p\left(\frac{Z_s - Z_t^*}{b}\right)}{b \, \sum_{r=1}^{n} K_p\left(\frac{Z_r - Z_t^*}{b}\right)}$$

$$\hat{f}(y|Z_t^*) = \frac{\sum_{s=1}^{n} K\left(\frac{Y_s - y}{b}\right) K_p\left(\frac{Z_s - Z_t^*}{b}\right)}{b \, \sum_{r=1}^{n} K_p\left(\frac{Z_r - Z_t^*}{b}\right)}$$

(3) Using X_t^*, Y_t^*, and Z_t^* compute the bootstrap statistic S_n^* using one of the distances defined above. (4) Repeat steps (1) and (2) I times to obtain I statistics $\{S_{ni}^*\}_{i=1}^{I}$. (5) The p-value is then obtained by:

$$p \equiv \frac{\sum_{i=1}^{I} 1\{S_{ni}^* > S_n\}}{I},$$

where S_n is the statistic obtained from the original data using one of the distances defined above, and $1\{\cdot\}$ denotes an indicator function taking value one if the expression between brackets is true and zero otherwise.

6.3 Monte Carlo Study

6.3.1 Simulation Design

To compare the aforementioned test procedures we assess their performance in both *size* and *power*. To identify size properties, the hypothesis H_0 of independence or conditional independence must hold everywhere. Data generating processes (DGPs) for which H_0 is true are named size-DGPs. The null hypothesis ($H_0 : X \perp\!\!\!\perp Y|Z$) may apply for three reasons. Either (i) there is no connection at all between X

Fig. 6.1 DAG 1

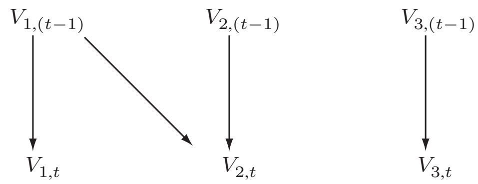

Fig. 6.2 DAG 2

and Y, or (ii) there exists a causal relation between X and Y but only via a set of variables Z, or still (iii) Z constitutes a common cause for X and Y in absence of any other connection besides (ii). In these latter two cases Z is said to *screen off* X from Y (Reichenbach 1956). To illustrate, let us represent the DGP via a Directed Acyclic Graph (DAG).[1] In absence of any causal relation between X and Y the corresponding DAG does not contain any edge or path between X and Y. In case of screening-off, there is a path connecting X and Y via variables Z. Take, for instance, the DAG represented in Fig. 6.1. Here, V_2 screens off (or *d-separates*) V_1 from V_3 such that $V_1 \perp\!\!\!\perp V_3|V_2$. Analogously, $V_2 \perp\!\!\!\perp V_4|V_3$ and $V_1 \perp\!\!\!\perp V_4|V_2, V_3$.

While for size DGPs $H_0 : X \perp\!\!\!\perp Y|Z$ holds everywhere, Z may obviously form causal relations with X and/or Y. These causal relations may take on different functional forms and represent the touchstone for the testing procedures emphasized in this work. To systematically vary nonlinearity and its impact we characterize the causal relation between, say, z_1 and y, in a polynomial form, i.e., via $y = f(z_1)+e$, where $f = \sum_{j=0}^{p} b_j z_1^j$. Herein j would reflect the degree of nonlinearity while b_j would capture the impact nonlinearity exerts. For polynomials of any degree only $b_p \neq 0$. An additive error term e completes the specification. In case of $p = 1$ we also examine the impact of an error entering the causal relation in a multiplicative manner, i.e., $y = b_1 z_1 \cdot e$.

Besides varying the functional form we distinguish an i.i.d. and a time-series case. The latter proves interesting since kernel smoothers generally show notoriously little sensitivity to i.i.d. violations (Welsh et al. 2002). Hence, the alternative procedures put forth before may not be subject to the usual overrejection of H_0 entailed by non-i.i.d. structures (Chlaß and Krüger 2007). For the i.i.d. case, realizations $\{X_t, Y_t, Z_t\}_{t=1}^n$ are generated from a (serially) independent and identical distribution, i.e., $corr(X_t, X_s) = 0$ for $s \neq t$. For the time-series case, each element of $\{X_t, Y_t, Z_t\}_{t=1}^n$ follows an AR(1) process with coefficient $a_1 = 0.5$ and error term $e_t \sim N(0, 1)$, i.e., $X_t = a_1 X_{t-1} + e_{X,t}$. For an illustration, take the DAG displayed in Fig. 6.2 representing such a time series case. Here, $V_{1,t} \perp\!\!\!\perp V_{2,t}|V_{1,(t-1)}$, since $V_{1,(t-1)}$ d-separates $V_{1,t}$ from $V_{2,t}$, while $V_{2,t} \perp\!\!\!\perp V_{3,s}$, for any t and s.

Within the i.i.d. and AR(1) scenarios we vary the number of variables that may establish conditional independence between X_t and Y_t. Either zero, one, but

[1] For definition and properties of DAGs see Spirtes et al. (2000: Chapter 2).

Table 6.1 Simulated cases for size properties

		No causal relations	Causal relations with screening-off $p = 1$	$p = 2$	$p = 3$
$\{X, Y, Z\} \sim$ i.i.d.	$\#Z = 0$	S0.1			
	$\#Z = 1$	S1.1	S1.2 S1.5*	S1.3	S1.4
	$\#Z = 2$	S2.1	S2.2 S2.5*	S2.3	S2.4
$\{X, Y, Z\} \sim AR(1)$	$\#Z = 0$	S0.2			
	$\#Z = 1$	S1.6		S1.7	
	$\#Z = 2$	S2.6		S2.7	

Note: *Non-additive errors.

Table 6.2 Simulated cases for power properties

		$p = 1$	$p = 2$	$p = 3$
$\{X, Y, Z\} \sim$ i.i.d.	$\#Z = 0$	P0.1[1] P0.2[2] P0.7*	P0.3[1] P0.4[2]	P0.5[1] P0.6[2]
	$\#Z = 1$	P1.1 P1.4*	P1.2 P1.5	P1.3
	$\#Z = 2$	P2.1 P2.5*	P2.2 P2.3	P2.4
$\{X, Y, Z\} \sim AR(1)$	$\#Z = 0$	P1.8	P1.9	
	$\#Z = 1$		P1.6	
	$\#Z = 2$		P2.6	

Note: *Non-additive errors; $^1b_p = 0.4$, $^2b_p = 0.8$.

maximally two variables may form the set $Z = \{Z_1, \ldots, Z_d\}$ of conditioned variables; hence Z has cardinality $\#Z = \{0, 1, 2\}$. Table 6.1 reviews all cases for which size properties are investigated.

Power properties of the tests proposed were assessed using DGPs such that H_0 does not hold anywhere, i.e., $X \not\perp\!\!\!\perp Y | Z$. The latter is guaranteed by either (i) a direct path between X and Y which does not include Z, (ii) a common cause for X and Y which is not an element of Z or (iii) a "collider" between X and Y belonging to Z.[2] As before, we vary the functional form f of these causal paths polynomially. In a very stylized manner we design three further phenomena that often arise jointly with nonlinearity. First, we investigate the impact of a non-additive, i.e., multiplicative error term when $b_{j=p} = 0.5$ and $p = 1$. Second, we relinquish the i.i.d. assumption as before and induce X_t and Y_t as two time series of the aforementioned AR(1)-structure. X_t now furthermore depends on Y_t while the functional, i.e., polynomial form of this dependence writes either $\{b_{j \neq 1} = 0, b_1 = 0.5, p = 1\}$ or $\{b_{j \neq 2} = 0, b_2 = 0.5, p = 2\}$. Third, we investigate different cardinalities $\#Z = \{0, 1, 2\}$ of the set of variables that establishes conditional independence between X_t and Y_t. The latter is done to challenge nonparametric procedures in higher dimensional settings where they are known to weakly perform.[3] Table 6.2 reviews all cases for which size properties are investigated.

[2] An example for a collider is displayed in Fig. 6.2: $V_{2,t}$ forms a collider between $V_{1,(t-1)}$ and $V_{2,(t-1)}$. In this case $V_{1,(t-1)} \not\perp\!\!\!\perp V_{2,(t-1)} | V_{2,t}$ although $V_{1,(t-1)} \perp\!\!\!\perp V_{2,(t-1)}$.

[3] For an introduction to the so-called curse of dimensionality see, e.g., Yatchew (2003, p. 676).

6.4 Results

Table 6.3 reports our simulation results for the case where Z is empty ($\#Z = 0$). Hence, Y and X are marginally independent. Rejection frequencies are reported for three different tests, both at the 0.05 and 0.1 level of significance. Take the first line depicting the case S0.1. Here, X and Y were generated 1,000 times from two independent white noise processes. We find a proportion of rejections that is 0.048 at the 0.05 confidence level and 0.096 at the 0.1 confidence level. In other words, for 48 simulation runs out of 1,000 the p-value was greater than 0.05 and for 96 simulation runs out of 1,000 the p-value was greater than 0.1. The Energy test behaves quite well for all cases, since it tends not to reject H_0 when it holds (size DGPs) and it tends to reject H_0 when it is violated (power DGPs). Only when X linearly depends upon Y with a low coefficient (case P0.1), the rejection frequency is not as high as in the other cases. The Cramer test does not produce correct results for: P0.1, P0.3, P0.5, P0.8, P0.9. Let us compare these two nonparametric tests with the Fisher test proposed by Spirtes et al. (2000). We find this test to perform well in some nonlinear cases (P0.2, P0.5, P0.6). However, the percentage of rejection is too high for independent time series (S0.2) and too low for several forms of nonlinear dependence (P0.3, P0.4, P0.7, P0.9). To summarize the case without conditioned variables, the Energy test outperforms both the Cramer and the Fisher test.

Results for the one-conditioned-variable case (Z consisting of one variable) are reported in Table 6.4. Here, for each simulated realization of the process, we apply the local bootstrap procedure described in Section 6.2. To save computation time, we lower the number of iteration to 200. We assess the nonparametric tests described in Section 6.2 and compare them with the parametric Fisher's z. The label "Euclid" comprises Energy and Cramer tests based on the Euclidean distance

Table 6.3 Proportion of rejection of H_0 (no conditioned variables)

	Energy	Cramer	Fisher	Energy	Cramer	Fisher
	Level of significance 5%			*Level of significance* 10%		
Size DGPs						
S0.1 (ind. white noises)	0.048	0.000	0.046	0.096	0.000	0.096
S0.2 (ind. time series)	0.065	0.000	0.151	0.122	0.000	0.213
Power DGPs						
P0.1 (linear, coefficient = 0.4)	0.675	0.024	0.972	0.781	0.047	0.988
P0.2 (linear, coefficient = 0.8)	0.999	0.663	1	1	0.821	1
P0.3 (quadratic, coef. = 0.4)	0.855	0.023	0.165	0.897	0.093	0.240
P0.4 (quadratic, coef. = 0.8)	0.999	0.598	0.282	1	0.790	0.383
P0.5 (cubic, coefficient = 0.4)	0.865	0.025	1	0.915	0.105	1
P0.6 (cubic, coefficient = 0.8)	1	0.605	1	1	0.805	1
P0.7 (non-additive, coef. = 0.5)	1	0.969	0.279	1	0.996	0.376
P0.8 (time series linear)	0.959	0.308	0.999	0.981	0.462	1
P0.9 (time series non-linear)	0.986	0.255	0.432	0.997	0.452	0.521

Note: Length series (n) = 100; number of iterations = 1,000.

Table 6.4 Proportion of rejection of H_0 (one conditioned variable)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	Level of significance 5%			*Level of significance* 10%		
Size DGPs						
S1.1 (ind. white noises)	0.035	0.035	0.053	0.070	0.085	0.100
S1.2 (linear)	0.030	0.025	0.050	0.050	0.055	0.099
S1.3 (quadratic)	0.015	0.005	0.220	0.015	0.005	0.315
S1.4 (cubic)	0.000	0.000	0.375	0.000	0.000	0.436
S1.5 (non-additive)	0.005	0.545	0.221	0.020	0.600	0.313
S1.6 (time series)	0.035	0.035	0.062	0.090	0.060	0.103
S1.7 (time series nonlinear)	0.040	0.020	0.048	0.065	0.035	0.104
Power DGPs						
P1.1 (linear)	0.735	0.745	0.997	0.825	0.820	1
P1.2 (quadratic)	0.865	0.870	0.187	0.925	0.925	0.278
P1.3 (cubic)	0.995	1	1	1	1	1
P1.4 (non-additive)	1	1	0.260	1	1	0.352
P1.5 (quadratic)	0.965	0.975	0.204	0.995	0.990	0.285
P1.6 (time series nonlinear)	0.905	0.895	0.416	0.940	0.950	0.504

Note: n = 100; number of iterations = 200; number of bootstrap iterations (I) = 200.

formulated in equation 6.6. P-values for the Hellinger and Energy/Cramer tests are obtained using the local bootstrap procedure described in Section 6.2 with $I = 200$. The upper part of the table refers to size DGPs for which the hypothesis of conditional independence ($H_0 : X \perp\!\!\!\perp Y|Z$) always holds. For instance, when *X, Y*, and Z follow independent white noise processes, $H_0 : X \perp\!\!\!\perp Y|Z$ is rejected in 3.5% of all simulation runs using the Hellinger test (same result for the Energy/Cramer test) at the 0.05 level of significance and rejected in 7% of all runs at the 0.1 level of significance. That is, the p-value obtained for this case is greater than 0.05 in 3.5% of all simulations and greater than 0.1 for 7% of the simulations. Our results show that the Hellinger distance test (supported by the local bootstrap) performs quite well in all cases, except for the case of linear dependence. Therein, the frequency of rejection is satisfactory while not as high the one for the Fisher test. Such a result was to be expected since the linear case satisfies the assumptions required by the Fisher test. Both Energy and Cramer test (labeled "Euclid" in the table) perform quite similarly to the Hellinger test. In some cases they even slightly outperform the Hellinger test with somewhat lower rejection frequencies for size DGPs S1.2 and S1.7 and relatively higher rejection frequencies in many power DGPs (P1.1, P1.2, P1.3, P1.5). However, neither Energy nor Cramer test detect conditional independence in case of non-additive errors (S1.5). The results also confirm that we are led astray when applying the Fisher test in presence of nonlinear dependencies. In many of these cases (P1.2, P1.4, P1.5, P1.6) the power of the test turns out unsatisfactory. A better strategy proves to apply the Hellinger or, in case of additive errors, the Energy/Cramer test.

Table 6.5 finally displays our results for the case of two conditioned variables ($\#Z = 2$). As previously, columns "Hellinger", "Euclid", and "Fisher" refer to the

Table 6.5 Proportion of rejection of H_0 (two conditioned variables)

	Hellinger	Euclid	Fisher	Hellinger	Euclid	Fisher
	Level of significance 5%			*Level of significance* 10%		
Size DGPs						
S2.1 (independent white noises)	0.040	0.070	0.059	0.060	0.100	0.109
S2.2 (linear)	0.000	0.007	0.056	0.000	0.047	0.108
S2.3 (quadratic)	0.000	0.000	0.336	0.000	0.000	0.434
S2.4 (cubic)	0.000	0.000	0.028	0.007	0.000	0.068
S2.5 (non-additive)	0.960	0.253	0.190	0.993	0.340	0.268
S2.6 (time series linear)	0.006	0.020	0.050	0.033	0.046	0.102
S2.7 (time series non-linear)	0.000	0.010	0.035	0.000	0.040	0.087
Power DGPs						
P2.1 (linear)	1	1	1	1	1	1
P2.2 (quadratic)	1	1	1	1	1	1
P2.3 (quadratic)	0.273	0.573	1	0.320	0.673	1
P2.4 (cubic)	1	1	0.999	1	1	1
P2.5 (non-additive)	1	1	0.246	1	1	0.336
P2.6 (time series non-linear)	0.170	0.960	0.338	0.250	0.980	0.411

Note: n = 100; number of iterations = 150; number of bootstrap iterations (I) = 100.

Hellinger distance test, the Energy/Cramer test, and the Fisher's z test respectively. The Hellinger and Energy/Cramer tests here are based on four dimensional density functions. To save computational time, we lower the number of test iterations to 150 and the number of bootstrap iterations (I) to 100. All nonparametric tests perform well except for some cases. The Hellinger distance test fails in presence of nonadditive-errors (S2.5), quadratic dependencies (P2.3) and time series (P2.6). The Energy/Cramer test rejects somewhat less often in the S2.5 case, though still too frequently. Moreover, the power of the Energy/Cramer test outperforms the Hellinger test in the quadratic (P2.3) and time series case (P2.6). Fisher's z test does not produce satisfactory results for: S2.3, S2.5, P2.5, P2.6. To sum up, in absence of any information about the functional form, using the Energy/Cramer test proves the better strategy.

6.5 Concluding Remarks

We have assessed the performance of conditional independence tests to be used for graphical causal inference in continuous settings. Hitherto, the latter was based on parametric formulations of conditional independence, i.e., vanishing partial correlations. Such measures do, however, prove restrictive since they require linearity in the underlying dependencies and normally distributed errors. Here, we stress and compare nonparametric procedures operating on the distances between conditional kernel densities and on a local bootstrap. On one hand, our findings show these tests to reach a performance comparable to Fisher's z given linearity and normal errors. On the other hand, parametric tests perform very poorly given nonlinear data generating processes whereas nonparametric procedures still yield correct results. In continuous settings graphical causal inference hence cannot generally be based on the independence tests used so far. Their results lead astray when the functional form is not known and/or likely to be nonlinear. Any constraint-based causal discovery method, (Spirtes et al. 2000; Pearl 2000; Moneta 2008), can be applied on the basis of the tests proposed in this paper.

References

Baringhaus L, Franz C (2004) On a new multivariate two-sample test. J Multivariate Anal 88(1):190–206

Chlaß N, Krüger J (2007) The Wilcoxon signed rank test with discontinuous and dependent observations. Jena Econ Res Pap 032

Moneta A (2008) Graphical causal models and VARs: An assessment of the real business cycles hypothesis. Empirical Econ 35(2):275–300

Paparoditis E, Politis DN (2000) The local bootstrap for Kernel estimators under general dependence conditions. Ann Inst Statist Math 52:139–159

Pearl J (2000) Causality. Models, reasoning, and inference. Cambridge University Press, Cambridge

Reichenbach H (1956) The direction of time. University of California Press, Berkeley

Scheines R, Spirtes P, Glymour C, Meek C, Richardson T (1996) Tetrad III, tools for causal modeling. Erlbaum, New York

Spirtes P, Glymour C, Scheines R (2000) Causation, prediction, and search. MIT Press, Cambridge

Su L, White H (2008) A nonparametric Hellinger metric test for conditional independence. Economet Theor 24:829–864

Szekely GJ, Rizzo ML (2004) Testing for equal distributions in high dimension. InterStat November (5)

Wand MP, Jones MC (1995) Kernel smoothing. Chapman and Hall, London

Welsh AH, Lin X, Carroll RJ (2002) Marginal longitudinal nonparametric regression: Locality and efficiency of spline and Kernel methods. J Am Statis Assoc 97(458):482–492

Yatchew A (2003) Semiparametric regression for the applied econometrician. Cambridge University Press, Cambridge