Science & Philosophy Volume 8(1), 2020, pp. 3-20

Updating Statistical Measures of Causal Strength

H. D. Vinod *

Abstract

We address Northcott's (2005) criticism of Pearson's correlation coefficient 'r' in measuring causal strength by replacing Pearson's linear regressions by nonparametric nonlinear kernel regressions. Although new proof shows that Suppes' intuitive causality condition is neither necessary nor sufficient, we resurrect Suppes' probabilistic causality theory by using nonlinear tools. We use asymmetric generalized partial correlation coefficients from Vinod (2014) as our third criterion (denoted as Cr3) in addition to two more criteria (denoted Cr1 and Cr2). We aggregate the three criteria into one unanimity index, $UI \in [-100, 100]$, quantifying causal strengths associated with causal paths: $X_i \rightarrow X_j$, $X_j \rightarrow X_i$, and $X_i \leftrightarrow X_j$.
Keywords: kernel regression, generalized correlations
2010 AMS subject classifications: 97U99. [1]

*H. D. Vinod, Professor of Economics, Fordham University, Bronx, New York, USA 10458. E-mail: vinod@fordham.edu.

[1]Received on February 12th, 2020. Accepted on May 3rd, 2020. Published on June 30th, 2020. doi: 10.23756/sp.v8i1.497. ISSN 2282-7757; eISSN 2282-7765.

1 Introduction

Scientists and philosophers have debated the measurement of causal directions and strengths for a very long time. In pharmaceutical research, the Food and Drug Administration (FDA) has favored randomized controlled trials as the gold standard for assessing whether a drug is safe and effective. Many social scientists and philosophers have noted that controlled trials are often impractical, costly, time-consuming, and ethically unsuitable. Cartwright [2007] goes beyond the narrow scope of controlled experiments to argue that ultimately their basis remains deductive, implying that when assumptions fail to hold precisely, the causal strength measurements from controlled trials also become suspect.

Practitioners have long used Pearson's correlation coefficient 'r' developed in the 1890s as an indicator of causal strength. Rodgers and Nicewander [1988] list 13 interpretations of r including: a type of mean, a type of variance, the ratio of two means, the ratio of two variances, the slope of a line, the cosine of an angle, the tangent to an ellipse, and so forth. Building upon Sober [1988], and focusing on only two interpretations of r, Northcott [2005] highlights its limitations for measuring causal strengths. This paper refers to a newer generalized correlation coefficient r^* from Vinod [2014] and Vinod [2017b] implemented in a free software package Vinod [2017a] to overcome those limitations and lead to a newer practical measure of causal strength.

We limit the scope here by considering only those noisy causal strengths which can be computed in terms r or r^*. Hence, this paper bypasses the considerable literature dealing with Directed Acyclic Graphs (DAGs) or Pearl's 'do' operator, Pearl [2010]. We shall see that a study of r^* can include "variation in the features under study" representing counter-factuals described by Cartwright [2003]. Salmon [1977] [p. 151] suggests replacing the probabilities of *events* by causally connected *processes* defined as "spatio-temporally continuous entities" having their own physical status. Hence, our exploratory assessment of causal relations in this paper is assumed to be describable by regression equations between passively observed data generating processes (DGPs)–not between events. Our scope also excludes deterministic causal relations expressed as functional relations without random components. Thus, for example, Boyle's law (pressure *volume = a constant) where all component variables (pressure and volume) can be independently controlled in a laboratory, is beyond our scope.

Northcott distinguishes between absolute causal effectiveness CE_{abs}, what he calls "Gallalean idealization" isolating the contribution of a particular cause on the one hand, and CE_{rel}, or relative causal effectiveness representing the proportion of total noise explained by the cause. He argues that we are rarely interested in the relative concept, and that Pearson's r cannot measure CE_{abs}. One aim of this paper is to show how a generalized r, and two additional criteria together can

indeed quantify CE_{abs}.

Consider a set of p random variables $V = (X_1, \ldots X_p)$, with subscripts from the index set, $V_I = \{1, \ldots p\}$, and their joint density, $f(V)$. Unlike Cartwright, it is convenient here to let the same symbol X_i stand for observable proxies, as well as, the true underlying (cause or effect) variables. The conditional density is, by definition, the joint density divided by the marginal density, $f(X_i)$. The latter is the Radon-Nikodym derivative of the joint density, either with respect to the Lebesgue or the counting measure. The idea that X_i causes X_j is conveniently denoted by the causal path $X_i \to X_j$. For concreteness, the reader can think of X_i as a treatment, X_j as an outcome, and X_k as a set of background conditions often called control variables (e.g., age, gender, ethnicity), which affect the outcome.

Assume we have T observations denoted by $V_t = (X_{1t}, \ldots X_{pt}), t = 1, 2, \ldots T$. Let us denote by $LjRik$ a model having X_j on the left-hand side (LHS), and X_i plus a set of variable(s) combined into generic X_k on the right-hand side (RHS).

$$LjRik: \quad X_{jt} = f(X_{it}, X_{kt}) + \epsilon_{j|ik}, \tag{1}$$

where we use a generic f to denote a possibly nonlinear function, and ϵ denotes unobserved shocks or errors. According to one of 13 interpretations mentioned above, Pearson's r is a signed square root of the coefficient of determination R^2_{LjRik} of regression with linear f in (1).

Northcott [2005] lists the following two ambiguities associated with using Pearson's r.

(i) **Variable level versus its variance and covariance**: Northcott uses the example of X_j as the level of stock market price, and X_i as some variable affecting the stock price to argue that R^2 is focused on the variance of the stock price, which is of interest only to hedge funds focused on volatility, ignoring the stock price levels which are of interest to most investors. Our second criterion (Cr2) will explicitly consider absolute values of regression residuals $|\hat{\epsilon}_{j|ik}|$, obviously affected by levels of (X_i, X_j, X_k) even if we standardize the data variables to have zero mean and unit standard deviation.

(ii) r **cannot handle counter-factuals**. Northcott himself suggests a solution to the second problem by comparing (1) with an auxiliary regression representing a "baseline counter-factual" using the complete absence of X_i from the right-hand side as in the model:

$$LjRk: \quad X_{jt} = f(X_{kt}) + \epsilon_{j|k}. \tag{2}$$

Denote by R^2_{LjRk} the coefficient of determination of regression (2). Now Northcott implicitly suggests a simple extension of Pearson's r for computing the causal strength of $X_i \to X_j$ by computing

$$(X_i \to X_j)_{abs} = R^2_{LjRik} - R^2_{LjRk}. \tag{3}$$

When two more (original and auxiliary) regression equations are analogously defined by flipping X_i and X_j, in equations (1) and (2), we can define the strength of the causal path in the opposite direction by:

$$(X_j \rightarrow X_i)_{abs} = R^2_{LiRjk} - R^2_{LiRk}. \tag{4}$$

The strengths of the two paths from (3), and (4) will not, in general, be the same. Thus, even though the matrix of Pearson's r coefficients is symmetric, we have

$$(X_j \rightarrow X_i)_{abs} \neq (X_i \rightarrow X_j)_{abs}, \quad a.e., \tag{5}$$

where $a.e.$ denotes "almost everywhere" in a relevant measure space, in the sense that any violations of (5) are a 'set of measure zero' or extremely rare. Remark 1 in the sequel exploits the difference in the absolute magnitudes of the two sides of (5) to help assess the causal direction.

Vinod [2014] proves that a matrix of generalized correlation coefficients r^* is asymmetric. However, we cannot entirely rely on (4) or (5) as the sole criterion, because Northcott's item (i) listed above regarding variable levels remains unaddressed.

Mandel [2017] Table 4 summarizes ten definitions and concepts of causality from the literature, of which the 5-th is Suppes' probabilistic causality discussed in Hitchcock [2018]. It defines: X_i probabilistically causes X_j, if the information that X_i occurred increases the likelihood of X_j occurrence. The intuition behind "probabilistic theory of causation," Suppes [1970], is that if the causal path $(X_i \rightarrow X_j)$ holds, we should have:

$$P(X_j|X_i) > P(X_j) \quad a.e., \tag{6}$$

which is 'probabilistic,' because it holds $(a.e.)$, meaning that it may be violated on rare occasions. In modern probability (measure) theory, the notion of rare violations is described by the expression $(a.e.)$. We require the number of violations of the inequality (6) comprise 'a set of measure zero.'

Eells [1986] questions the intuition behind the inequality (6) by showing that a genuine cause need not raise the probability of a genuine effect when interacting with a third factor might be present. Some philosophers including Salmon [1977] have long criticized Suppes' theory with examples showing its logical failures. It is convenient to refer to those criticisms collectively as "old proofs," allowing us to state that the following Lemma provides a formal new proof.

Lemma 1: Suppes' condition (6) is neither necessary nor sufficient for causality

Proof:

Let X_k denote an additional omitted cause, which might be a confounder. By definition, 'confounder' X_k is a plausible underlying cause behind the apparent cause. Now, it is possible to construct counterexamples where the true causal paths are: $(X_k \to X_i)$, and $(X_k \to X_j)$. For example, let X_i denote the event of an atmospheric barometer falling sharply, and let X_j denote a weather storm event. These X_i, X_j events satisfy eq. (6). However, the barometer gauge itself does not cause the storm! The true cause X_k 'falling atmospheric pressure' is hidden from (6). Since barometer reading X_i is not necessary for the storm event X_j, this is a counterexample. Thus the "necessity" of (6) is rejected.

We reject sufficiency by using the definition of conditional probability as follows. Since conditional probability equals joint divided by marginal, we can rewrite (6) as

$$\frac{P(X_i \cap X_j)}{P(X_i)} > P(X_j),$$

or upon multiplying both sides by $P(X_i) > 0$ as:

$$P(X_i \cap X_j) > P(X_i) * P(X_j).$$

The inequality's sense remains intact if we divide both sides by a positive quantity, $P(X_j) > 0$, to yield the inequality:

$$\frac{P(X_i \cap X_j)}{P(X_j)} > P(X_i).$$

Thus we must always have

$$P(X_i|X_j) > P(X_i) \quad a.e. \tag{7}$$

This puzzling implication proves that Suppes' test satisfies conditions for $X_j \to X_i$ as well as $X_i \to X_j$ at all times. A result finding bidirectional causality $X_i \leftrightarrow X_j$ all the time means that the condition (6) is logically flawed, insufficient to distinguish between $X_i \leftrightarrow X_j$, and $X_i \to X_j$. **QED**.

Some philosophers and economists (e.g., Clive Granger) have suggested that the path $X_i \to X_j$ should further require that X_i must occur chronologically before X_j occurs, to help achieve a desirable asymmetry property. However, this is needlessly restrictive and inapplicable for human agents (who read newspapers) acting strategically at time t in anticipation of future events at time $t+1, t+2, \ldots$. Sayre [1977] also argues that temporal directionality is not needed.

Remark 1: Asymmetry via flipped models

Logically consistent probabilistic causality theory must retain robust asymmetry even when our causality testing condition(s) are stressed by flipping the cause

and effect (X_i and X_j). Since equations (6) and (7) suggesting opposite causal directions are proved to coexist, we need to go beyond the inequality signs, and consider the relative magnitudes of the differences: $(P(X_j|X_i) - P(X_j))$, and $(P(X_i|X_j) - P(X_i))$, in order to generalize Suppes' non-deterministic theory.

Remark 2: Confounders and controls distinguished

The causal path $X_i \rightarrow X_j$ assessment is often affected by two types of often present related events X_k. It is convenient to distinguish between two types of X_k: (i) 'confounder' and (ii) 'control' variables, even though the two may be synonymous for many readers. Recall that, by definition, 'confounder' X_k is a plausible underlying cause behind the apparent cause X_i for the outcome X_j. For example, the true cause of weather events X_j is 'atmospheric pressure' X_k and not 'barometer reading' as X_i. Second, we define X_k as a 'control' event if both (X_i, X_k) may be causing X_j, but we are interested in knowing if X_i causes X_j over and above the effect of X_k. For example, let X_j be health outcome, X_i be some medicine, then X_k, the patient's age, is commonly used as a control. A confounder can be treated as a control, but the converse may not hold true.

This paper develops a practical probabilistic theory of causality. Our Theorem 1 in the sequel proves that the revised theory does not suffer from the logical problems with Suppes' theory. When we try to develop computational methods for implementing the revised theory, we find that there are at least three coequal empirical criteria, denoted by Cr1, Cr2, and Cr3, which quantify the support for the causal path $X_i \rightarrow X_j$. It is not possible to prove why one criterion should dominate others. Hence, let us use the familiar 'preponderance of evidence' standard by computing a weighted sum of the three criteria denoted by $ui \in [-100, 100]$, and call it a sample unanimity index. Given a 5% threshold, say, (or $\tau = 5$), the index allows us to propose the following decision rules:

Rule 1: If $(ui < -\tau)$ the causal path is: $X_i \rightarrow X_j$.

Rule 2: If $(ui > \tau)$ the causal path is: $X_j \rightarrow X_i$.

Rule 3: If $(|ui| \leq \tau)$ we obtain bi-directional causality: $X_i \leftrightarrow X_j$, that is, the variables are jointly dependent.

Vinod [2017b] reports simulations showing good performance of these rules from a journal specializing in simulations.

Complete computational details for using these decision rules on any given data set are a part of an open source and a free software package called 'generalCorr', Vinod [2017a], in the computer language called R. It is readily available in an open forum for further criticism and development. The package comes with three vignettes that provide technical details about the algorithms used along with examples and citations to additional relevant papers, including Zheng et al. [2012].

A referee has pointed out that underlying ideas were partially anticipated by Yule in the late 1890s.

An outline of the remaining paper is as follows. Section 2 uses these Remarks 1 and 2 while avoiding logical problems with Suppes' theory to propose a revised version comprising our Theorem 1. We must translate the necessary and sufficient (iff) conditions for revised probabilistic causality into decision rules using the available data in the form of DGPs (X_i, X_j, X_k). This requires some sophisticated statistical tools. The intuition behind these tools is discussed in Section 3 without technical details. Section 4 briefly reviews kernel regressions for fitting nonlinear nonparametric functions. The residuals of these regressions are used to quantify Theorem 1 in Section 5. Section 4 specifies our three criteria Cr1 to Cr3. Section 6 develops a quantification of our Cr1 to Cr3, leading to a sample unanimity index $ui \in [-100, 100]$, summarizing the three criteria into a single number. Section 7 contains our final remarks.

2 Revised Probabilistic Theory of Causality Among DGPs

Following Remark 2 we define the set of variables X_k as containing both confounder and control variables. Since there are situations where X_k variables are completely out of the picture, we need two versions of the following result to accommodate both situations.

Theorem 1: Revised probabilistic Causality
(Version a) Assuming data on X_k are available, the causal path $X_i \to X_j$ holds if and only if (iff)

$$(P(X_j|X_i, X_k) - P(X_j|X_k)) > (P(X_i|X_j, X_k) - P(X_i|X_k)), \quad a.e. \tag{8}$$

(Version b) Assuming data on X_k are available, the causal path $X_i \to X_j$ holds if and only if (iff)

$$(P(X_j|X_i, X_k) - P(X_j)) > (P(X_i|X_j, X_k) - P(X_i)), \quad a.e. \tag{9}$$

Proof: Our proof removes the obstacles to proving 'necessity' described in Lemma 1 above by explicitly including X_k variables, which belong in the set of conditional variables in both versions of Theorem 1. Logical problems with Suppes' condition arise from the simultaneous existence of equations (6) and (7). We remove it by using flipped models (Remark 1) to impose asymmetry and focusing on relative sizes of inequalities of flipped models.

The iff condition from Theorem 1(Version a) becomes

$$(f(X_j|X_i, X_k) - f(X_j|X_k)) > (f(X_i|X_j, X_k) - f(X_i|X_k)) \quad a.e. \tag{10}$$

The slightly simpler iff condition from Theorem 1(Version b) becomes

$$(f(X_j|X_i, X_k) - f(X_j)) > (f(X_i|X_j, X_k) - f(X_i)) \quad a.e. \tag{11}$$

Since we eschew consideration of 'events' and focus on probabilities of DGPs, we can use widely accepted multiple regression to remove the effect of X_k, not readily available if we were to study probabilities of events. Besides DGPs, a further novelty here is to use nonlinear nonparametric kernel regressions (instead of linear regressions). Its advantages explained later include greater realism and superior statistical fits.

3 The Intuition behind Empirical evaluation of Theorem 1

The iff conditions established in Theorem 1 involve various probability density functions $f(.)$. Now consider the construction of these densities from data in the form of $t = 1, \ldots, T$ observations on DGPs for (X_i, X_j, X_k). We describe the intuition behind modern statistical methods used here, quantifying equations (10) and (11) starting with the simplest.

(a) **Marginal Densities:** Elementary statistics teaches us how to classify the data series X_i into a few class intervals and draw histograms. More advanced statistics papers describe smoothing of histograms into empirical approximations to marginal density functions $f(X_i)$, separately for each variable.

(b) **Joint Densities:** When we study two or more variables simultaneously, placing them along two or more axes, we need simultaneous class intervals for joint variation arising from multi-way histograms based on a joint binning of the data. Instead of histogram smoothing, quantification of joint densities of two or more variables is handled by using kernel weighting in higher dimensional spaces using modern computer-intensive methods.

(c) **Conditional Densities:** One defines conditional densities as 'joint density' divided by 'marginal density,' similar to conditional probabilities.

(d) **Standardization:** Recall that a typical evaluation of iff conditions requires quantification of the difference between two (conditional) densities. For example, the left-hand side of (10) is: $(f(X_j|X_i, X_k) - f(X_j|X_k))$. Since units of measurement of variables in different DGPs are likely to be distinct, it is intuitively obvious that one cannot simply subtract one density from another. We must first

standardize all variables to have zero means and unit standard deviations, technically known as requiring densities to have numerically comparable supports.

(e) **Stochastic Dominance:** Financial economists need tools to choose between two or more risky assets (e.g., buying Facebook or Amazon stocks) using the probability distributions of their expected future returns, extrapolated from data on past stock returns. When comparing two or more densities, we need to examine their distinct features exemplified by local mean, variance, skewness, and kurtosis. There are well-developed methods in Finance based on the concept of stochastic dominance of four orders to compare the four moments of the density, respectively. Empirical versions of stochastic dominance methods yield four sets of numbers comparing four orders of integrals of two 'empirical cumulative density functions' being compared. One can compute such four sets of numbers when comparing any two densities.

(f) **Conditional Expectation Functions & Kernel regressions:** Since joint densities $f(X_i, X_j, X_k)$ have three or more dimensions, they are difficult to quantify into one set of T numbers. Typical conditional density functions needed in our iff conditions such as $f(X_j|X_i, X_k)$, obtained from ratios of joint and marginal densities, are also multi-dimensional and difficult to quantify. Hence, we use T estimates of fitted values of kernel regression models associated with conditional *expectation* functions $Ef(X_j|X_i, X_k)$ evaluated at each $(t = 1, \ldots T)$.

In traditional linear regressions, conditional expectation functions contain regression coefficients that remain constant for all t and yield fitted values. Moreover, flipped linear regressions, $X_j = a + bX_i$ and $X_i = \tilde{a} + \tilde{b}X_j$, have identical R^2 values. Hence, we cannot assess causal directions from measures of goodness of fit of the flipped linear regressions. Therefore, we must use more sophisticated nonlinear kernel regressions described in the next section.

4 Kernel Regression Review

The linearity of the regression model is often a matter of convenience rather than an evidence-based choice. Back in 1784, the German philosopher Kant said: "Out of the crooked timber of humanity no straight thing was ever made." Since social sciences and medicine deal with human agents, evidence supporting linearity is often missing.

The main reason for using nonparametric nonlinear kernel regression in applied work is to avoid misspecification of the functional form. The best-fitting kernel regression line is often jagged, which does not have any polynomial or sinusoidal form. However, it provides a superior fit (higher R^2) by not assuming a functional form.

A disadvantage used to be its computational difficulty, which has recently dis-

appeared. The remaining disadvantages are that kernel regressions fail to provide partial derivatives and that out-of-sample forecasts can be poor. Fortunately, partial derivatives and out-of-sample forecasts are irrelevant for determining causal structures.

Let us replace the generic function f from (1) by the population conditional expectation function $G_1(X)$ when X_{kt} variables are omitted for ease of exposition without loss of generality. It will be estimated by nonlinear and nonparametric kernel regression. The sample estimate of G_1 is:

$$g_1(X) = \frac{\sum_{t=1}^{T} X_{jt} K(\frac{X_{it}-X}{h})}{\sum_{t=1}^{T} K(\frac{X_{it}-X}{h})}, \tag{12}$$

where $K(.)$ is the well known Gaussian kernel function, and h is the bandwidth chosen by leave-one-out cross-validation. The exposition becomes complicated when several regressors are involved, each needing a separate bandwidth of its own as described in Li and Racine [2007]. It is well known that kernel regression fits are superior to OLS.

Assuming that g_1 in eq. (12) belongs to $\mathscr{B}$, a class of Borel measurable functions having a finite second moment, then g_1 is an optimal predictor of X_j given X_i, in the sense that it minimizes the mean squared error (MSE) in the class of Borel measurable functions, [Li and Racine, 2007, Theorem 2.1]. Also note that kernel regression estimates are proved to be consistent and asymptotically normal (CAN) under certain assumptions, partly considered in the next subsection.

4.1 Kernel regression consistency

The sample kernel regression estimate g_1 of the population conditional expectation function G_1 is consistent provided true unknown errors in eq. (1) are orthogonal to the regressors. That is, the following probability limit should be zero, or:

$$\text{plim}_{T\to\infty}(\epsilon_{j|ik} X_{it})/T = 0. \tag{13}$$

We note in passing that nonlinear nonparametric kernel regressions prevent inconsistency induced in linear regressions by hidden nonlinearities. Assume that we approximate the hidden nonlinearity by a high order polynomial. Now a researcher using a linear model is implicitly letting high order polynomial terms merge into the regression error. Since the merged error will be correlated with the regressor due to misspecification, it will induce inconsistency.

4.2 Implicit Counter-factuals in Cross-validation:

Counter-factuals are defined as "what has not happened but could happen" in the avaliable data. Since experimental manipulation is often not an option, especially in social sciences, many authors use virtual manipulation involving counter-factuals, implicit in cross-validation described next.

Considering $\{X_{it}, X_{jt}, X_{kt}\}$ data, when we pretend that t-th observation is absent, even though it is present, we have a counter-factual. Now leave-one-out cross-validation used to determine bandwidth h appearing in (12) of kernel regressions minimizes a weighted error sum of squares

$$min_h \quad \frac{1}{T}\Sigma_t[Y_t - \hat{g}_{1,-tL}]^2 W(X_t), \tag{14}$$

where $W(.)$ is a weight function, subscript $(-t)$ denotes omitting t-th observation, and where the subscript (L) refers to a local linear fit. We employ cross-validation as a counter-factual in our determination of (g_1) conditional expectation functions, which will eventually determine our causal direction and its strength.

5 Quantification of Theorem 1 from residuals

Recall that numerical quantification of the causal path $(X_i \to X_j)$ between standardized DGPs from Theorem 1(b), requires that we evaluate the four-part inequality: $[(P1 - P2) > (P3 - P4)]$, where the four parts P1 to P4 are readily seen from (10) to be:

$$(f(X_j|X_i) - f(X_j)) > (f(X_i|X_j) - f(X_i)).$$

Quantification of the first part, P1:

In randomized controlled experiments, the conditioning variables are randomly assigned to experimental units $(X_{it}, t = 1, \ldots, T)$, and the researcher records corresponding values of X_{jt}. Whether experimental or passively observed, one can construct T_i class intervals and record the probabilities (relative frequencies) of various values of X_j in each class interval (group). Midpoints of class intervals and associated relative frequencies as probabilities yield T_i numbers representing the conditional density P1 $=f(X_j|X_i)$. The range of observed values of X_i represents the 'support' of P1 density. If each class interval is to have at least five observations on an average for a reasonable estimate of conditional density, T will be five times larger than T_i. In general, T_i class intervals will be far fewer than T, expressed by $T_i << T$. We like to avoid cumbersome construction of class intervals, which always involves loss of information since they are not 'sufficient' statistics.

Paragraph (f) of Section 3 notes that conditional expectation functions from kernel regression of X_j on X_i can yield consistent estimates of fitted values of $\hat{X}_{jt|i}$ containing T numbers. We have noted before that fitted values obtained by using bandwidths from leave-one-out cross-validation perform a form of counterfactual, relevant for conditional density estimation.

Next, consider the **quantification of P2**, which is $f(X_j)$. Note that data on X_{jt} can readily construct an empirical cumulative distribution function (ECDF) whose Radon-Nikodym derivative is X_{jt}, providing T numbers which represent the density P2.

Regression residuals are defined as 'observed minus fitted' values: $e_{jt|i} = X_{jt} - \hat{X}_{jt|i}$. Using quantified P1 and P2 described above, **quantification of P1–P2** is available from negative residuals or $-e_{jt|i}$.

Consider **quantification of P3–P4**. Note that this is completely analogous to P1–P2, when we simply flip i and j. Hence, it is easy to verify that P3–P4 is quantified by the negative of the residuals of a flipped kernel regression or $-e_{it|j}$.

Finally, we can assess whether the inequality $P1 - P2 > P3 - P4$ in Theorem 1(b) holds by computing the following inequality, where we replace negative residuals by positive ones, change the sense of the inequality from (>) to (<) and also insert absolute value symbols to yield:

$$|e_{j|i}| = |X_j - E[\hat{f}(X_j|X_i)]| \; < \; |X_i - E[\hat{f}(X_i|X_j)]| = |e_{i|j}|, \quad a.e. \tag{15}$$

It is easy to verify that the densities in Theorem 1(a) are similar to the P1 to P4 discussed above, except that we have to condition all densities on control variables X_{kt}. Thus causal path $X_i \rightarrow X_j$ after removing the effects of confounder or control variable(s) X_k can be approximately assessed by

$$|e_{j|ik}| = |X_j - E[\hat{f}(X_j|X_i, X_k)]| \; < \; |X_i - E[\hat{f}(X_i|X_j, X_k)]| = |e_{i|jk}|, \quad a.e., \tag{16}$$

An intuitive interpretation of this inequality is that the causal path $X_i \rightarrow X_j$ requires kernel regression LjRik with causal variable X_i on the RHS should have a superior fit compared to the flipped regression LiRjk.

The inequalities (15) and (16) are considered fuzzy since they hold with some exceptions, almost everywhere but not everywhere. A concrete example is illustrated in Vinod [2017a] using European data. It has the crime rate as X_i, police deployment rate as X_j. The causal path from high crime to high police deployment, $X_i \rightarrow X_j$, requires that regression residuals for the model with X_i on RHS are "smaller" in some fuzzy sense than the flipped model with X_j on RHS.

6 Criteria Cr1 to Cr3 using residuals

The discussion so far has quantified Theorem 1 as amounting to certain fuzzy inequalities among T numbers obtained from residuals of kernel regressions with flipped i and j in equations (15) and (16). Our next task is to develop three criteria (Cr1 to Cr3) using these residuals to quantify the strength of the causal path, which will eventually yield our unanimity index ui.

Our first criterion Cr1 described next evaluates finite sample implications of consistency of conditional expectation functions in kernel regressions described in Section 4.1. We plug observable residuals into the consistency condition (13), yielding two sets of probability limit ('plim') expressions. We simply compute absolute values of T multiplicative expressions: $|e_{j|ik}X_{it}|$ and $|e_{i|jk}X_{jt}|$. Our Cr1 exploits the theoretical result that closeness to zero of these expressions implies faster convergence.

6.0.1 First criterion Cr1 for $X_i \to X_j$

Long ago, Koopmans [1950] formulated the consistency requirement of eq. (13) as exogeneity of X_i and went on to require that each right-hand side (RHS) variable should be exogenous in the sense that it should "approximately cause" the LHS variable. Being the oldest, we let this criterion based on comparing residuals provide our first criterion.

Since Kernel regressions are consistent, the conditional expectation functions are also consistent. Since speeds of convergence can differ, one should prefer the conditioning with a faster convergence rate. The obvious finite sample indicators of speeds of convergence are available from eq. (13) when we replace the true unknown errors by residuals. If the conditional expectation function when X_i on RHS converges faster to its true value than its converse, the T values implied by the 'plim' expression of the LjRi model should be closer to zero than the similar 'plim' expression of the flipped LiRj model.

Hence, the condition for the causal path $X_i \to X_j$ based on the faster convergence rate of the LjRik model, than that of the flipped LiRjk model, is the inequality:

$$\text{Cr1}: \quad |e_{j|ik}X_{it}| < |e_{i|jk}X_{jt}|, \quad a.e., \tag{17}$$

where the residuals are comparable in numerical magnitudes because we have standardized all variables.

Working with residuals overcomes a criticism of 'r' defined in terms of variances and covariance by Northcott [2005] mentioned before, that 'r' ignores data 'levels.' Residuals are obviously sensitive to levels.

Note that we have T numbers for each side of the inequality, and we want to compare whether one set is "larger" than the other, analogous to the investor's

problem in choosing one distribution of two risky investment options. An old solution to the problem uses stochastic dominance methods for four orders approximating local mean, variance, skewness and kurtosis of the two distributions, yielding four sets of numbers for a thorough and sophisticated comparison of two investment options. Here, we simply apply the tools from Finance to our problem.

6.0.2 Second criterion Cr2 for $X_i \rightarrow X_j$

Our second criterion simply restates the fuzzy inequality (16) quantifying our Theorem 1. Recall that it checks whether independent changes in X_i lead to (dependent) changes in X_j, leading to LjRi model providing a better fit than LiRj. Hence we require:

$$\text{Cr2}: \quad |e_{j|ik}| < |e_{i|jk}|, \quad a.e. \tag{18}$$

Note that Cr2 is similar to Cr1, requiring a comparison of two sets of T numbers. Hence we quantify Cr2 by using stochastic dominance of four orders.

6.0.3 Third criterion Cr3 for $X_i \rightarrow X_j$

Following Vinod [2014] an aggregate manifestation of the 'a.e.' inequality (16) involving residuals: $e_{j|ik}, e_{i|jk}$, can be stated in terms of a higher coefficient of determination R^2 for one of the two flipped models. The effect of X_k variable(s), if any, on X_i, X_j is netted out in these computations to yield our third criterion:

$$\text{Cr3}: \quad R^2_{j|i,k} = 1 - \frac{\Sigma(e_{j|ik})^2}{(\text{TSS})} > 1 - \frac{\Sigma(e_{i|jk})^2}{(\text{TSS})} = R^2_{i|j,k}, \tag{19}$$

where TSS denotes the total sum of squares, which is $(T-1)$ for standardized data, and where we denote the conditioning in the two models by subscripts to R^2.

An equivalent requirement using generalized partial correlation coefficients from Vinod [2017a] for $X_i \rightarrow X_j$ is:

$$|r^*_{(j|i;\,k)}| > |r^*_{(i|j;\,k)}|. \tag{20}$$

Two advantages of Cr3 are that it involves a simple comparison to two summary statistics that can be computed without having to standardize the data.

R package 'generalCorr' reports the generalized partial correlation coefficients in eq. (20), if desired. The R function `pacorMany` provides partial correlation coefficients of the first column paired with all others after removing the effect of a specified list of control variables. Recall that Theorem 1(a) eq. (10) considers netting out of the confounders X_k from both causal X_i and outcome X_j variables, exactly as it is implemented in computing (20).

Since quantification of Cr1 and Cr2 yields four numbers for four orders of stochastic dominance, we need a weighting scheme to get one summary number each denoted by $N_{cri}, i = 1, 2$. Cr3 yields only one sign leading to N_{cr3}. Our unanimity index transforms a weighted sum of N_{cri} into $ui \in [-100, 100]$ needed for the decision rules Rule 1 to Rule 3. Vinod [2017b] reports successful simulations using the rules and tools for bootstrap-based statistical inference to study sampling variability of ui.

7 Summary and Final Remarks

This paper supports Northcott [2005] view that the Pearson correlation coefficient 'r' is unsuitable for measuring causal strength or absolute causal effectiveness (CE_{abs}). Instead of 'r' we suggest recently developed generalized partial correlation coefficients $r^*_{i,jk} \neq r^*_{j,ik}$ based on nonparametric kernel regressions, Vinod [2014]. Partial correlations help quantify our third criterion Cr3. Since Cr3 is sensitive to variances and covariances but not data 'levels,' we overcome Northcott's criticism of 'r' by not relying on Cr3 alone. We use additional coequal criteria, Cr1 and Cr2, based on certain residuals sensitive to data levels.

Our Lemma 1 in Section 1 provides a new proof showing that the condition $P(X_j|X_i) > P(X_j)$ proposed in Suppes' "probabilistic causality theory" is logically faulty for the causal path $X_i \rightarrow X_j$, because it always coexists with the opposite path $X_j \rightarrow X_i$.

Our Theorem 1 incorporates additional variables X_k if confounding variables vitiating the causal paths are present and goes on to update Suppes' theory by providing new iff conditions in the form of inequalities involving conditional and marginal densities. Direct quantification of inequalities among conditional densities using limited, nonexperimental, and passively observed data (without manipulation) is obviously difficult. Hence, we suggest a deeper study of two nonparametric kernel regressions, LjRik: $X_{jt} = f(X_{it}, X_{kt}) + \epsilon_{j|ik}$, and a flipped regression where i and j are interchanged, LiRjk.

We employ three coequal empirical criteria Cr1 to Cr3 to quantify support for the causal path $X_i \rightarrow X_j$ using the absolute values of residuals of two regressions LjRik and LiRjk. Sophisticated comparisons need to look at all locally defined mean, variance, skewness, and kurtosis properties of empirical cumulative distribution functions implied by absolute values of residuals. We employ readily available tools from Financial Economics to quantify our criteria and develop a unanimity index ui used to provide decision rules for choosing between three causal paths $X_i \rightarrow X_j$, $X_, \rightarrow X_i$, and $X_i \leftrightarrow X_j$.

Our methods are completely transparent (open source, free) described in the R package 'generalCorr' and its three vignettes, providing examples, simulations,

and all details. One R command, `causeSummary` is worthy of further attention as an approximate implementation of Theorem 1. The 'generalCorr' decision rules are recently used in Lister and Garcia [2018] to conclude that global warming causes anthropod deaths, in Allen and Hooper [2018] to explore causes of volatility in stock prices, in Mlynczak and Krysztofiak [2019] to study causal links between sports and cardio-respiratory issues faced by elite athletes, and in Fousekis [2020] to study US commodity futures markets.

References

David E Allen and Vince Hooper. Generalized correlation measures of causality and forecasts of the VIX using non-linear models. *Sustainability*, 10((8): 2695): 1–15, 2018. doi: 10.3390/su10082695. URL https://www.mdpi.com/2071-1050/10/8/2695.

Nancy Cartwright. Two theorems on invariance and causality. *Philosophy of Science*, 70(1):203–224, 2003.

Nancy Cartwright. Are RCTs the gold standard? *BioSocieties*, 2(1):1–20, 2007.

Ellery Eells. Probabilistic causal interaction. *Philosophy of Science*, 53(1):52–64, 1986.

P. Fousekis. Returns and volume: Kernel causality in the US futures markets for agricultural, energy and metal commodities. *Studies in Economics and Finance*, 2020. URL https://doi.org/10.1108/SEF-10-2019-0416.

Christopher Hitchcock. Probabilistic causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018. URL https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/.

Tjalling C Koopmans. When is an equation system complete for statistical purposes. Technical report, Yale University, 1950. URL http://cowles.econ.yale.edu/P/cm/m10/m10-17.pdf.

Q Li and J S Racine. *Nonparametric Econometrics*. Princeton University Press, 2007.

Bradford C. Lister and Andres Garcia. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proceedings of the National*

Academy of Sciences, Oct. 15:1–10, 2018. URL http://www.pnas.org/content/early/2018/10/09/1722477115.full.pdf.

Igor Mandel. Troublesome dependency modeling: Causality, inference, statistical learning. *SSRN eLibrary*, 2017. URL https://ssrn.com/abstract=2984045.

Marcel Mlynczak and Hubert Krysztofiak. Cardiorespiratory temporal causal links and the differences by sport or lack thereof. *Frontiers in Physiology*, 10 (45):1–14, 2019. doi: 10.3389/fphys.2019.00045.

Robert Northcott. Pearson's wrong turning: Against statistical measures of causal efficacy. *Philosophy of Science*, 72(5):900–912, 2005.

Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, pages 1–59, 2010. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836213/.

Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

Wesley C. Salmon. An "at-at" theory of causal influence. *Philosophy of Science*, June, 44(2):215–224, 1977. URL http://www.unige.ch/lettres/baumgartner/docs/kausa/protect/salmon.pdf.

Kenneth M. Sayre. Statistical models of causal relations. *Philosophy of Science*, 44(2):203–214, 1977.

Elliott Sober. Apportioning causal responsibility. *Journal of Philosophy*, 85(6): 303 – 318, 1988. URL https://www.jstor.org/stable/2026721.

Patrick Suppes. *A probabilistic theory of causality*. Amsterdam: North-Holland, 1970.

H. D. Vinod. Causal paths and exogeneity tests in generalCorr package for air pollution and monetary policy, 2017a. URL https://cloud.r-project.org/web/packages/generalCorr/vignettes/generalCorr-vignette3.pdf. A vignette accompanying the package "generalCorr" of R.

Hrishikesh D Vinod. Matrix algebra topics in statistics and economics using R. In M. B. Rao and C. R. Rao, editors, *Handbook of Statistics: Computational Statistics with R*, volume 34, chapter 4, pages 143–176. North-Holland, Elsevier Science, New York, 2014.

Hrishikesh D. Vinod. Generalized correlation and kernel causality with applications in development economics. *Communications in Statistics - Simulation and Computation*, 46(6):4513–4534, 2017b. URL https://doi.org/10.1080/03610918.2015.1122048. Available online: 29 Dec 2015.

Shurong Zheng, Ning-Zhong Shi, and Zhengjun Zhang. Generalized measures of correlation for asymmetry, nonlinearity, and beyond. *Journal of the American Statistical Association*, 107(499):1239–1252, September 2012.