

Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables

Lionel Barnett*

*Centre for Computational Neuroscience and Robotics, School of Informatics, University of Sussex,
Brighton BN1 9QJ, United Kingdom*

Adam B. Barrett[†] and Anil K. Seth[‡]

*Sackler Centre for Consciousness Science, School of Informatics, University of Sussex, Brighton BN1 9QJ, United Kingdom
(Received 23 October 2009; published 4 December 2009)*

Granger causality is a statistical notion of causal influence based on prediction via vector autoregression. Developed originally in the field of econometrics, it has since found application in a broader arena, particularly in neuroscience. More recently *transfer entropy*, an information-theoretic measure of time-directed information transfer between jointly dependent processes, has gained traction in a similarly wide field. While it has been recognized that the two concepts must be related, the exact relationship has until now not been formally described. Here we show that for Gaussian variables, Granger causality and transfer entropy are entirely equivalent, thus bridging autoregressive and information-theoretic approaches to data-driven causal inference.

DOI: 10.1103/PhysRevLett.103.238701

PACS numbers: 89.70.Cf, 05.40.-a, 87.10.Mn, 87.19.1o

The problem of inferring causal interactions from data has challenged scientists and philosophers for centuries [1]. One approach that has become increasingly popular over recent years was introduced originally by Wiener [2], and formalized in terms of linear autoregression by Granger [3]. According to Wiener-Granger causality (*G* causality), given sets of interdependent variables X and Y , it is said that “ Y *G*-causes X ” if, in an appropriate statistical sense, Y assists in predicting the future of X beyond the degree to which X already predicts its own future. Importantly, identification of a *G*-causality interaction is *not* identical to identifying a physically instantiated causal interaction in a system. Although the two descriptions are intimately related [4,5], physically instantiated causal structure can only be unambiguously identified by perturbing a system and observing the consequences [1]. Nonetheless, *G* causality is pragmatic, well defined, and has delivered many insights into the functional connectivity of systems in a variety of fields, particularly in neuroscience [6].

The information-theoretic notion of *transfer entropy* was formulated by Schreiber [7] as a measure of directed (time-asymmetric) information transfer between joint processes. In contrast to *G* causality, transfer entropy is framed not in terms of prediction but in terms of *resolution of uncertainty*. One can say that “the transfer entropy from Y to X ” is the degree to which Y disambiguates the future of X beyond the degree to which X already disambiguates its own future. There is therefore an attractive symmetry between the notions (“predicts” \leftrightarrow “disambiguates”) which has been noted previously (see, e.g., [8]) but never explicitly specified. In this Letter we show that under Gaussian assumptions *G* causality and transfer entropy are in fact entirely equivalent. Our results therefore provide

a framework for inferring causality which unifies information-theoretic and autoregressive approaches.

We use a standard mathematical vector-matrix notation in which bold type generally denotes vector quantities and uppercase type denotes matrices or random variables, according to context. All vectors are considered to be *row* vectors. The symbol \top denotes the transpose operator and \oplus denotes *concatenation* of vectors, so that for $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_m)$, $\mathbf{x} \oplus \mathbf{y}$ is the $1 \times (n + m)$ vector $(x_1, \dots, x_n, y_1, \dots, y_m)$.

Given jointly distributed multivariate random variables (i.e., random vectors) X, Y , we denote by $\Sigma(X)$ the $n \times n$ matrix of covariances $\text{cov}(X_i, X_j)$ and by $\Sigma(X, Y)$ the $n \times m$ matrix of cross-covariances $\text{cov}(X_i, Y_\alpha)$. We then use $\Sigma(X|Y)$ to denote the $n \times n$ matrix

$$\Sigma(X|Y) \equiv \Sigma(X) - \Sigma(X, Y)\Sigma(Y)^{-1}\Sigma(X, Y)^\top \quad (1)$$

defined when $\Sigma(Y)$ is invertible. $\Sigma(X|Y)$ appears as the covariance matrix of the residuals of a linear regression of X on Y [cf. Eq. (3) below]; thus, by analogy with *partial correlation* [9] we term $\Sigma(X|Y)$ the *partial covariance* [10] of X given Y .

Suppose we have a multivariate stochastic process X_t in discrete time [11] (i.e., the random variables X_{ti} are jointly distributed). We use the notation $X_t^{(p)} \equiv X_t \oplus X_{t-1} \oplus \dots \oplus X_{t-p+1}$ to denote X itself, along with $p - 1$ lags, so that $X_t^{(p)}$ is a $1 \times pn$ random vector for each t . Given the lag p , we use the shorthand notation $X_t^- \equiv X_{t-1}^{(p)}$ for the lagged variable.

Let X, Y be jointly distributed random vectors and consider the linear regression

$$X = \alpha + Y \cdot A + \epsilon, \quad (2)$$

where the $m \times n$ matrix A comprises the regression coefficients, $\alpha = (\alpha_1, \dots, \alpha_n)$ are the constant terms, and the random vector $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ comprises the residuals. The mean squared error may then be written in terms of the covariance matrix of the residuals as $E^2 = \text{trace}[\Sigma(\epsilon)]$. E^2 is just the sum of the variances of the ϵ_i , sometimes known as the *total variance*. Performing an ordinary least squares (OLS) to find the coefficients A that minimize E^2 yields [assuming $\Sigma(Y)$ invertible] $A = \Sigma(Y)^{-1}\Sigma(X, Y)^T$ and we find that for the least squares fit the covariance matrix of the residuals is given by

$$\Sigma(\epsilon) = \Sigma(X|Y) \quad (3)$$

with $\Sigma(X|Y)$ the partial covariance as defined by (1). We note that the same coefficients A which minimize the total variance E^2 also minimize the *generalized variance* $|\Sigma(\epsilon)|$ [12], where $|\cdot|$ denotes the determinant (this procedure is sometimes referred to as “least generalized variance”; see, e.g., [13]).

If the residuals ϵ can be taken to be uncorrelated with the regressors Y in (2)—as would be the case, for instance, for a multivariate autoregressive (MVAR) model—the residual covariance matrix can be derived directly from (2). Taking the covariance of both sides of (2) yields

$$\Sigma(X) = A^T \Sigma(Y) A + \Sigma(\epsilon). \quad (4)$$

Since the residuals and regressors are uncorrelated, we also have

$$\begin{aligned} 0 &= \Sigma(Y, \epsilon) \\ &= \Sigma(Y, X - \alpha - Y \cdot A) \\ &= \Sigma(X, Y)^T - \Sigma(Y) A. \end{aligned} \quad (5)$$

Solving (5) for A and substituting in (4) we recover Eq. (3) for $\Sigma(\epsilon)$. We note that Eqs. (4) and (5) are essentially *Yule-Walker* equations [6] for the regression (2).

Suppose now we have three jointly distributed, stationary [14] multivariate stochastic processes X_t, Y_t, Z_t (“variables” for brevity). Consider the regression models

$$X_t = \alpha_t + (X_{t-1}^{(p)} \oplus Z_{t-1}^{(r)}) \cdot A + \epsilon_t, \quad (6)$$

$$X_t = \alpha'_t + (X_{t-1}^{(p)} \oplus Y_{t-1}^{(q)} \oplus Z_{t-1}^{(r)}) \cdot A' + \epsilon'_t, \quad (7)$$

so that the “predictee” variable X is regressed first on the previous p lags of itself plus r lags of the conditioning variable Z and second, in addition, on q lags of the “predictor” variable Y [15]. The G causality of Y to X given Z is a measure of the extent to which inclusion of Y in the second model (7) reduces the prediction error of the first model (6).

The standard measure of G causality in the literature is defined for *univariate* predictor and predictee variables Y and X , and is given by the natural logarithm of the ratio of the residual variance in the restricted regression (6) to that

of the unrestricted regression (7). In our notation [16]

$$\begin{aligned} \mathcal{F}_{Y \rightarrow X|Z} &\equiv \ln \left(\frac{\text{var}(\epsilon_t)}{\text{var}(\epsilon'_t)} \right) = \ln \left(\frac{|\Sigma(\epsilon_t)|}{|\Sigma(\epsilon'_t)|} \right) \\ &= \ln \left(\frac{|\Sigma(X|X^- \oplus Z^-)|}{|\Sigma(X|X^- \oplus Y^- \oplus Z^-)|} \right), \end{aligned} \quad (8)$$

where the last equality follows from the general formula (3). By stationarity this expression does not depend on time t , so we drop the subscript when there is no danger of confusion. Note that the residual variance of the first regression will always be larger than or equal to that of the second, so that $\mathcal{F}_{Y \rightarrow X|Z} \geq 0$ always. As regards statistical inference, it is known that the corresponding maximum likelihood estimator $\hat{\mathcal{F}}_{Y \rightarrow X|Z}$ will have (asymptotically for large samples) a χ^2 distribution under the null hypothesis $\mathcal{F}_{Y \rightarrow X|Z} = 0$ [17,18] and a noncentral χ^2 distribution under the alternative hypothesis $\mathcal{F}_{Y \rightarrow X|Z} > 0$ [19,20].

Although rarely considered in the literature, there is no requirement in principle that either the predictee or predictor variable be univariate. In this Letter we address the general case where all variables are allowed to be *multivariate*; see [21,22] for motivation and discussion regarding this generalization. For the case of a multivariate predictor, Eq. (8) above (with Y replaced by the bold-type Y) is a valid and consistent formula for G causality. However, generalization to the case of a multivariate predictee is less clear-cut and there does not yet appear to be a standard definition for G causality in the literature. Here we use an extension first proposed by Geweke [19], in which the residual variance $\text{var}(\epsilon_t) = \Sigma(\epsilon_t)$ is replaced by the generalized variance $|\Sigma(\epsilon_t)|$:

$$\begin{aligned} \mathcal{F}_{Y \rightarrow X|Z} &\equiv \ln \left(\frac{|\Sigma(\epsilon_t)|}{|\Sigma(\epsilon'_t)|} \right) \\ &= \ln \left(\frac{|\Sigma(X|X^- \oplus Z^-)|}{|\Sigma(X|X^- \oplus Y^- \oplus Z^-)|} \right). \end{aligned} \quad (9)$$

This formula always produces a non-negative quantity, and for a univariate predictee reduces to (8). Moreover, its estimator is also asymptotically χ^2 distributed. Geweke [19] lists a number of motivations for this choice, to which we add the result presented in this Letter. [An alternative formulation for multivariate G causality is proposed in [21], although see [22] for more detailed discussion and further motivation for the form (9).]

With X_t, Y_t, Z_t as before, the *transfer entropy* of Y to X given Z [7,23] is defined as the difference between the entropy of X conditioned on its own past and the past of Z , and its entropy conditioned, in addition, on the past of Y :

$$\mathcal{T}_{Y \rightarrow X|Z} \equiv H(X|X^- \oplus Z^-) - H(X|X^- \oplus Y^- \oplus Z^-), \quad (10)$$

where $H(\cdot)$ denotes entropy and $H(\cdot|\cdot)$ conditional entropy.

Again, by stationarity transfer entropy does not depend on time t , and $\mathcal{T}_{Y \rightarrow X|Z} \geq 0$ always. $\mathcal{T}_{Y \rightarrow X|Z}$ may be understood as the degree of uncertainty of X resolved by the past of Y over and above the degree of uncertainty of X resolved by its own past. As with Granger causality, the transfer entropy literature generally deals only with univariate variables, although in this case the extension (10) to the multivariate case is unproblematic.

We now turn to the equivalence with G causality. For a multivariate Gaussian random variable X we have the well-known expression [24]

$$H(X) = \frac{1}{2} \ln(|\Sigma(X)|) + \frac{1}{2} n \ln(2\pi e)$$

for entropy in terms of the determinant of the covariance matrix, where n is the dimension of X . We now show that the conditional entropy $H(X|Y)$ for two jointly multivariate Gaussian variables may be expressed in terms of the determinant of the corresponding partial covariance matrix:

$$H(X|Y) = \frac{1}{2} \ln(|\Sigma(X|Y)|) + \frac{1}{2} n \ln(2\pi e). \quad (11)$$

To see this, we have

$$\begin{aligned} H(X|Y) &\equiv H(X \oplus Y) - H(Y) \\ &= \frac{1}{2} \ln(|\Sigma(X \oplus Y)|) - \frac{1}{2} \ln(|\Sigma(Y)|) + \frac{1}{2} n \ln(2\pi e). \end{aligned}$$

Now

$$\Sigma(X \oplus Y) = \begin{pmatrix} \Sigma(X) & \Sigma(X, Y) \\ \Sigma(X, Y)^\top & \Sigma(Y) \end{pmatrix}$$

and from the block determinant identity [25]

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| |A - BD^{-1}C|$$

we have

$$|\Sigma(X \oplus Y)| = |\Sigma(Y)| \cdot |\Sigma(X|Y)|$$

from which we obtain (11) [26].

If, then, the processes X_t , Y_t , Z_t are jointly multivariate Gaussian (i.e., any finite subset of the component variables X_{it} , $Y_{\alpha t}$, $Z_{\alpha t}$ has a joint Gaussian distribution) it follows from (11) that the expression (10) for transfer entropy becomes [27]

$$\mathcal{T}_{Y \rightarrow X|Z} \equiv \frac{1}{2} \ln \left(\frac{|\Sigma(X|X^- \oplus Z^-)|}{|\Sigma(X|X^- \oplus Y^- \oplus Z^-)|} \right). \quad (12)$$

Comparing (12) with (9) leads directly to our central result: if all processes are jointly Gaussian, then

$$\mathcal{F}_{Y \rightarrow X|Z} = 2\mathcal{T}_{Y \rightarrow X|Z} \quad (13)$$

so that Granger causality and transfer entropy are equivalent up to a factor of 2. This result holds, in particular, for a univariate predictee X with the standard definition (8) of G causality.

Empirically, numerical equivalence between G causality and transfer entropy will depend on the method used to estimate the transfer entropy in sample. If it is assumed at the outset that the data may be reasonably modeled as Gaussian—and that, consequently, conditional entropies may be estimated from the appropriate sample covariance matrices—then, of course, numerical equivalence will be guaranteed. If, however, conditional entropies are estimated directly from sampled probability distributions, results will vary with the estimation technique. It is known that naive estimation of transfer entropy by partitioning of the state space is problematic [7] and that such estimators frequently fail to converge to the correct result [23]. In practice, more sophisticated techniques such as kernel [28] or k -nearest neighbor estimators [29,30], will need to be deployed; however, such techniques may entail their own assumptions about the empirical distribution of the data (see [23] for a good discussion on these points). Furthermore, unlike G causality, for which the (asymptotic) distribution of the sample statistic is known, we are not aware of any such general result for transfer entropy. Thus, in particular, significance testing for transfer entropy estimates is likely to be hard.

Our result (13) provides for the first time a unified framework for data-driven causal inference that bridges information-theoretic and autoregressive methods. In particular, it opens new research possibilities in transforming findings originally developed in one domain into the other. For example, an advantage of the autoregressive approach is that it admits a straightforward decomposition by frequency [6,19]. Our result now provides a foundation for the development of spectral implementations of transfer entropy. In the opposite direction, the invariance of information-theoretic quantities under general nonlinear transformations [23] could potentially prove useful in the identification of appropriate nonlinear autoregressive models [31,32]. Preliminary work by the authors indicates, perhaps surprisingly, that under Gaussian assumptions there is nothing extra to account for by nonlinear extensions to G causality, since a stationary Gaussian AR process is necessarily linear [22]. This finding has practical significance because sensitivity to nonlinear data features is often presented as a reason to prefer transfer entropy to G causality (see, e.g., [33]).

As regards Gaussian assumptions, although their appropriateness may be disputed in the context of specific physical systems, they are, nevertheless, widely employed in neuroscience, econometrics, and beyond, frequently in the role of an analytical benchmark for subsequent more physically motivated analysis. In practice, given empirical data it is likely to be difficult to establish the extent to which Gaussian assumptions are tenable, particularly for highly multivariate data sets and limited sample sizes. Further research is thus required to characterize—both analytically and in sample—the manner in which the

equivalence (13) breaks down when Gaussian assumptions fail. As a starting point it is known, at least, that in the generic (non-Gaussian) case, nonzero G causality implies nonzero transfer entropy [34].

More generally, G causality is typically implemented within the well-understood and easily applicable framework of MVAR modeling. This implementation, however, implies many assumptions about how to model the data. Transfer entropy by contrast, although on a theoretical level “model agnostic” (in the sense that it involves no presumptions about the joint statistical distribution of the data), may present severe difficulties in empirical application. Investigators, then, are free to use whichever practical methods best suit their data. Numerical issues aside, the analytical equivalence (13) furnishes the essential point that—under Gaussian assumptions— G causality has a natural interpretation as transfer entropy and vice-versa.

A.K.S. received financial support from the EPSRC Grant No. EP/G007543/1, which also supports the work of A. B. B.

*L.C.Barnett@sussex.ac.uk

†abb22@sussex.ac.uk

‡A.K.Seth@sussex.ac.uk

- [1] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, UK, 1999).
- [2] N. Wiener, in *Modern Mathematics for Engineers*, edited by E. Beckenbach (McGraw-Hill, New York, 1956).
- [3] C. Granger, *Econometrica* **37**, 424 (1969).
- [4] A. K. Seth and G. Edelman, *Neural Comput.* **19**, 910 (2007).
- [5] A. J. Cadotte, T. B. DeMarse, P. He, and M. Ding, *PLoS ONE* **3**, e3355 (2008).
- [6] M. Ding, Y. Chen, and S. Bressler, in *Handbook of Time Series Analysis*, edited by S. Schelter, M. Winterhalder, and J. Timmer (Wiley, Weinheim, 2006), pp. 438–460.
- [7] T. Schreiber, *Phys. Rev. Lett.* **85**, 461 (2000).
- [8] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, *Phys. Rev. E* **63**, 046211 (2001).
- [9] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics, Inference and Relationship Vol. 2* (Griffin, London, 1979).
- [10] This is to be distinguished from the *conditional covariance*, which will in general be a random variable—although later we note that for *Gaussian* variables the notions coincide.
- [11] While our analysis may be extended to *continuous* time we focus here on the discrete time case.
- [12] S. S. Wilks, *Biometrika* **24**, 471 (1932).
- [13] J. Davidson, *Econometric Theory* (Wiley-Blackwell, Oxford, UK, 2000).
- [14] The analysis carries through for the nonstationary case, but for simplicity we assume here that all processes are stationary.
- [15] Equations (6) and (7) are not to be interpreted as specifying actual MVAR processes; indeed, as such they could not be consistent due to the $Y_{t-1}^{(q)}$ dependence in (7). Furthermore, we note that the variables X_t , Y_t , Z_t may depend on *latent* (unknown) or *exogenous* (unmeasured) variables not included in the regressions. Rather, we should view the regressions purely as *predictive models* specified by the parameters A and A' , respectively, which are then to be fitted by an OLS or equivalent procedure.
- [16] Note that even though X and Y are univariate, the *lagged* variables X^- and Y^- will generally be multivariate (at least if $p, q > 1$); hence they are written in bold type.
- [17] C. W. J. Granger, *Inf. Control* **6**, 28 (1963).
- [18] P. Whittle, *J. R. Stat. Soc. Ser. B* **15**, 125 (1953).
- [19] J. Geweke, *J. Am. Stat. Assoc.* **77**, 304 (1982).
- [20] A. Wald, *Trans. Am. Math. Soc.* **54**, 426 (1943).
- [21] C. Ladrone, S. Guo, K. Kendrick, and J. Feng, *PLoS ONE* **4**, e6899 (2009).
- [22] L. Barnett, A. B. Barrett, and A. K. Seth (unpublished).
- [23] A. Kaiser and T. Schreiber, *Physica (Amsterdam)* **166D**, 43 (2002).
- [24] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 2002), 4th ed.
- [25] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, UK, 1985).
- [26] We note that for jointly multivariate Gaussian variables X , Y a standard result states that the covariance matrix of X given $Y = y$ does not depend on the value of y , so that the *conditional covariance* $\text{cov}(X|Y)$ is a well-defined (non-random) covariance matrix, which is just the partial covariance $\Sigma(X|Y)$ of (1).
- [27] This is essentially a multivariate, conditional version of the formula given in [23], Eq. (19).
- [28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability Vol. 26 (Chapman & Hall, London, 1986).
- [29] A. Kraskov, H. Stögbauer, and P. Grassberger, *Phys. Rev. E* **69**, 066138 (2004).
- [30] S. Frenzel and B. Pompe, *Phys. Rev. Lett.* **99**, 204101 (2007).
- [31] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, *Phys. Lett. A* **324**, 26 (2004).
- [32] N. Ancona, D. Marinazzo, and S. Stramaglia, *Phys. Rev. E* **70**, 056221 (2004).
- [33] O. Sporns and M. Lungarella, *PLoS Comp. Biol.* **2**, e144 (2006).
- [34] D. Marinazzo, M. Pellicoro, and S. Stramaglia, *Phys. Rev. Lett.* **100**, 144103 (2008).