Non-parametric Wiener-Granger Causality in Partially Observed Systems

Mehrdad Jafari-Mamaghani
Division of Mathematical Statistics, Department of Mathematics, Stockholm University, Stockholm, Sweden
E-mail: mjm@math.su.se

***Abstract*—Wiener's definition of causality, commonly known as Wiener-Granger causality, has become a frequently used quantification of temporally resolved causality in numerous fields of science. In many empirical studies, the system of interest cannot be observed in its entirety and relevant information may reside outside of the sampled observations. To this end, partial Wiener-Granger causality has been developed to circumvent this issue. In this paper, we extend partial Wiener-Granger causality to the non-parametric case and discuss different approaches to estimate it.**

***Keywords*—Wiener-Granger causality, Conditional mutual information, Kernel canonical correlations.**

I. Introduction

The 20th century has witnessed series of attempts to define causality in a quantitative manner. Here, we will focus on Norbert Wiener's general concept of causality established in 1956 [1]:

> *For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.*

Motivated by inquiries in econometric research, Clive W.J. Granger, a post-doctoral student at Princeton in the early 1960s, formalized the idea above to a definition that has become known as Wiener-Granger causality (WGC) [2], [3]. Granger acknowledged Wiener's contribution during his keynote speech as the recipient of the Nobel Memorial Prize in Economic Sciences in 2003 [4]. During the same speech Granger observed the two concepts underpinning the definition of Wiener-Granger causality: the cause temporally precedes the effect, and the cause contains unique information about the effect.

Since its introduction, WGC has been employed in a wide array of scientific disciplines. Particularly, applications of WGC in fields such as econometrics [5], finance [6], neurophysiology [7], climatology [8], and cell biology [9], have helped to elucidate systems of causal interaction in these fields, and furthermore led to a rich spectrum of developments on measures and analytical frameworks of WGC [10], [11].

Common frameworks for analysis of WGC include linear models, utilizing e.g. linear regression, and non-parametric models, based on e.g. functions of Shannon entropy. The choice of model is usually a consequence of properties of sampled data and assumptions on complexity of causal interactions. In many real-world applications however, the sampled observations do not represent the system of interest in a complete manner as a result of practical limitations. In neurophysiology, this concern has lead to the introduction of *partial* Wiener-Granger causality based on models of linear regression [12].

The aim of this paper is to introduce a corresponding non-parametric multivariate measure of partial Wiener-Granger causality. Furthermore, we will use definitions from information theory and machine learning to present different approaches for estimating non-parametric partial WGC.

II. Methods

A. Wiener-Granger Causality

Let us define a temporally resolved system from which we have observed three continuous stochastic processes $\{X\}_{-\infty}^{t}$, $\{Y\}_{-\infty}^{t}$ and $\{Z\}_{-\infty}^{t}$, where t denotes the present state of the processes. We say that $\{Y\}$ Wiener-Granger causes $\{X\}$ if we can reject the hypothesis:

$$H_0 : X_t \perp\!\!\!\perp Y_{t-1}, \ldots, Y_{t-k} | X_{t-1}, \ldots, X_{t-k}, Z_{t-1}, \ldots, Z_{t-k}, \tag{1}$$

where k denotes the number of included lags in the model, and $\perp\!\!\!\perp$ stands for probabilistic independence. As a result, in the absence of Wiener-Granger causality under the null hypothesis, the following conditional probability density functions are equal [13]:

$$\begin{aligned} &f(X_t | X_{t-1}, \ldots, X_{t-k}, Z_{t-1}, \ldots, Z_{t-k}) = \\ &f(X_t | Y_{t-1}, \ldots, Y_{t-k}, X_{t-1}, \ldots, X_{t-k}, Z_{t-1}, \ldots, Z_{t-k}). \end{aligned} \tag{2}$$

A common approach to test the hypothesis in (1), or the equality in (2), goes via the employment of linear regression techniques and measures of residual variation, such as the F-test [3], or Geweke's variance test [14]. Of these, we will use the latter in the remainder of this paper.

In order to define the linear measures, consider the following

IEEE 2014 Conference on Norbert Wiener in the 21st Century
Manuscript received May 11, 2014.
978-1-4799-4562-7/14/$31.00 2014 IEEE.

regression models:

$$H_0 : X_t = \alpha + \mathbf{X}^- \cdot A_1 + \mathbf{Z}^- \cdot A_2 + \epsilon_t \quad (3)$$

$$H_A : X_t = \beta + \mathbf{X}^- \cdot B_1 + \mathbf{Z}^- \cdot B_2 + \mathbf{Y}^- \cdot B_3 + \eta_t, \quad (4)$$

where A and B are the regression parameter matrices, the vectors ϵ and η denote Gaussian residual vectors with constant variance, and $\mathbf{X}^- = \{X_{t-1}, \ldots, X_{t-k}\}$. Definitions for the multivariate random vectors $\mathbf{Y}^-$ and $\mathbf{Z}^-$ follow by analogy. The hypothetical contribution of $\mathbf{Y}^-$ to the prediction of X_t given the information in $\mathbf{X}^-$ and $\mathbf{Z}^-$ can be measured using the test statistic [14]:

$$\mathcal{G}_{Y \to X|Z} = \log\left(\frac{\mathrm{Var}(\epsilon_t)}{\mathrm{Var}(\eta_t)}\right), \quad (5)$$

which is χ^2-distributed under the null hypothesis, and follows a non-central χ^2 distribution under the alternate hypothesis. The measure in (5) can be generalized to the multivariate case, i.e. multivariate effect/response variable in the regression models, by using the *total* variance or the *generalized* variance. Here, we will focus on the latter.

B. Partial Wiener-Granger Causality

Let us now assume that we have only observed part of a system that is comprised of more than three stochastic processes. In other words, there are other latent processes that can influence the causal interactions between the observed variables. Given this background, we are interested in re-framing WGC so that we can account for the incomplete nature of our sampled observations. This problem arises in many real-world situations and in neurophysiology has lead to the introduction of partial Wiener-Granger causality [12]. In the same spirit as the previous section, consider the following linear models:

$$Z_t = \widetilde{\alpha} + \mathbf{X}^- \cdot \tilde{A}_1 + \mathbf{Z}^- \cdot \tilde{A}_2 + \tilde{\epsilon}_t \quad (6)$$

$$Z_t = \widetilde{\beta} + \mathbf{X}^- \cdot \tilde{B}_1 + \mathbf{Z}^- \cdot \tilde{B}_2 + \mathbf{Y}^- \cdot \tilde{B}_3 + \tilde{\eta}_t, \quad (7)$$

where $\tilde{A}$ and $\tilde{B}$ represent the regression parameter matrices, and similar to (6) and (7), the residual vectors $\tilde{\epsilon}$ and $\tilde{\eta}$ follow Gaussian distributions with constant variance.

Now, let us decompose the residual vector in model (6) according to the following scheme:

$$\tilde{\epsilon}_t = \tilde{\epsilon}_t^E + f(\tilde{\epsilon}_t^L) + \tilde{\epsilon}_t^0, \quad (8)$$

where $\tilde{\epsilon}_t^E$ denotes the influence of exogenous inputs (environmental influence), $f(\tilde{\epsilon}_t^L)$ represents a function of influences from latent (unobserved) variables, and $\tilde{\epsilon}_t^0$ stands for noise caused by other factors. Decomposition of the residual vectors in models (3), (4) and (7) follows by analogy.

To define a measure of partial WGC, which accounts for the existence of exogenous inputs and latent variables, we have to modify the variances used in (5) so that the influences of $\tilde{\epsilon}_t$ and $\tilde{\eta}_t$ are removed from ϵ_t and η_t, respectively:

$$\mathcal{G}^{(\mathrm{partial})}_{Y \to X|Z} = \log\left(\frac{\mathrm{Var}(\epsilon_t|\tilde{\epsilon}_t)}{\mathrm{Var}(\eta_t|\tilde{\eta}_t)}\right). \quad (9)$$

Given Gaussian variables, the expression in (9) can be evaluated as:

$$\mathcal{G}^{(\mathrm{partial})}_{Y \to X|Z} = \log\left(\frac{\mathrm{Var}(\epsilon_t) - \mathrm{Cov}(\epsilon_t, \tilde{\epsilon}_t)\,\mathrm{Var}(\tilde{\epsilon}_t)^{-1}\,\mathrm{Cov}(\tilde{\epsilon}_t, \epsilon_t)}{\mathrm{Var}(\eta_t) - \mathrm{Cov}(\eta_t, \tilde{\eta}_t)\,\mathrm{Var}(\tilde{\eta}_t)^{-1}\,\mathrm{Cov}(\tilde{\eta}_t, \eta_t)}\right). \quad (10)$$

However, as the assumption of Gaussianity is scarcely met in real-world applications, in the next section we will present a non-parametric framework to estimate partial WGC.

III. Results

A. Partial Conditional Mutual Information

Information theory has been the dominant contributor to non-parametric frameworks of analysis in the context of WGC. Specifically, transfer entropy, introduced in 2000 by Thomas Schreiber [15], has become a frequently used measure in applications of WGC where linear models are believed to be inadequate. It has been shown that, under certain model specifications, transfer entropy, conditional mutual information, and the Jensen-Shannon divergence can be regarded as identical measures [16]. With this background, and given the fact that the term 'partial transfer entropy' has already been used to refer to a case of conditional transfer entropy [17], we will henceforth use the terms conditional mutual information and *partial* conditional mutual information without any loss of generality.

Following the same notational routines as the previous section, we define the *partial* conditional mutual information (PCMI) as:

$$\mathcal{P}_{Y \to X|Z} = H(X_t|\mathbf{X}^-, \mathbf{Z}^-, Z_t) - H(X_t|\mathbf{X}^-, \mathbf{Z}^-, \mathbf{Y}^-, Z_t), \quad (11)$$

where the terms on the right-hand side denote conditional Shannon entropies. In correspondence to the case for standard WGC, it can be shown that for Gaussian variables (see Appendix A):

$$\mathcal{G}^{(\mathrm{partial})}_{Y \to X|Z} = 2 \cdot \mathcal{P}_{Y \to X|Z}. \quad (12)$$

Given that $H(X|Y) = H(X,Y) - H(Y)$, the measure in (11) can be estimated using estimates of differential Shannon entropy:

$$\widehat{H}(X) = -\int_{\mathcal{X}} \hat{f}(X) \cdot \log \hat{f}(X) dx, \quad (13)$$

where $\mathcal{X}$ denotes the space on which X is defined, and $\hat{f}(X)$ represents an estimate of the probability density function of X.

B. Kernel Density Estimation

A common technique to obtain estimates of probability density functions in (13) goes via kernel density estimation [18] (for a technical account in the context of information theoretical measures of WGC see [16]). Although kernel

density estimation (KDE) has been used extensively in the context of WGC, its utilization comes with a number of issues that can be of potential concern.

Firstly, the choices of kernel function and kernel bandwidth have a paramount impact on the resulting estimates [16]. Specifically, the choice of bandwidth is crucial in determining the bias-variance trade-off in estimates of density functions [19].

Secondly, estimation of entropy-based measures via KDE can be a computationally expensive task for large high-dimensional data samples; an inconvenience that is amplified when conducting tests of significance based on permutation resampling.

Most importantly however, it follows from [20] that conditional mutual information estimated via KDE is less sensitive to non-linear signals; an observation that is counter-intuitive to the non-parametric premise of conditional mutual information.

C. Partial Kernel Canonical Correlations

Motivated by the reasons above, here we will introduce an alternative path to estimate PCMI using kernel canonical correlations.

However, note that in contrast to KDE where the kernel functions are an instrument to create smooth sampling windows, the term kernel in the case of kernel canonical correlations refers to the usage of kernel functions as a means to project data from the input space to a higher-dimensional feature space [21].

Using the Gaussian kernel function $\Phi(X) = \exp(-X^2/2)/\sqrt{2\pi}$, we propose to estimate PCMI via kernel canonical correlations (KCC) using:

$$\mathcal{P}^{(\mathrm{KCC})}_{Y\to X|Z} = -\frac{1}{2}\log\prod_{i=1}^{d}\left[1-\rho_i(X_t,\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)^2\right], \tag{14}$$

where d is the number of kernel canonical correlations, and ρ_i denotes the ith kernel canonical correlation of X_t and $\mathbf{Y}$ when the influence of $\mathbf{X}^-, \mathbf{Z}^-$ and Z_t has been removed. More specifically, the kernel canonical correlations used in (14) are defined as [22]:

$$\rho_i(X_t,\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t) = \underset{\gamma,\delta}{\arg\max}\ \frac{\langle\Phi(X_t|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\gamma,\Phi(\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\delta\rangle}{\Psi(\gamma)^{1/2}\Psi(\delta)^{1/2}}, \tag{15}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product function, $\Phi(\cdot)$ is used to denote the higher-dimensional projections of input data using the Gaussian kernel function, and γ and δ denote the canonical transformation vectors used to maximize between-set correlations (i.e. linear projections used in canonical correlation analysis). Furthermore, the two terms in the denominator, using norm penalization with regularization parameter ζ to counterbalance overfitting, are defined as:

$$\begin{aligned}\Psi(\gamma) =&\langle\Phi(X_t|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\gamma,\Phi(X_t|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\gamma\rangle \\ &+\zeta\|\langle\gamma,\Phi(X_t|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\gamma\rangle\|^2, \qquad (16)\\ \Psi(\delta) =&\langle\Phi(\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\delta,\Phi(\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\delta\rangle \\ &+\zeta\|\langle\delta,\Phi(\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)\delta\rangle\|^2. \qquad (17)\end{aligned}$$

The canonical correlations in (15) are obtained as solutions to the generalized eigenvalue problem formulated by partial derivatives of (15) with respect to γ and δ (see [22] for thorough details).

1) PCMI estimated via KCC: The expression in (14) presents a considerably more efficient estimate for PCMI. The relationship between PCMI and KCC is supported by the link between mutual information and kernel generalized variance [23]. In the context of WGC, this topic is discussed further in [22], and a more elaborate examination of this link in a more general sense is given in [23].

As for its properties, the measure in (14) offers a number of appealing advantages in real-world applications. Firstly, as the Gaussian kernel function used in the evaluation of (14) qualifies as a *Mercer* kernel [24], the resulting images of data in the higher-dimensional feature space will qualify as positive-semidefinite *Gram* matrices, which as a result are amenable to matrix factorization using e.g. Cholesky decomposition. This property allows for substantially enhanced computational efficiency in estimations and tests of KCC as the Gram matrices can be decomposed to efficient low-rank matrices using e.g. *incomplete* Cholesky decomposition [25]. In other words, by tuning the size of inner product matrices used in the evaluation of KCC in (15) via incomplete matrix factorization, we are able to control the main bottleneck in computation of KCC and corresponding tests of significance based on permutation resampling.

Moreover, as the measure in (14) is founded on canonical information and regularized to counter overfitting, it has been shown to outperform KDE-estimated transfer entropy in large high-dimensional settings and in presence of noise [22].

IV. Conclusion

Wiener's notion of causality offers a powerful framework to test a general type of time-resolved causality and has reached a vast audience in many areas of quantitative sciences.

In many real-world situations, given a specified question, the sampled observations do not comprise the entirety of the system of interest, and may be driven by factors that are not observable. These concerns have lead to the development of partial WGC using linear models, which in this paper have been extended to the non-parametric case by presenting partial conditional mutual information.

Furthermore, we have proposed to estimate our non-parametric measure of partial WGC, using kernel canonical correlations as a result of its superior performance over other estimates based on density estimation. We believe that this approach offers a number of appealing properties in terms of computational efficiency and enhanced sensitivity to non-linear

and non-monotonous signals.

As a final remark, it should be noted that in empirical analysis of WGC, the choice of model has to be in concert with the correct understanding of the problem, and quality and size of data. Moreover, any discovery of statistically significant causal links has to be interpreted in the specific context of the observed system and sampled data.

APPENDIX

In this section, we will look closer at the multivariate extensions of (5) and (9), and some resulting properties.

The test statistic for standard WGC in (5) can be generalized to the multivariate case using the *generalized* variance, the determinant of the covariance matrix:

$$\mathcal{G}_{Y\to X|Z} = \log\left(\frac{|\Sigma(\epsilon_t)|}{|\Sigma(\eta_t)|}\right) \tag{18}$$

$$= \log\left(\frac{|\Sigma(X_t|\mathbf{X}^-,\mathbf{Z}^-)|}{|\Sigma(X_t|\mathbf{X}^-,\mathbf{Y}^-,\mathbf{Z}^-)|}\right) \tag{19}$$

$$= \log\left(\frac{|\Sigma(X_t,\mathbf{X}^-,\mathbf{Z}^-)|\cdot|\Sigma(\mathbf{X}^-,\mathbf{Y}^-,\mathbf{Z}^-)|}{|\Sigma(X_t,\mathbf{X}^-,\mathbf{Y}^-,\mathbf{Z}^-)|\cdot|\Sigma(\mathbf{X}^-,\mathbf{Z}^-)|}\right), \tag{20}$$

where $\Sigma(\cdot)$ denotes the covariance function, and where the second equality stems from the solution for Yule-Walker equations under the null hypothesis, and the last equality follows from the block determinant identity [26].

1) Multivariate Partial WGC, Case I*:* Analogously, generalizing $\mathrm{Var}(\epsilon_t|\tilde{\epsilon}_t)$ to the multivariate case using $|\Sigma(\epsilon_t|\tilde{\epsilon}_t)|$, the corresponding extension of (9) to the multivariate case uses the equalities [27]:

$$\Sigma(\epsilon_t|\tilde{\epsilon}_t) = \Sigma(X_t|\mathbf{X}^-,\mathbf{Z}^-,Z_t), \tag{21}$$

$$\Sigma(\eta_t|\tilde{\eta}_t) = \Sigma(X_t|\mathbf{X}^-,\mathbf{Y}^-,\mathbf{Z}^-,Z_t). \tag{22}$$

Let us for the sake of convenience use the following substitutions: $\mathbf{V} = \{\mathbf{X}^-,\mathbf{Z}^-\}$ and $\mathbf{W} = \{\mathbf{X}^-,\mathbf{Y}^-,\mathbf{Z}^-\}$. As a result, the linear multivariate measure for partial WGC is defined as:

$$\mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} = \log\left(\frac{|\Sigma(X_t|\mathbf{V},Z_t)|}{|\Sigma(X_t|\mathbf{W},Z_t)|}\right) \tag{23}$$

$$= \log\left(\frac{|\Sigma(X_t,\mathbf{V},Z_t)|\cdot|\Sigma(\mathbf{W},Z_t)|}{|\Sigma(X_t,\mathbf{W},Z_t)|\cdot|\Sigma(\mathbf{V},Z_t)|}\right). \tag{24}$$

Furthermore, given Gaussian random vectors, we have the following expression for the Shannon entropy:

$$H^{(\mathrm{Gauss})}(\mathbf{X}) = \frac{1}{2}\log\left(|2\pi e\cdot\Sigma(\mathbf{X})|\right). \tag{25}$$

Analogously from the case for standard WGC, it follows that:

$$\mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} = 2\cdot\mathcal{P}_{Y\to X|Z}. \tag{26}$$

2) Multivariate Partial WGC, Case II*:* As an alternative approach, again given Gaussian-distributed variables, we can reformulate (9) as:

$$\mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} = \log\left(\frac{|\Sigma(\epsilon_t|\tilde{\epsilon}_t)|}{|\Sigma(\eta_t|\tilde{\eta}_t)|}\right) \tag{27}$$

$$= \log\left(\frac{|\Sigma\big([X_t|\mathbf{V}]\,|\,[Z_t|\mathbf{V}]\big)|}{|\Sigma\big([X_t|\mathbf{W}]\,|\,[Z_t|\mathbf{W}]\big)|}\right) \tag{28}$$

$$= \log\left(\frac{|\Sigma\left(X_t|\mathbf{V},Z_t|\mathbf{V}\right)|\cdot|\Sigma\left(Z_t|\mathbf{W}\right)|}{|\Sigma\left(X_t|\mathbf{W},Z_t|\mathbf{W}\right)|\cdot|\Sigma\left(Z_t|\mathbf{V}\right)|}\right), \tag{29}$$

By using the expression in (25), we can write (29) in terms of Shannon entropies:

$$\mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} = 2\cdot\big(H(X_t|\mathbf{V},Z_t|\mathbf{V}) + H(Z_t|\mathbf{W}) - H(X_t|\mathbf{W},Z_t|\mathbf{W}) - H(Z_t|\mathbf{V})\big), \tag{30}$$

where the joint entropy of two conditional random processes, using the definition of mutual information $I(X;Y) = H(X) - H(X,Y)$, can be rewritten as:

$$\begin{aligned} H(X_t|\mathbf{V},Z_t|\mathbf{V}) =& H(X_t|\mathbf{V}) + H(Z_t|\mathbf{V}) - I(X_t|\mathbf{V};Z_t|\mathbf{V}) \\ =& H(X_t,\mathbf{V}) + H(Z_t,\mathbf{V}) - I(X_t|\mathbf{V};Z_t|\mathbf{V}) \\ & - 2\cdot H(\mathbf{V}) \\ =& H(X_t,Z_t,\mathbf{V}) + I(X_t|\mathbf{V};Z_t|\mathbf{V}) + H(\mathbf{V}) \\ & - I(X_t|\mathbf{V};Z_t|\mathbf{V}) - 2\cdot H(\mathbf{V}) \\ =& H(X_t,Z_t,\mathbf{V}) - H(\mathbf{V}). \end{aligned} \tag{31}$$

Inserting the result from (31) in (30) and using the identity $H(X|Y) = H(X,Y) - H(Y)$ gives us:

$$\begin{aligned} \mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} &= 2\cdot\big(H(X_t,Z_t,\mathbf{V}) - H(\mathbf{V}) + H(Z_t|\mathbf{W}) \\ &\quad - H(X_t,Z_t,\mathbf{W}) + H(\mathbf{W}) - H(Z_t|\mathbf{V})\big) \\ &= 2\cdot\big(H(X_t,Z_t,\mathbf{V}) - H(\mathbf{V}) + H(Z_t,\mathbf{W}) \\ &\quad - H(\mathbf{W}) - H(X_t,Z_t,\mathbf{W}) + H(\mathbf{W}) \\ &\quad - H(Z_t,\mathbf{V}) + H(\mathbf{V})\big) \\ &= 2\cdot\big(H(X_t,Z_t,\mathbf{V}) + H(Z_t,\mathbf{W}) - H(X_t,Z_t,\mathbf{W}) \\ &\quad - H(Z_t,\mathbf{V})\big) \\ &= 2\cdot\big(H(X_t|Z_t,\mathbf{V}) - H(X_t|Z_t,\mathbf{W})\big) \\ &= 2\cdot\ \mathcal{P}_{Y\to X|Z}\ , \end{aligned}$$

in accordance with the equality in (26) and the definition in (11).

3) Multivariate Partial WGC, Case III*:* Lastly, under the same assumption of Gaussian-distributed variables, and vectors X_t and $\mathbf{Y}^-$ of same dimensionality d, it can be shown that:

$$\mathcal{G}^{(\mathrm{partial})}_{Y\to X|Z} = -\log\prod_{i=1}^{d}\left[1 - r_i(X_t,\mathbf{Y}^-|\mathbf{X}^-,\mathbf{Z}^-,Z_t)^2\right], \tag{32}$$

where r_i denotes the ith canonical correlation of X_t and $\mathbf{Y}^-$ when the influence of $\{\mathbf{X}^-,\mathbf{Z}^-,Z_t\}$ has been removed [22].

ACKNOWLEDGMENT

Support by the Magnusson Fund of the Royal Swedish Academy of Sciences is gratefully acknowledged.

REFERENCES

[1] N. Wiener, "The theory of prediction," *Modern mathematics for engineers*, pp. 165–190, 1956.

[2] C. W. Granger, "Economic processes involving feedback," *Information and Control*, vol. 6, no. 1, pp. 28–48, 1963.

[3] ——, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–38, 1969.

[4] ——, "Time series analysis, cointegration, and applications," *Nobel Lecture in Stockholm*, 2004.

[5] C. A. Sims, "Money, Income, and Causality," *Amer Econ Rev*, vol. 62, no. 4, pp. 540–552, 1972.

[6] A. Zaremba and T. Aste, "Measures of Causality in Complex Datasets with Application to Financial Data," *arXiv preprint arXiv:1401.1457*, 2014.

[7] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[8] J. Kang and R. Larsson, "What is the link between temperature and carbon dioxide levels? A Granger causality analysis based on ice core data," *Theoretical and Applied Climatology*, pp. 1–12, 2013.

[9] J. G. Lock, M. Jafari-Mamaghani, H. Shafqat-Abbasi, X. Gong, J. Tyrcha, and S. Strömblad, "Plasticity in the Macromolecular-Scale Causal Networks of Cell Migration," *PLOS ONE*, vol. 9, no. 2, p. e90593, 2014.

[10] S. Guo, C. Ladroue, and J. Feng, "Granger causality: Theory and applications," in *Frontiers in Computational and Systems Biology*, ser. Computational Biology, J. Feng, W. Fu, and F. Sun, Eds. Springer London, 2010, vol. 15, pp. 83–111.

[11] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[12] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng, "Partial Granger causalityeliminating exogenous inputs and latent variables," *Journal of neuroscience methods*, vol. 172, no. 1, pp. 79–93, 2008.

[13] J.-P. Florens and M. Mouchart, "A note on noncausality," *Econometrica: Journal of the Econometric Society*, pp. 583–591, 1982.

[14] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, p. 304, 1982.

[15] T. Schreiber, "Measuring Information Transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[16] M. Jafari-Mamaghani, "Non-parametric Analysis of Granger Causality Using Local Measures of Divergence," *Applied Mathematical Sciences*, vol. 7, no. 83, pp. 4107–4136, 2013.

[17] V. A. Vakorin, O. A. Krakovska, and A. R. McIntosh, "Confounding effects of indirect connections on causality estimation," *Journal of neuroscience methods*, vol. 184, no. 1, pp. 152–160, 2009.

[18] M. P. Wand and M. C. Jones, *Kernel smoothing*. Crc Press, 1994, vol. 60.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, corrected ed. Springer, Aug. 2009.

[20] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information." *Physical review. E*, vol. 69, no. 6 Pt 2, Jun. 2004.

[21] B. Schölkopf and A. J. Smola, *Learning with kernels*. The MIT Press, 2002.

[22] M. Jafari-Mamaghani, "Robust Non-linear Wiener-Granger Causality For Large High-dimensional Data," *Submitted manuscript*, 2014.

[23] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *The Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2003.

[24] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. The MIT Press, 2012.

[25] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *The Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2002.

[26] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables," *Phys. Rev. Lett.*, vol. 103, p. 238701, Dec 2009.

[27] A. B. Barrett, L. Barnett, and A. K. Seth, "Multivariate Granger causality and generalized variance," *Physical Review E*, vol. 81, no. 4, p. 041907, 2010.