JMLR: Workshop and Conference Proceedings 12 (2011) 30–64

Robust Statistics for Describing Causality in Multivariate Time Series.

Florin Popescu FLORIN.POPESCU@FIRST.FRAUNHOFER.DE
Fraunhofer Institute FIRST
Kekulestr. 7, Berlin 12489 Germany

Editor(s): Florin Popescu and Isabelle Guyon

Abstract

A widely agreed upon definition of time series causality inference, established in the seminal 1969 article of Clive Granger (1969), is based on the relative ability of the history of one time series to predict the current state of another, conditional on all other past information. While the Granger Causality (GC) principle remains uncontested, its literal application is challenged by practical and physical limitations of the process of discretely sampling continuous dynamic systems. Advances in methodology for time-series causality subsequently evolved mainly in econometrics and brain imaging: while each domain has specific data and noise characteristics the basic aims and challenges are similar. Dynamic interactions may occur at higher temporal or spatial resolution than our ability to measure them, which leads to the potentially false inference of causation where only correlation is present. Causality assignment can be seen as the principled partition of spectral coherence among interacting signals using both auto-regressive (AR) modeling and spectral decomposition. While both approaches are theoretically equivalent, interchangeably describing linear dynamic processes, the purely spectral approach currently differs in its somewhat higher ability to accurately deal with mixed additive noise.

Two new methods are introduced 1) a purely auto-regressive method named Causal Structural Information is introduced which unlike current AR-based methods is robust to mixed additive noise and 2) a novel means of calculating multivariate spectra for unevenly sampled data based on cardinal trigonometric functions is incorporated into the recently introduced phase slope index (PSI) spectral causal inference method (Nolte et al., 2008). In addition to these, PSI, partial coherence-based PSI and existing AR-based causality measures were tested on a specially constructed data-set simulating possible confounding effects of mixed noise and another additionally testing the influence of common, background driving signals. Tabulated statistics are provided in which true causality influence is subjected to an acceptable level of false inference probability.

Keywords: Causality, spectral decomposition, cross-correlation, auto regressive models.

1. Introduction

Causality is the *sine qua non* of scientific inference methodology, allowing us, among other things to advocate effective policy, diagnose and cure disease and explain brain function. While it has recently attracted much interest within Machine Learning, it bears reminding that a lot of this recent effort has been directed toward *static* data rather than time series. The 'classical' statisticians of the early 20^{th} century, such as Fisher, Gosset and Karl Pearson, aimed at a rational and general recipe for causal inference and discovery

(Gigerenzer et al., 1990) but the tools they developed applied to simple types of inference which required the pres-selection, through consensus or by design, of a handful of candidate causes (or 'treatments') and a handful of subsequently occurring candidate effects. Numerical experiments yielded tables which were intended to serve as a technician's almanac (Pearson, 1930; Fisher, 1925), and are today an essential part of the vocabulary of scientific discourse, although tables have been replaced by precise formulae and specialized software. These methods rely on *removing* possible causal links at a certain 'significance level', on the basic premise that a twin experiment on data of similar size generated by a hypothetical non-causal mechanism would yield a result of similar strength only with a known (small) probability. While it may have been hoped that a generalization of the statistical test of difference among population means (e.g. the t-test) to the case of time series causal structure may be possible using a similar almanac or recipe book approach, in reality causality has proven to be a much more contentious - and difficult - issue.

Time series theory and analysis immediately followed the development of classical statistics (Yule, 1926; Wold, 1938) and was spurred thereafter by exigence (a severe economic boom/bust cycle, an intense high-tech global conflict) as well as opportunity (the post-war advent of a machine able to perform large linear algebra calculations). From a wide historical perspective, Fisher's 'almanac' has rendered the industrial age more orderly and understandable. It can be argued, however, that the 'scientific method', at least in its accounting/statistical aspects, has not kept up with the explosive growth of data tabulated in history, geology, neuroscience, medicine, population dynamics, economics, finance and other fields in which causal structure is at best partially known and understood, but is needed in order to cure or to advocate policy. While it may have been hoped that the advent of the computer might give rise to an automatic inference machine able to 'sort out' the ever-expanding data sphere, the potential of a computer of any conceivable power to condense the world to predictable patterns has long been proven to be shockingly limited by mathematicians such as Turing (Turing, 1936) and Kolmogorov (Kolmogorov and Shiryayev, 1992) - even before the ENIAC was built. The basic problem reduces itself to the curse of dimensionality: being forced to choose among combinations of members of a large set of hypotheses (Lanterman, 2001). Scientists as a whole took a more positive outlook, in line with post-war boom optimism, and focused on accessible automatic inference problems. One of these was scientists was Norbert Wiener, who, besides founding the field of cybernetics (the precursor of ML), introduced some of the basic tools of modern time-series analysis, a line of research he began during wartime and focused on feedback control in ballistics. The time-series causality definition of Granger (1969) owes inspiration to earlier discussion of causality by Wiener (1956). Granger's approach blended spectral analysis with vector auto-regression, which had long been basic tools of economics (Wold, 1938; Koopmans, 1950), and appeared nearly at the same time as similar work by Akaike (1968) and Gersch and Goddard (1970).

It is useful to highlight the differences in methodological principle and in motivation for static *vs.* time series data causality inference, starting with the former as it comprises a large part of the pertinent corpus in Machine Learning and in data mining. Static causal inference is important in the sense that any classification or regression presumes some kind of causality, for the resulting relation to be useful in identifying elements or features of the data which 'cause' or predict target labels or variables and are to be selected at the

exclusion of other confounding 'features'. In learning and generalization of static data, sample ordering is either uninformative or unknown. Yet order is implicitly relevant to learning both in the sense that some calculation occurs in the physical world in some finite number of steps which transform independent inputs (*stimuli*) to dependent output (*responses*), and in the sense that generalization should occur on expected *future stimuli*. To ably generalize from a limited set of samples implies making accurate causal inference. With this priority in mind prior NIPS workshops have concentrated on feature selection and on graphical model-type causal inference (Guyon and Elisseeff, 2003; Guyon et al., 2008, 2010) inspired by the work of Pearl (2000) and Spirtes et al. (2000). The basic technique or underlying principle of this type of inference is vanishing partial correlation or the inference of *static* conditional independence among 3 or more random variables. While it may seem limiting that no unambiguous, generally applicable causality assignment procedure exists among single *pairs* of random variables, for large ensembles the ambiguity may be partially resolved. Statistical tests exist which assign, with a controlled probability of false inference, random variable X_1 as dependent on X_2 given no other information, but as independent on X_2 given X_3, a conceptual framework proposed for time-series causality soon after Granger's 1969 paper using partial *coherence* rather than static correlation (Gersch and Goddard, 1970). Applied to an ensemble of observations $X_1..X_N$, efficient polynomial time algorithms have been devised which combine information about pairs, triples and other sub-ensembles of random variables into a complete dependency graph including, but not limited to, a directed acyclical graph (DAG). Such inference algorithms operate in a nearly deductive manner but are not guaranteed to have unique, optimal solution. Underlying predictive models upon which this type of inference can operate includes linear regression (or structural equation modeling) (Richardson and Spirtes, 1999; Lacerda et al., 2008; Pearl, 2000) and Markov chain probabilistic models (Scheines et al., 1998; Spirtes et al., 2000). Importantly, a previously unclear conceptual link between the notions of time series causality and static causal inference has been formally described: see White and Lu (2010) in this volume.

Likewise, algorithmic and functional relation constraints, or at least likelihoods thereof, have been proposed as to assign causality for co-observed random variable pairs (i.e. simply by analyzing the scatter plot of X_1 vs. X_2) (Hoyer et al., 2009). In general terms, if we are presented a scatter plot X_1 vs. X_2 which looks like a noisy sine wave, we may reasonably infer that X_2 causes X_1, since a given value of X_2 'determines' X_1 and not vice versa. We may even make some mild assumptions about the noise process which superimposes on a functional relation ($X_2 = X_1$ + additive noise which is independent of X_1) and by this means turn our intuition into a proper *asymmetric* statistic, i.e. a controlled probability that X_1 does *not* determine X_2, an approach that has proven remarkably successful in some cases where the presence of a causal relation was known but the direction was not (Hoyer et al., 2009). The challenge here is that, unlike in traditional statistics, there is not simply the case of the null hypothesis and its converse, but one of 4 mutually exclusive cases. A) X_1 is independent of X_2 B) X_1 causes X_2 C) X_2 causes X_1 and D) X_1 and X_2 are observations of dependent and non-causally related random variables (bidirectional information flow or feedback). The appearance of a symmetric bijection (with additive noise) between X_1 and X_2 does not mean absence of causal relation, as asymmetry in the apparent relations is merely a clue and not a determinant of causality. Inference over static data is not without ambiguities without additional assumptions and requires observations of interacting triples

(or more) of variables as to allow somewhat reliable descriptions of causal relations or lack thereof (see Guyon et al. (2010) for a more comprehensive overview). Statistical evaluation requires estimation of relative likelihood of various candidate models or causal structures, including a null hypothesis of non-causality. In the case of complex multidimensional data theoretical derivation of such probabilities is quite difficult, since it is hard to analytically describe the class of dynamic systems we may be expected to encounter. Instead, common ML practice consists in running toy experiments in which the 'ground truth' (in our case, causal structure) is only known to those who run the experiment, while other scientists aim to test their discovery algorithms on such data, and methodological validity (including error rate) of any candidate method rests on its ability to predict responses to a set of 'stimuli' (test data samples) available only to the scientists organizing the challenge. This is the underlying paradigm of the Causality Workbench (Guyon, 2011). In time series causality, we fortunately have far more information at our disposal relevant to causality than in the static case. Any type of reasonable interpretation of causality implies a physical mechanism which accepts a modifiable input and performs some operations in some finite time which then produce an output and includes a source of randomness which gives it a stochastic nature, be it inherent to the mechanism itself or in the observation process. Intuitively, the structure or connectivity among input-output blocks that govern a data generating process are related to causality no matter (within limits) what the exact input-output relationships are: this is what we mean by structural causality. However, not all structures of data generating processes are obviously causal, nor is it self evident how structure corresponds to Granger (*non*) causality (GC), as shown in further detail by White and Lu (2010). Granger causality is a measure of relative predictive information among variables and not evidence of a direct physical mechanism linking the two processes: no amount of analysis can exclude a latent unobserved cause. Strictly speaking the GC statistic is not a measure of causal relation: it is the possible non-rejection of a null hypothesis of time-ordered independence.

Although time information helps solve many of the ambiguities of static data several problems, and despite the large body of literature on time-series modeling, several problems in time-series causality remain vexing. Knowledge of the structure of the overall multivariate data generating process is an indispensable aid to inferring causal relationships: but how to infer the structure using weak *a priori* assumptions is an open research question. Sections 3, 4 and 5 will address this issue. Even in the simplest case (the bivariate case) the observation process can introduce errors in time-series causal inference by means of co-variate observation noise (Nolte et al., 2010). The bivariate dataset NOISE in the Causality Workbench addresses this case, and is extended in this study to the evaluation datasets PAIRS and TRIPLES. Two new methods are introduced: an autoregressive method named Causal Structural Information (Section 7) and a method for estimating spectral coherence in the case of unevenly sampled data (Section 8.1). A principled comparison of different methods as well as their performance in terms of type I, II and III errors is necessary, which addresses both the presence/absence of causal interaction and directionality. In discussing causal influence in real-world processes, we may reasonably expect that not inferring a potentially weak causal link may be acceptable but positing one where none is missing may be problematic. Sections 2, 6, 7 and 8 address robustness of bivariate causal inference, introducing a pair of novel methods and evaluating them along with existing ones. Another common source of argument in discussions of causal structure is the case of false inference

by neglecting to condition the proposed causal information on other background variables which may explain the proposed effect equally well. While the description of a general deductive method of causal connectivity in multivariate time series is beyond the scope of this article, Section 9 evaluates numerical and statistical performance in the tri-variate case, using methods such as CSI and partial coherence based PSI which can apply to bivariate interactions conditioned by an arbitrary number of background variables.

2. Causality statistic

Causality inference is subject to a wider class of errors than classical statistics, which tests independence among variables. A general hypothesis evaluation framework can be:

$$\begin{gathered}
\text{Null Hypothesis} = \text{No causal interaction} \quad H_0 = A \perp_C B \,|C \\
\text{Hypothesis 1a} = \text{A } \textit{drives} \text{ B} \quad H_a = A \to B \,|C \\
\text{Hypothesis 1b} = \text{B } \textit{drives} \text{ A} \quad H_b = B \to A \,|C \\
\\
\text{Type I error prob. } \alpha = P\left(\hat{H}_a \text{ or } \hat{H}_b |\, H_0\right) \qquad (1) \\
\\
\text{Type II error prob. } \beta = P\left(\hat{H}_0 |\, H_a \text{ or } H_b\right) \\
\text{Type III error prob. } \gamma = P\left(\, \hat{H}_a \,|H_b \text{ or } \hat{H}_b \,|H_a\right)
\end{gathered}$$

The notation $\hat{H}$ means that our statistical estimate of the estimated likelihood of H exceeds the threshold needed for our decision to confirm it. This formulation carries some caveats the justification for which is pragmatic and will be expounded upon in later sections. The main one is the use of the term '*drives*' in place of '*causes*'. The null hypothesis can be viewed as equivalent to *strong* Granger non-causality (as it will be argued is necessary), but it does not mean that the signals **A** and **B** are independent: they may well be correlated to one another. Furthermore, we cannot realistically aim at statistically supporting *strict* Granger causality, i.e. strictly one-sided causal interaction, since asymmetry in bidirectional interaction may be more likely in real-world observations and is equally meaningful. By '*driving*' we mean instead that the history of one time series element **A** is more useful to predicting the current state of **B** than vice-versa, and not that the history of **B** is irrelevant to predicting **A**. In the latter case we would specify '*G-causes*' instead of '*drives*' and for H_0 we would employ non-parametric independence tests of Granger non causality (GNC) which have already been developed as in Su and White (2008) and Moneta et al. (2010). Note that the definition in (1) is different from that recently proposed in White and Lu (2010), which goes further than GNC testing to make the point that structural causality inference must also involve a further conditional independence test: Conditional Exogeneity (CE). In simple terms, CE tests whether the innovations process of the potential effect is conditionally independent of the cause (or, by practical consequence, whether the innovations processes are uncorrelated). White and Lu argue that if both GNC and CE fail we ought not make any decision regarding causality, and combine the power of both tests

in a principled manner such that the probability of false causal inference, or non-decision, is controlled. The difference in this study is that the concurrent failure of GNC and CE is *precisely* the difficult situation requiring additional focus and it will be argued that methods that can cope with this situation can also perform well for the case of CE, although they require stronger assumptions. In effect, it is assumed that real-world signals feature a high degree of non-causal correlation, due to aliasing effects as described in the following section, and that strong evidence to the contrary is required, i.e. that non-decision is equivalent to inference of non-causality. The precise meaning of 'driving' will also be made explicit in the description of Causal Structural Information, which is implicitly a proposed definition of H_0. Also different in Definition (1) than in White and Lu is the accounting of potential error in causal *direction* assignment under a framework which forces the practitioner to make such a choice if GNC is rejected.

One of the difficulties of causality inference methodology is that it is difficult to ascertain what true causality in the real world ('ground truth') is for a sufficiently comprehensive class of problems (such that we can reliably gage error probabilities): hence the need for extensive simulation. A clear means of validating a causal hypothesis would be *intervention* Pearl (2000), i.e. modification of the presumed cause, but in instances such as historic and geological data this is not feasible. The basic approach will be to assume a non-informative probability distribution of the degree degree of mixing, or non-causal dynamic interactions, as well as over individual spectra and compile inference error probabilities over a wide class of coupled dynamic systems. In constructing a 'robust causality' statistic there is more than simply null-hypothesis rejection and accurate directionality to consider, however. In scientific practice we are not only interested to know that **A** and **B** are causally related or not, but which is the *main* driver in case of bidirectional coupling, and among a time series vector **A**, **B**, **C**, **D**... it is important to determine which of these factors are the main causes of the target variable, say **A**. The relative effect size and relative causal influence strength, lest the analysis be misused (Ziliak and McCloskey, 2008). The rhetorical and scientific value of effect size in no way devalues the underlying principle of robust statistics and controlled inference error probabilities used to quantify it.

3. Auto-regression and aliasing

A simple multivariate time series model is the multivariate auto-regressive model (abbreviated as MVAR or VAR). It assumes that the data generating process (DGP) that created the observations is a linear dynamic model and, as such, it contains poles only i.e. the numerator of the transfer function between innovations process and observation is a scalar. The more complex auto-regressive moving average model (ARMA) includes zeros as well. Despite the rather stringent assumptions of VAR, a time-series extension of ordinary least squares linear regression, it has been hugely successful in applications from neuroscience to engineering to sociology and economics. Its familiar VAR (or VARX) formulation is:

$$y_i = \sum_{k=1}^{K} A_k y_{i-k} + Bu + w_i \tag{2}$$

Where $\{y_{i,d=1..D}\}$ is a real valued vector of dimension D. Notice the absence of a subscript in the exogenous input term u. This is because a general treatment of exogenous inputs requires a lagged sum, i.e. $\sum_{k=1}^{K} B_k u_{i-k}$. Since exogenous inputs are not explicitly addressed in the following derivations the general linear operator placeholder Bu is used instead and can be re-substituted for subsequent use.

Granger non-causality for this system, expressed in terms of conditional independence, would place a relation among elements of y subject to knowledge of u. If $D = 2$, for all i

$$y_{1,i} \perp y_{2,i-1..i-K} \mid y_{1,i-1..i-K} \tag{3}$$

If the above is true, we would say that y_2 does not finite-order G cause y_1. If the world was made exclusively of linear VARs, it would not be terribly difficult to devise a reliable statistic for G causality. We would, given a sequence of N data points, identify the maximum-likelihood parameters A and B via ordinary least squares (OLS) linear regression after having, via some model selection criterion, determined the order K. Furthermore we would choose another criterion (e.g. test and p-value) which tells us whether any particular coefficient is likely to be statistically indistinguishable from 0, which would correspond to a vanishing partial correlation. If all A's are lower triangular G non-causality is satisfied (in one direction but not the converse). It is however very rare that the physical mechanism we are observing is indeed the embodiment of a VAR, and therefore even in the case in which G non-causality can be safely rejected, it is not likely that the best VAR approximation of the data observed is strictly lower/upper triangular. The necessity of a distinction between strict causality, which has a structural interpretation, and a causality statistic, which does not measure independence in the sense of Granger-non causality, but rather *relative* degree of dependence in both directions among two signals (driving) is most evident in this case. If the VAR in question had very small (and statistically observable) upper triangular elements would a discussion of causality of the observed time series be rendered moot?

One of the most common physical mechanisms which is incompatible with VAR is aliasing, i.e. dynamics which are faster than the (shortest) sampling interval. The standard interpretation of aliasing is the false representation of frequency components of a signal due to sub-Nyquist frequency sampling: in the multivariate time-series case this can also lead to spurious correlations in the observed innovations process (Phillips, 1973). Consider a continuous bivariate VAR of order 1 with Gaussian innovations in which the sampling frequency is several orders of magnitude smaller than the Nyquist frequency. In this case we would observe a covariate time independent Gaussian process since for all practical purposes the information travels 'instantaneously'. In economics, this effect could be due to social interactions or market reactions to news which happen faster than the sampling interval (be it daily, hourly or monthly). In fMRI analysis sub- sampling interval brain dynamics are observed over a relatively slow time convolution process of hemodynamic response of neural activity (for a detailed exposition of causality inference in fMRI see Roebroeck et al. (2011) in this volume). Although 'aliasing' normally refers to temporal aliasing, the same process can occur *spatially*. In neuroscience and in economics the observed variables are summations (dimensionality reductions) of a far larger set of interacting agents, be they individuals or neurons. In electroencephalography (EEG) the propagation of electrical potential from cortical axons arrives via multiple pathways to the same recording location on the scalp: the summation of micrometer scale electric potentials on the scalp at centimeter

scale. Once again there are spurious observable correlations: this is known as the *mixing* problem. Such effects can be modeled, albeit with significant information loss, by the same DGP class which is a superset of VAR and known in econometrics as SVAR (structural vector auto-regression, the time series equivalent of structural equation modeling (SEM), often used in static causality inference (Pearl, 2000)). Another basic problem in dynamic system identification is that we not only discard much information from the world in sampling it, but that our observations are susceptible to additive noise, and that the randomness we see in the data is not entirely the randomness of the mechanism we intend to study. One of the most problematic of additive noise models is *mixed colored noise,* in which there are structured correlations both in time and across elements of the time-series, but not in any causal way: there is only a linear transformation of colored noise, sometimes called mixing, due to spatial aliasing. Mixing may occur due to temporal aliasing in sampling a coupled continuous-variable VAR system. In EEG analysis mixed colored noise models the background electrical activity of the brain. In other domains such as economics, one can imagine the influence of unpredictable events such as natural cataclysms or macroenomic cycles which are not white noise and which are reflect nearly 'instantaneously' but to varying degree in all our measurements. In this case, since each additive noise component is colored (it has temporal auto- correlation), its past helps predict its current value. Since the observation is a linear mixture of noise components, all current observations are correlated, and the past of any component can help predict the current state of any other. In this case, the strict definition of Granger causality would not make practical sense, since this cross-predictability is not meaningful.

It should be noted on this point that the literature contains (sometimes inconsistent) sub-classifications of Granger Causality, such as *weak* and *strong* Granger causality. One definition which is particularly pertinent to this work is that given in Caines (1976) and Solo (2006) and is that strong Granger causality allows instantaneous dependence and that weak Granger causality does not (i.e. it is strictly time ordered). We are aiming in this work at strong Granger causality inference, i.e. one which is robust to aliasing effects such as colored noise. While we should *account* for instantaneous interactions, we do not have to assign causal interpretations to them, since they are symmetric (the cross-correlation of independent mixed signals is symmetric).

4. Auto-regression, learning and Granger Causality

Learning is the process of discovering predictable patterns in the real world, where a 'pattern' is described by an algorithm or an automaton. Besides the object of learning, i.e. the algorithm which we infer and which maps stimuli to responses, we need to consider the algorithm which performs the learning process and outputs the former. The third algorithm we should consider is the algorithm embodied in the real world, which we do not know, which generates the data we observe, and which we hope to be able to recover, or at least approximate. How can we formally describe it? A Data Generating Process (DGP) can be a machine or automaton: an algorithm that performs every operation deterministically in a *finite* number of steps, but which contains an oracle that generates perfectly random numbers. It is sufficient that this oracle generate **1**'s and **0**'s only: all other computable probability distributions can be calculated from it. A DGP contains rational valued param-

eters (rational as to comply with finite computability), in this case the integer K and all elements of the matrices A. Last but not least a DGP specification may limit the set of admissible parameter values and probability distributions of the oracle-generated values. The set of all possible outputs of a DGP corresponds to the set of all probability distributions generated by it over all admissible parameter values, which we shall call the DGP class.

Definition 1 *Let $i \in \mathbb{N}$ and let s_a, s_w, p_w be finite length prefix-free binary strings. Furthermore let $\boldsymbol{y}$ and $\boldsymbol{u}$ be rational valued matrices of size $N \times i$ and $M \times i$, and $\boldsymbol{t}$ be rational valued vector with distinct elements, of length $\boldsymbol{i}$. Let $\boldsymbol{a}$ also be a finite rational valued vector. A Data Generating Process is a quintuple $\{s_a,p_w,T_a,T_w\}$ where T_a, T_w are finite time Turing machines which perform the following operations: Given an input of the incompressible string p_w the machine T_w calculates a rational valued matrix $\boldsymbol{w}$. The machine T_a when given matrices $\boldsymbol{y}$, $\boldsymbol{a}$, $\boldsymbol{u}$, $\boldsymbol{t}$, $\boldsymbol{w}$ and a positive rational Δt outputs a vector y_{i+1} which is assigned for future operations to the time $t_{i+1} = max(t) + \Delta t$*

The definition is somewhat unusual in terms of the definition of stochastic systems as embodiments of Turing machines, but it is quite standard in terms of defining an innovations term w, a probability distribution thereof p_w, a state y, a generating function p_a with parameters a and an exogenous input u. The motivation for using the terminology of algorithmic information theory is to analyse causality assignment as a computational problem. For reasons of finite description and computability our variables are rational, rather than real valued. Notice that there is no real restriction on how the time series is to be generated, recursively or otherwise. The initial condition in case of recursion is implicit, and time is specified as distinct and increasing but otherwise arbitrarily distributed - it does not necessarily grow in constant increments (it is asynchronous). The slight paradox about describing stochastic dynamical systems in algorithmic terms is the necessity of postulating a random number generator (an oracle) which in some ways is our main tool for abstracting the complexity of the real world, but yet is a physical impossibility (since such an oracle would require infinite computational time see Li and Vitanyi (1997) for overview). Also, the Turing machines we consider have finite memory and are time restricted (they implement a predefined maximum number of operations before yielding a default output). Otherwise the rules of algebra (since they perform algebraic operations) apply normally. The *cover* of a DGP can be defined as:

Definition 2 *The cover of a Data Generating Process (DGP) class is the cover of the set of all outputs $\boldsymbol{y}$ that a DGP calculates for each member of the set of admissible parameters $\boldsymbol{a,u,t,w}$ and for each initial condition y_1. Two DGPs are stochastically equivalent if the cover of the set of their possible outputs (for fixed parameters) is the same.*

Let us now attempt to define a Granger Causality statistic in algorithmic terms. Allowing for the notation $j..k = \{j-1, j-2.., k+1, k\}$ if $j > k$ and in reverse order if $j < k$

$$\frac{1}{i}\sum_{j=1}^{i} K(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1}) - K(y_{1,j} \mid y_{2,j-1..1}, y_{1,j-1..1}, u_{j-1..1}) \tag{4}$$

This differs from Equation (3) in two elemental ways: it is not a statement of independence but a number (statistic), namely the average difference (rate) of conditional (or prefix)

Kolmogorov complexity of each point in the presumed effect vector when given both vector histories or just one, and given the exogenous input history. It is a generalized conditional entropy rate, and may be reasonably be normalized as such:

$$\mathcal{F}^{K}_{2\to 1|u} = \frac{1}{i}\sum_{j=1}^{i}\left(1 - \frac{K(y_{1,j} \mid y_{2,j-1..1}, y_{1,j-1..1}, u_{j-1..1})}{K(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1})}\right) \tag{5}$$

which is a fraction ranging from 0 - meaning no influence of y_1 by y_2 - to 1, corresponding to complete determination of y_1 by y_2 and can be transformed into a statistic comparing different data sets and processes, and which gives probabilities of spurious results. Another difference with Equation (3) is that we do not refer to finite-order G causality but simply G causality (in the general case we do not know the maximum lag order but must infer it). For a more in depth look at DGPs, structure and G-causality, see White and Lu (2010). The larger the value $\mathcal{F}^{K}_{2\to 1|u}$, the more likely that y_2 G-causes y_1. The definition is one of conditional information and it is one of an averaged process rather than a single instance (time point). However, Kolmogorov complexity is incomputable, and as such Granger (non) causality must also be, in general, incomputable. A detailed look at this issue is beyond the scope of this article, but in essence, we can never test all possible models that could tell us wether the history of a time series helps or does not help predict (compress) another, and the set of finite running time Turing machines is not enumerable. We've partially circumvented the halting problem since we've specified finite-state, finite-operation machines as the basis of DGPs but have not specified a search procedure over all DGPs that enumerates them. Even if we limit ourselves to DGPs which are MVAR, the necessary computational time to calculate the description length (instead of $K(.)$) is NP-complete, i.e. it requires an enumeration of all possible parameters of a DGP class, barring any special properties thereof: finding the optimal model order requires such a search (keep in mind VAR estimation is convex only once we know the model order and AR structure).

In practice, we should limit the class of DGPs we consider within out statistic to one which allows the possibility of polynomial time computation. Let us take Equation (2), and further make the common assumption that the input vector w is an *i.i.d.* normally distributed sequence independent along dimension d, we've specified the linear VAR Gaussian DGP class (which we shall shorten as VAR class). This DGP class, again, has proven remarkably useful in cases where nothing else except the time series vector y is known. Re-writing (2):

$$y_i = \sum_{k=1}^{K} A_k y_{i-k} + \mathcal{D}\ w_{i-1}\ , \mathcal{D}_{ii} > 0, \mathcal{D}_{ij} = 0 \tag{6}$$

The matrix $\mathcal{D}$ is a positive diagonal matrix containing the scaling, or effective standard deviations of the innovation terms. The standard deviation of each element of the innovations term w **is assumed hereafter to be equal to 1**.

5. Equivalence of auto-regressive data generation processes.

In econometrics the following formulation is familiar (SVAR):

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu + \mathcal{D}w_i \tag{7}$$

The difference between this and Equation (6) is the presence of a 0-lag matrix A_0 which, for easy tractability has zero diagonal entries and is sometimes present on the LHS. This 0-lag matrix is meant to model the sub-sampling interval dynamic interactions among observations, which appear instantaneous, see Moneta et al. (2011) in this volume. Let us call this form *zero lag SVAR*. In electric- and magneto- encephalography (EEG/MEG) we often encounter the following form:

$$x_i = \sum_{k=1}^{K} {}_{\mu}A_k x_{i-k} +_{\mu} Bu + \mathcal{D}w_i,$$

$$y_i = Cx_i \tag{8}$$

Where C represents the observation matrix, or *mixing matrix* and is determined by the conductivity/permeability of tissue, and accounts for the superposition of the electromagnetic fields created by neural activity, which happens at nearly the speed of light and therefore appears instantaneous. Let us call this *mixed output SVAR*. Finally, in certain engineering applications we may see structured disturbances:

$$y_i = \sum_{k=1}^{K} {}_{\theta}A_k y_{i-k} +_{\theta} Bu + D_w w_i \tag{9}$$

Which we shall call *covariate innovations SVAR* (D_w is a general nonsingular matrix unlike $\mathcal{D}$ which is diagonal). Another final SVAR form to consider would be one in which the 0-lag matrix ${}_{\triangleleft}A_0$ is strictly upper triangular (*upper triangular zero lag* SVAR):

$$y_i =_{\triangleleft} A_0 y_i + \sum_{k=1}^{K} A_k y_{i-k} +_{\triangleleft} Bu + \mathcal{D}w_i \tag{10}$$

Finally, we may consider a upper or lower triangular co-variate innovations SVAR:

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu +_{\triangleleft} Dw_i \tag{11}$$

Where ${}_{\triangleleft}D$ is upper/lower triangular. The SVAR forms (6)-(10) may look different, and in fact each of them may uniquely represent physical processes and allow for direct interpretation of parameters. From a statistical point of view, however, all four SVAR DGPs introduced above are equivalent since they have identical cover.

Lemma 3 *The Gaussian covariate innovations SVAR DGP has the same cover as the Gaussian mixed output SVAR DGP. Each of these sets has a redundancy of* $2^N N!$ *for instances in which the matrices* D_w *is the product of and unitary and diagonal matrices, the matrix* C *is a unitary matrix and the matrix* A_0 *is a permutation of an upper triangular matrix.*

Proof Staring with the definition of covariate innovations SVAR in Equation (9) we use the variable transformation $y = D_w x$ and obtain the mixed-output form (trivial). The set of Guassian random variables is closed under scalar multiplication (and hence sign change) and addition. This means that the variance if the innovations term in Equation (9) can be written as:

$$\Sigma_w = D_w^T D_w = D_w^T U^T U D_w$$

Where U is a unitary (orthogonal, unit 2-norm) matrix. Since all innovations term elements are zero mean, the covariance matrix is the sole descriptor of the Gaussian innovations term. This in turn means that any other matrix $D_w^{'} = D_w^T U^T$ substituted into the DGP described in Equation (9) amounts to a stochastically equivalent DGP. The matrix $D_w^{'}$ can belong to a number of general sets of matrices, one of which is the set of nonsingular upper triangular matrices (the transformation is achievable through the QR decomposition of Σ_w). Another such set is lower triangular matrix set. Both are subsets of the set of matrices sometimes named 'psychologically upper triangular', meaning a permutation of an upper triangular matrix.

If we constrain D_w to be of the form $D_w = U\mathcal{D}$, i.e. such that (by polar decomposition) it is the product of a unitary and a diagonal positive definite matrix, the only stochastically equivalent transformations of D_w are a symmetry preserving permutation of its rows/columns and a sign change in one of the columns (this is a property of orthogonal matrices such as U). There are $N!$ such permutations and 2^N possible sign changes. For the general case, in which the input u has no special properties, there are no other redundancies in the SVAR model (since changing any parameter in A and B will otherwise change the output). Without loss of generality then, we can write the transformation from covariate innovations to mixed output SVAR form as:

$$y_i = \sum_{k=1}^{K} {}_\theta A_k y_{i-k} +_\theta Bu + U\mathcal{D}_w w_i$$

$$x_i = \sum_{k=1}^{K} U^T({}_\theta A_k) U x_{i-k} + U^T({}_\theta B)u + \mathcal{D}_w w_i$$

$$y_i = U^T x_i$$

Since the transformation U is one to one and invertible, and since this transformation is what allows a (restricted) a covariate noise SVAR to map, one to one, onto a mixed output SVAR, the cardinality of both covers is the same.

Now consider the zero-lag SVAR form:

$$y_i = \sum_{k=0}^{K} A_k y_{i-k} + Bu + \mathcal{D} w_i$$

$$\mathcal{D}^{-1}(1 - A_0)\, y_i = \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + \mathcal{D}^{-1} Bu + w$$

Taking the singular value decomposition of the (nonsingular) matrix coefficient on the LHS:

$$U_0 S V_0^T y_i = \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + \mathcal{D}^{-1} B u + w_i$$

$$V_0^T y_i = S^{-1} U_0^T \sum_{k=1}^{K} \mathcal{D}^{-1} A_k y_{i-k} + S^{-1} U_0^T \mathcal{D}^{-1} B u + S^{-1} U_0^T w_i$$

Using the coordinate transformation $z = V_0^T y$. The unitary transformation U_0^Tcan be ignored due closure properties of the Gaussian. This leaves us with the mixed-output form:

$$z_i = \sum_{k=1}^{K} S^{-1} U_0^T \mathcal{D}^{-1} A_k V_0 z_{i-k} + S^{-1} U_0^T \mathcal{D}^{-1} B u + S^{-1} w_i^{'}$$

$$y = V_0 z$$

So far we've shown that for every zero-lag SVAR there is at least one mixed-output VAR. Let us for a moment consider the covariate noise SVAR (after pre-multiplication)

$$D_w^{-1} y_i = \sum_{k=1}^{K} D_w^{-1}{}_\theta A_k y_{i-k} + D_w^{-1}{}_\theta B u + w_i$$

We can easily then write it in terms of zero lag:

$$y_i = \left(I - D_w^{-1}\right) y_i + \sum_{k=1}^{K} D_w^{-1}{}_\theta A_k y_{i-k} + D_w^{-1}{}_\theta B u + w_i$$

However, the entries of $I - D_w^{-1}$ are not zero (as required by definition). This can be done by scaling by the diagonal:

$$diag(D_w^{-1}) y_i = (diag(D_w^{-1}) - D_w^{-1}) y_i + \sum_{k=1}^{K} D_w^{-1}{}_\theta A_k y_{i-k} + D_w^{-1}{}_\theta B u + w_i$$

$$\mathcal{D}_0 \triangleq diag(\mathcal{D}_w^{-1})$$

$$y_i = (I - \mathcal{D}_0^{-1} D_w^{-1}) y_i + \sum_{k=1}^{K} \mathcal{D}_0^{-1} D_w^{-1}{}_\theta A_k y_{i-k} + \mathcal{D}_0^{-1} D_w^{-1}{}_\theta B u + \mathcal{D}_0^{-1} w_i$$

$$A_0 = (I - \mathcal{D}_0^{-1} D_w^{-1})$$

$$D_w^{-1} = diag(D_w^{-1})(I - A_0)$$

While the following constant relation preserves DGP equivalence:

$$(D_w^T D_w)^{-1} = \Sigma_w^{-1} = \mathcal{D}_w^{-1} \mathcal{D}_w^{-T} = \mathcal{D}_0 (I - A_0)(I - A_0)^T \mathcal{D}_0$$

$$A_0 = (I - \mathcal{D}_0^{-1} D_w^{-1})^T (I - \mathcal{D}_0^{-1} D_w^{-1})$$

The zero lag matrix is a function of D_w^{-1}, the inverse of which is an eigenvalue problem. However, as long as the covariance matrix or its inverse is constant, the DGP is unchanged and this allows $N(N-1)/2$ degrees of freedom. Let us consider only mixed input systems for which the innovations terms are of unit variance. There is no real loss of generality since a simple row-division by each element of $\mathcal{D}_0$ normalized the covariate noise form (to be regained by scaling the output). In this case the equivalence constraint on is one of in which:

$$(I -_{\triangleleft} A_0)^T (I -_{\triangleleft} A_0) = (I - A_0)^T (I - A_0)$$

If $(I - A_0)$ is full rank, a strictly upper triangular matrix ${}_{\triangleleft}A_0$ may be found that is equivalent (this would be the Cholesky decomposition of the inverse covariance matrix in reverse order). As D_W is equivalent to a unitary transformation $U\mathcal{D}_W$ this will include permutations and orthogonal rotations. Any permutation of D_W will imply a corresponding permutation of A_0, which (along with rotations) has $2^N N!$ solutions.

■

The non-uniqueness of SVAR and the problematic interpretation of AR coefficients with respect to variable permutation is a known problem Sims (1981), as is the fact that modeling zero-lag matrices is equivalent to covariance estimation for the Gaussian case in the other lag coefficients are zero. In fact, statistically vanishing elements of the covariance matrix are used in Structural Equation Modeling and are given causality interpretations Pearl (2000). It is not clear how robust such inferences are with respect to equivalent permutations. The point of the lemma above is to illustrate the ambiguity of interpretation if the structure of (sparse or full) AR systems in the case of covariate innovations, zero-lag, or mixed output, which are equivalent to each other. In the case of SVAR, one approach is to perform standard AR followed by a Cholesky decomposition of the covariance of the residuals and then pre-multiplying. In Popescu (2008), the upper triangular SVAR estimation is done directly by singular value decomposition after regression and the innovations covariance estimated from the zero-lag matrix.

Granger, in his 1969 paper, suggests that 'instantaneous' (i.e. covariate) effects be ignored and only the temporal structure be used. Whether or not we accept instantaneous causality depends on prior knowledge: in the case of EEG, the mixing matrix cannot have any physical 'causal' explanation even if it is sparse. Without additional *a priori* assumptions, either we infer causality on unseen and presumably interacting hidden variables (mixed output form, the case of EEG/MEG) or we assume a a non-causal mixed innovations input. Note also that the zero-lag system appears to be causal but can be written in a form which suggest the opposite difference causal influence (hence it is sometimes termed 'spurious causality'). In short, since instantaneous interaction in the Gaussian case cannot be resolved causally purely in terms of prediction and conditional information (as intended by Wiener and Granger), it is proposed that such interactions be accounted for but not given causal interpretation (as in 'strong' Granger non-causality) .

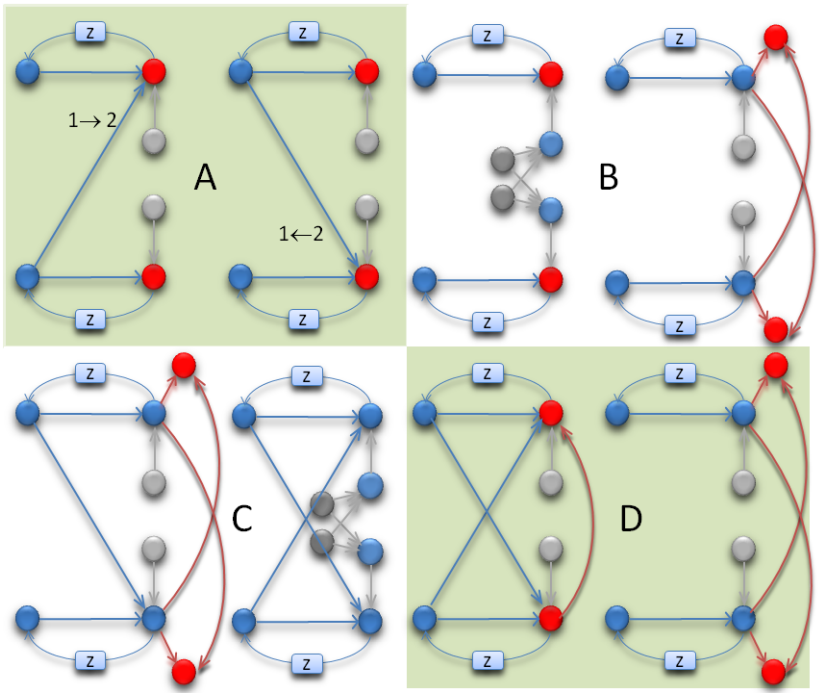


Figure 1: SVAR causality and equivalence. Structural VAR equivalence and causality. A) direct structural Granger causality (both directions shown). z stands for the delay operator. B) equivalent covariate innovations (left) and mixed output systems. Neither representation shows dynamic interaction C) sparse, one sided covariate innovations DAG is non sparse in the mixed output case (and vice-versa). D) upper triangular structure of the zero-lag matrix is not informative in the 2 variable Gaussian case, and is equivalent to a full mixed output system.

There are at least four distinct overall approaches to dealing with aliasing effects in time series causality. 1) is to make prior assumptions about covariance matrices and limit inference to domain relevant and interpretable posteriors, as in Bernanke et al. (2005) in economics and Valdes-Sosa et al. (2005) in neuroscience. 2) to allow for unconstrained graphical causal model type inference among covariate innovations, by either assuming Gaussianity or non-Gaussianity, the latter allowing for stronger causal inferences (see Moneta et al. (2011) in this volume). One possible drawback of this approach is that DAG-type inference, at least in the Gaussian case in which there is so-called 'Markov equivalence' among candidate graphs, is non-unique. 3) a physically interpretable mixed output or co-variate innovations is assumed and the inferred sparsity structure (or the intersection thereof over the nonzero lag coefficient matrices) as the connection graph. Popescu (2008) implemented such an approach by using the minimum description length principle to provide a universal prior over rational-valued coefficients, and was able to recover structure in the majority of simulated co-variate innovations processes of arbitrary sparsity. This approach is computationally laborious, as it is NP and non-convex, and moreover a system that is sparse in one form (covariate innovations or mixed-ouput) is not necessarily sparse in another equivalent SVAR form. Moreover completely dense SVAR systems may be non-causal (in the strong GC sense). 4) Causality is not interpreted as a binary value, but rather direction of interaction is determined as a continuous valued statistic, and one which is theoretically robust to covariate innovations or mixtures. This is the principle of the recently introduced phase slope index (PSI), which belongs to a class of methods based on spectral decomposition and partition of coherency. Although auto-regressive, spectral and impulse response convolution are theoretically equivalent representation of linear dynamics, they do differ numerically and spectral representations afford direct access to phase estimates which are crucial to the interpretation of lead and lag as it relates to causal influence. These methods are reviewed in the next section.

6. Spectral methods and phase estimation

Cross- and auto spectral densities of a time series, assuming zero-mean or de-trended values, are defined as:

$$\begin{aligned} \rho_{Lij}(\tau) &= E\left(y_i(t)y_j(t-\tau)\right) \\ S_{ij}\left(\omega\right) &= \mathcal{F}(\rho_{Lij}(\tau)) \end{aligned} \tag{12}$$

Note that continuous, linear, raw correlation values are used in the above definition as well as the continuous Fourier transform. Bivariate *coherency* is defined as:

$$C_{ij}(\omega) = \frac{S_{ij}\left(\omega\right)}{\sqrt{S_{ii}(\omega)S_{jj}(\omega)}} \tag{13}$$

Which consists of a complex numerator and a real-valued denominator. The coherence is the squared magnitude of the coherency:

$$c_{ij}(\omega) = C_{ij}(\omega)^* C_{ij}(\omega) \tag{14}$$

Besides various histogram and discrete (fast) Fourier transform methods available for the computation of coherence, AR methods may be also used, since they are also linear

transforms, the Fourier transform of the delay operator being simply $z^k = e^{-j2\pi\omega\tau_S}$ where τ_S is the sampling time and $k = \omega\tau_S$. Plugging this into Equation (9) we obtain:

$$X(j\omega) = \left(\sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right) X(j\omega) + BU(j\omega) + \mathcal{D}$$

$$Y(j\omega) = CX(j\omega) \tag{15}$$

$$Y(j\omega) = C\left(I - \sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right)^{-1} (BU(j\omega) + \mathcal{D}W(j\omega)) \tag{16}$$

In terms of a SVAR therefore (as opposed to VAR) the mixing matrix C does not affect stability, nor the dynamic response (i.e. the poles). The transfer functions from ith innovations to jth output are entries of the following matrix of functions:

$$H(j\omega) = C\left(I - \sum_{k=1}^{K} A_k e^{-j2\pi\omega\tau_S k}\right)^{-1} D \tag{17}$$

The spectral matrix is simply (having already assumed independent unit Gaussian noise):

$$S(j\omega) = H(j\omega)^* H(j\omega) \tag{18}$$

The coherency as the coherence following definitions above. The partial coherence considers the pair (i, j) of signals conditioned on all *other* signals, the (ordered) set of which we denote $\overline{(i,j)}$:

$$S_{i,j|\overline{(i,j)}}(j\omega) = S_{(i,j),(i,j)} + S_{(i,j),\overline{(i,j)}} S^{-1}_{\overline{(i,j)},\overline{(i,j)}} S_{\overline{(i,j)},(i,j)} \tag{19}$$

Where the subscripts refer to row/column subsets of the matrix $S(j\omega)$. The partial spectrum, substituted into Equation (13) gives us partial coherency $C_{i,j|\overline{(i,j)}}(j\omega)$ and correspondingly, partial coherence $c_{i,j|\overline{(i,j)}}(j\omega)$. These functions are symmetric and therefore cannot indicate direction of interaction in the pair (i, j). Several alternatives have been proposed to account for this limitation. Kaminski and Blinowska (1991); Blinowska et al. (2004) proposed the following normalization of $H(j\omega)$ which attempts to measure the relative magnitude of the transfer function from any innovations process to any output (which is equivalent to measuring the normalized strength of Granger causality) and is called the directed transfer function (DTF):

$$\gamma_{ij}(j\omega) = \frac{H_{ij}(j\omega)}{\sqrt{\sum_k |H_{ik}(j\omega)|^2}}$$

$$\gamma^2_{ij}(j\omega) = \frac{|H_{ij}(j\omega)|^2}{\sum_k |H_{ik}(j\omega)|^2} \tag{20}$$

A similar measure is called directed coherence Baccalá et al. (Feb 1991), later elaborated into a method complimentary to DTF, called partial directed coherence (PDC) Baccalá and Sameshima (2001); Sameshima and Baccalá (1999), based on the inverse of H:

$$\pi_{ij}(j\omega) = \frac{H_{ij}^{-1}(j\omega)}{\sqrt{\sum_k \left|H_{ik}^{-1}(j\omega)\right|^2}}$$

The objective of these coherency-like measures is to place a measure of directionality on the otherwise information-symmetric coherency. While SVAR is not generally used as a basis of the autoregressive means of spectral and coherence estimation, or of DTF/PDC is is done so in this paper for completeness (otherwise it is assumed $C = I$). Granger's 1969 paper did consider a mixing matrix (indirectly, by adding non-diagonal terms to the zero-lag matrix), and suggested ignoring the role of that part of coherency which depends on mixing terms as non-informative 'instantaneous causality'. Note that the ambiguity of the role and identifiability of the full zero lag matrix, as described herein, was fully known at the time and was one of the justifications given for separating sub-sampling time dynamics. Another measure of directionality, proposed by Schreiber (2000) is a Shannon-entropy interpretation of Granger Causality, and therefore will be referred to as GC herein. The Shannon entropy, and conditional Shannon entropy of a random process is related to its spectrum. The conditional entropy formulation of Granger Causality for AR models in the multivariate case is (where $\overline{(i)}$ denotes, as above, all other elements of the vector except i):

$$\mathcal{H}_{j\to i|u}^{GC} = \mathcal{H}(y_{i,t+1}|y_{:,t:t-K}, u_{:,t:t-K}) - \mathcal{H}(y_{i,t+1}|y_{\overline{(j)},t:t-K}, u_{:,t:t-K})$$

$$\mathcal{H}_{j\to i|u}^{GC} = \log \mathcal{D}_i - \log \mathcal{D}_i^{\overline{(j)}} \tag{21}$$

The Shannon entropy of a Gaussian random variable is the logarithm of its standard deviation plus a constant. Notice than in this paper the definition of Granger Causality is slightly different than the literature in that it relates to the innovations process of a mixed output SVAR system of closest rotation and not a regular MVAR. The second term $\mathcal{D}_i^{\overline{(j)}}$ is formed by computing a reduced SVAR system which omits the jth variable. Recently Barrett et al. have proposed an extension of GC, based on prior work by Geweke (1982) from interaction among pairs of variables to groups of variables, termed multivariate Granger Causality (MVGC) Barrett et al. (2010). The above definition is straightforwardly extensible to the group case, where I ad J are subsets of $1..D$, since total entropy of independent variables is the sum of individual entropies.

$$\mathcal{H}_{J\to I|u}^{GC} = \sum_{i\in I} \left(\log \mathcal{D}_i - \log \mathcal{D}_i^{\overline{(J)}}\right) \tag{22}$$

The Granger entropy can be calculated directly from the transfer function, using the Shannon-Hartley theorem:

$$\mathcal{H}_{j\to i}^{GCH} = -\sum_{\omega} \Delta\omega \, \ln\left(1 - \frac{|H_{ij}(\omega)|^2}{S_{ii}(\omega)}\right) \tag{23}$$

Finally Nolte (Nolte et al., 2008) introduced a method called Phase Slope Index which evaluates bilateral causal interaction and is robust to mixing effects (i.e. zero lag, observation or innovations covariance matrices that depart from MVAR):

$$PSI_{ij} = Im\left(\sum_{\omega} C^*_{ij}(\omega)\ C_{ij}(\omega + d\omega)\right) \tag{24}$$

PSI, as a method is based on the observation that pure mixing (that is to say, all effects stochastically equivalent to output mixing as outlined above) does not affect the imaginary part of the coherency C_{ij} just as (equivalently) it does not affect the antisymmetric part of the auto-correlation of a signal. It does not place a measure the phase relationship *per se*, but rather the slope of the coherency phase weighted by the magnitude of the coherency.

7. Causal Structural Information

Currently, Granger causality estimation based on linear VAR modeling has been shown to be susceptible to mixed noise, in the presence of which it may produce false causality assignment Nolte et al. (2010). In order to allow for accurate causality assignment in the presence of instantaneous interaction and aliasing the Causal Structural Information (CSI) method and statistic for causality assignment is introduced below.

Consider the SVAR lower triangular form in (11) for a set of observations y. The information transfer from i to j may be obtained by first defining the index re-orderings:

$$ij* \triangleq \{i, j, \overline{ij}\}$$

$$i* \triangleq \{i, \overline{ij}\}$$

This means that we reorder the (identified) mixed-innovations system by placing the target time series first and the driver series second, followed by all the rest. The same ordering, minus the driver is also useful. We define CSI as

$$CSI(j \to i|\overline{ij}) \triangleq \log({}_{\lhd ij*}D_{11}) - \log({}_{\lhd i*}D_{11}) \tag{25}$$

$$CSI(i, j|\overline{ij}) \triangleq CSI(j \to i|\overline{ij}) - CSI(i \to j|\overline{ij}) \tag{26}$$

Where the D is *upper-triangular* form in each instance. This Granger Causality formulation requires the identification of 3 different SVAR models, one for the entire time series vector, and one each for all elements except i and all elements except j. Via Cholesky decomposition, the logarithm of the top vertex of the triangle is proportional to the entropy rate (conditional information) of the innovations process for the target series given all other (past and present) information including the innovations process. While this definition is clearly an interpretation of the core idea of Granger causality, it is, like DTF and PDC, not an independence statistic but a measure of (causal) information flow among elements of a time-series vector. Note the anti-symmetry (by definition) of this information measure $CSI(i, j|\overline{ij}) = -CSI(j, i|\overline{ij})$. Note also that $CSI(j \to i|\overline{ij})$ and $CSI(i \to j|\overline{ij})$ may very conceivably have the same sign: the various triangular forms used to derive this measure are purely for calculation purposes, and do not carry intrinsic meaning. As a matter of fact

other re-orderings and SVAR forms may be employed for convenient calculation as well. In order to improve the explanatory power of the CSI statistic the following normalization is proposed, mirroring that defined in Equation (5) :

$$\mathcal{F}^{CSI}_{j\to i|\overline{ij}} \triangleq \frac{CSI(i,j|\overline{ij})}{\log(_{\lhd i*}D_{11}) + \log(_{\lhd j*}D_{11}) + \zeta} \tag{27}$$

This normalization effectively measures the ratio of causal to non-causal information, where ζ is a constant which depends on the number of dimensions and quantization width and is necessary to transform continuous entropy to discrete entropy.

8. Estimation of multivariate spectra and causality assignment

In Section 6 and a series of causality assignment methods based on spectral decomposition of a multivariate signal were described. In this section spectral decomposition itself will be discussed, and a novel means of doing so for unevenly sampled data will be introduced and evaluated along with the other methods for a bivariate benchmark data set.

8.1. The cardinal transform of the autocorrelation

Currently there are few commonly used methods for cross- *power* spectrum estimation (i.e. multivariate spectral power estimation) as opposed to univariate power spectrum estimation, and these methods average over repeated, or shifting, time windows and therefore require a lot of data points. Furthermore all commonly used multivariate spectral power estimation methods rely on synchronous, evenly spaced sampling, despite the fact that much of available data is unevenly sampled, has missing values, and can be composed of relatively short sequences. Therefore a novel method is presented below for multivariate spectral power estimation which can be estimated on asynchronous data.

Returning to the definition of coherency as the Fourier transform of the auto-correlation, which are both continuous transforms, we may extend the conceptual means of its estimation in the discrete sense as a regression problem (as a discrete Fourier transform, DFT) in the evenly sampled case as:

$$\Omega_n \triangleq \frac{n}{2\tau_0(N-1)}, \quad n = -\lfloor N/2 \rfloor ... \lfloor N/2 \rfloor \tag{28}$$

$$\hat{C}_{ij}(\omega)|_{\ \omega=\Omega} = a_{ij,n} + \mathrm{j} b_{ij,n} \tag{29}$$

$$\rho_{ji}(-k\tau) = \rho_{ij}(k\tau) = E(x_i(t)x_j(t+k\tau)) \tag{30}$$

$$\rho_{ij}(k\tau_0) \cong \frac{1}{N-k} \sum_{q=1:N-k} x_i(q)x_j(q+k) \tag{31}$$

$$\{a_{ij}, b_{ij}\} = \arg\min \sum_{k=-N/2}^{N/2} (\rho_{ij}(k\tau_0) - a_{ij,n} cos(2\pi\Omega_n\tau_0 k) - b_{ij,n} sin(2\pi\Omega_n\tau_0 k)\ )^2 \tag{32}$$

where τ_0 is the sampling interval. Note that for an odd number of points the regression above is actually a well determined set of equations, corresponding to the 2-sided DFT. Note also that by replacing the expectation with the geometric mean, the above equation can also be written (with a slight change in weighting at individual lags) as:

$$\{a_{ij}, b_{ij}\} = \arg\min \sum_{p,q \in 1..N} (x_{i,p}x_{j,q} - a_{ij,n}cos(2\pi\Omega_k(t_{i,p} - t_{j,q})) - b_{ij,n}sin(2\pi\Omega_n(t_{i,p} - t_{j,q}))\ )^2 \tag{33}$$

The above equation holds even for time series sampled at unequal (but overlapping) times (x_i, t_i) and (x_j, t_j) as long as the frequency basis definition is adjusted (for example $\tau_0 = 1$). It represents a discrete, finite approximation of the continuous, infinite auto-regression function of an infinitely long random process. It is a regression on the outer product of the vectors x_i and x_j. Since autocorrelations for finite memory systems tend to fall off to zero with increasing lag magnitude, a novel coherency estimate is proposed based on the cardinal sine and cosine functions, which also decay, as a compact basis:

$$\hat{C}_{ij}(\omega) = a_{ij,n} \sum \mathcal{C}(\Omega_n) + \mathrm{j}\, b_{ij,n}\mathcal{S}(\Omega_n) \tag{34}$$

$$\{a_{ij}, b_{ij}\} = \arg\min$$

$$\sum_{p,q \in 1..N} (x_{i,p}x_{j,q} - a_{ij,n}cosc(2\pi\Omega_k(t_{i,p} - t_{j,q})) - b_{ij,n}sinc(2\pi\Omega_n(t_{i,p} - t_{j,q}))\ )^2 \tag{35}$$

Where the sine cardinal is defined as $sinc(x) = sin(\pi x)/x$, and its Fourier transform is $\mathcal{S}(j\omega) = 1, |j\omega| < 1$ and $\mathcal{S}(j\omega) = 0$ otherwise. Also the Fourier transform of the cosine cardinal can be written as $\mathcal{C}(j\omega) = j\omega\ S(j\omega)$. Although in principle we could choose any complete basis as a means of Fourier transform estimation, the cardinal transform preserves the odd-even function structure of the standard trigonometric pair. Computationally this means that for autocorrelations, which are real valued and even, only *sinc* needs to be calculated and used, while for cross-correlation both functions are needed. As linear mixtures of independent signals only have symmetric cross-correlations, any nonzero values of the *cosc* coefficients would indicate the presence of dynamic interaction. Note that the Fast Fourier Transform earns its moniker thanks to the orthogonality of *sin* and *cos* which allows us to avoid a matrix inversion. However their orthogonality holds true only for infinite support, and slight correlations are found for finite windows - in practice this effect requires further computation (windowing) to counteract. The cardinal basis is not orthogonal, requires full regression and may have demanding memory requirements. For moderate size data this not problematic and implementation details will be discussed elsewhere.

8.2. Robustness evaluation based on the NOISE dataset

A dataset named NOISE, intended as a benchmark for the bivariate case, has been introduced in the preceding NIPS workshop on causality Nolte et al. (2010) and can be found online at *www.causality.inf.ethz.ch/repository.php*, along with the code that generated the data. It was awarded best dataset prize in the previous NIPS causality workshop and

challenge Guyon et al. (2010). For further discussion of Causality Workbench and current dataset usage see Guyon (2011). NOISE is created by the summation of the output of a strictly causal VAR DGP and a non-causal SVAR DGP which consists of mixed colored noise:

$$y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i} \tag{36}$$

$$x_{N,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix}_{N,k} x_{N,i-k} + w_{N,i}$$

$$y_{N,i} = Bx_{N,i} \tag{37}$$

$$y = (1 - |\beta|)y_N + |\beta| y_C \frac{\|y_N\|_F}{\|y_C\|_F} \tag{38}$$

The two sub-systems are pictured graphically as systems A and B in Figure 1. If $\beta < 0$ the AR matrices that create y_C are transposed (meaning that y_{1C} causes y_{2C} instead of the opposite). The coefficient β is represented in Nolte et al. (2010) by 'γ ' where $\beta = sgn(\gamma)(1 - |\gamma\ |)$. All coefficients are generated as independent Gaussian random variables of unit variance, and unstable systems are discarded. While both the causal and noise generating systems have the same order, note that the system that would generate the sum thereof requires an infinite order SVAR DGP to generate (it is not stochastically equivalent to any SVAR DGP but instead is a SVARMA DGP, having both poles and zeros). Nevertheless it is an interesting benchmark since the exact parameters are not fully recoverable via the commonly used VAR modeling procedure and because the causality interpretation is fairly clear: the sum of a strictly causal DGP and a stochastic noncausal DGP should retain the causality of the former.

In this study, the same DGPs were used as in NOISE but as one of the current aims is to study the influence of sample size on the reliability of causality assignment, signals of 100, 500, 1000 and 5000 points were generated (as opposed to the original 6000). This is the dataset referred to as PAIRS below, which only differs in numbers of samples per time series. For each evaluation 500 time series were simulated, with the order for each system of being uniformly distributed from 1 to 10. The following methods were evaluated:

- PSI (Ψ) using Welch's method, and segment and epoch lengths being equal and set to $\left\lceil \sqrt{N} \right\rceil$ and otherwise is the same as Nolte et al. (2010).
- Directed transfer function DTF. estimation using an automatic model order selection criterion (BIC, Bayesian Information Criterion) using a maximum model order of 20. DTF has been shown to be equivalent to GC for linear AR models (Kaminski, 2001) and therefore GC itself is not shown . The covariance matrix of the residuals is also included in the estimate of the transfer function. The same holds for all methods described below.
- Partial directed coherence PDC. As described in the previous section, it is similar to DTF except it operates on the signal-to-innovations (i.e. inverse) transfer function.

- Causal Structural Information. As a described above this is based on the triangular innovations equivalent to the estimated SVAR (of which there are 2 possible forms in the bivariate case) and which takes into account instantaneous interaction / innovations process covariance.

All methods were statistically evaluated for robustness and generality by performing a 5-fold jackknife, which gave both a mean and standard deviation estimate for each method and each simulation. All statistics reported below are mean normalized by standard deviation (from jackknife). For all methods the final output could be -1, 0, or 1, corresponding to causality assignment 1→2, no assignment, and causality 2 → 1. A true positive (TP) was the rate of correct causality assignment, while a false positive (FP) was the rate of incorrect causality assignment (type III error), such that TP+FP+NA=1, where NA stands for rate of no assignment (or neutral assignment). The TP and FP rates are be co-modulated by increasing/decreasing the threshold of the absolute value of the *mean/std* statistic, under which no causality assignment is made:

$$STAT = rawSTAT/\mathrm{std}(rawSTAT),\ rawSTAT\text{=}PSI,\ DTF,\ PDC\ ..$$

$$c = sign\,(STAT)\ \text{ if } STAT > TRESH,\ 0 \text{ otherwise}$$

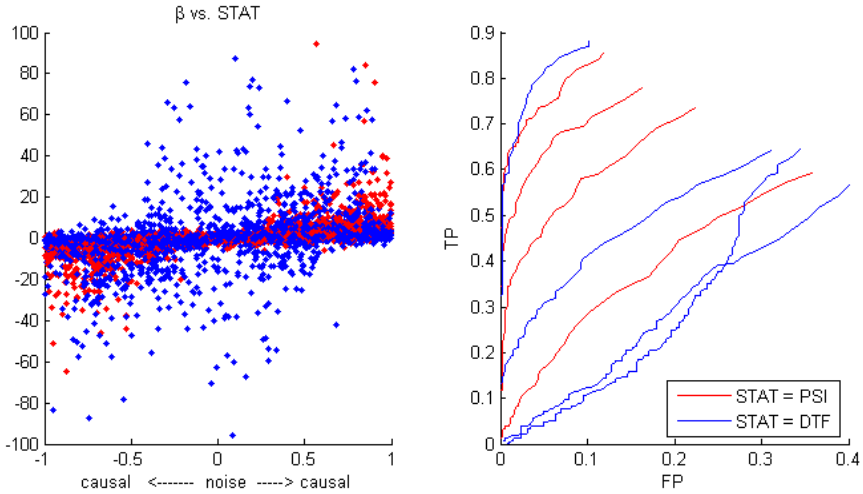
In Table 1 we see results of overall accuracy and controlled True Positive rate for the ***non-mixed*** colored noise case (meaning the matrix B above is diagonal). In Table 1 and Table 2 methods are ordered according to the mean TP rate over time series length (highest on top).

Table 1: Unmixed colored noise PAIRS

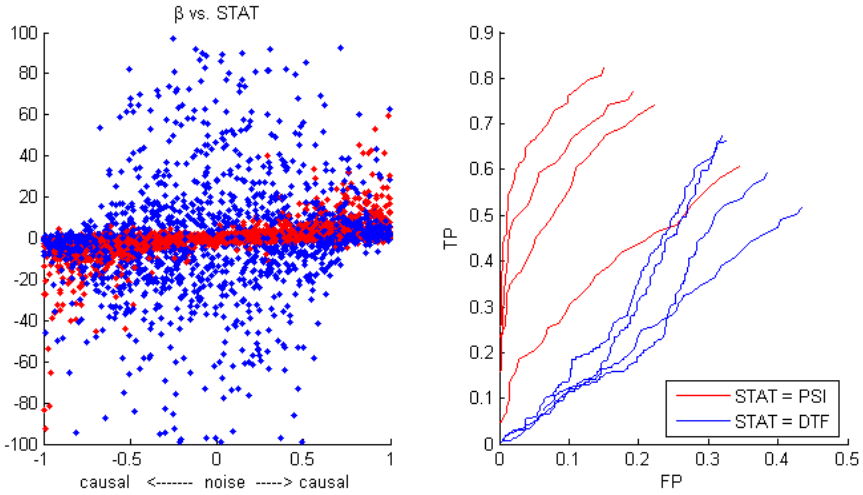
	Max. Accuracy				TP , FP < 0.10			
	100	**500**	**1000**	**5000**	**100**	**500**	**1000**	**5000**
Ψ	0.62	0.73	0.83	0.88	0.25	0.56	0.75	0.85
DTF	0.58	0.79	0.82	0.88	0.18	0.58	0.72	0.86
CSI	0.62	0.72	0.79	0.89	0.23	0.53	0.66	0.88
Ψ_C	0.57	0.68	0.81	0.88	0.19	0.29	0.70	0.87
PDC	0.64	0.67	0.75	0.78	0.23	0.33	0.48	0.57

In Table 2 we can see results for a PAIRS, in which the noise mixing matrix B is not strictly diagonal.

As we can see in both Figure 2 and Table 1, all methods are almost equally robust to unmixed colored additive noise (except PDC). However, while *addition of mixed colored noise* induces a mild gap in maximum accuracy, it creates a large gap in terms of TP/FP rates. Note the dramatic drop-off of the TP rate of VAR/SVAR based methods PDC and DTF. Figure 3 shows this most clearly, by a wide scatter of STAT outputs for DTF around $\beta = 0$ that is to say with no actual causality in the time series and a corresponding fall-off of TP vs. FP rates. Note also that PSI methods still allow a fairly reasonable TP rate determination at low FP rates of 10% even at 100 points per time-series, while the CSI

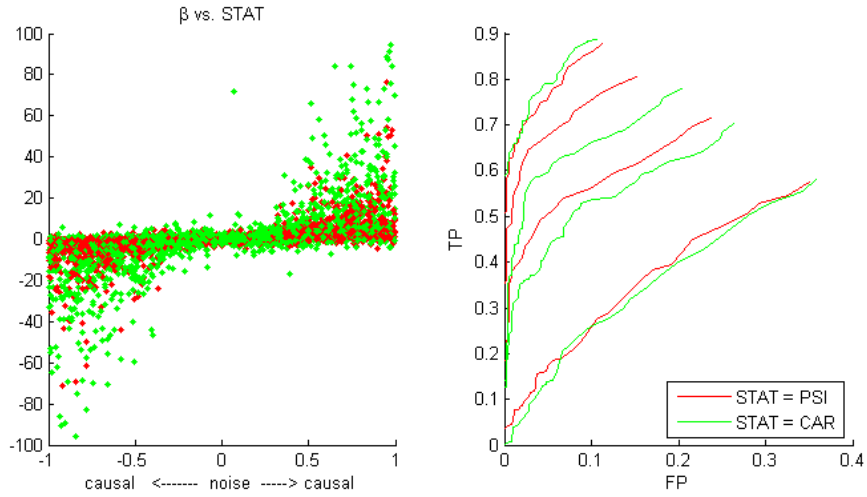


(*a*) Unmixed colored noise

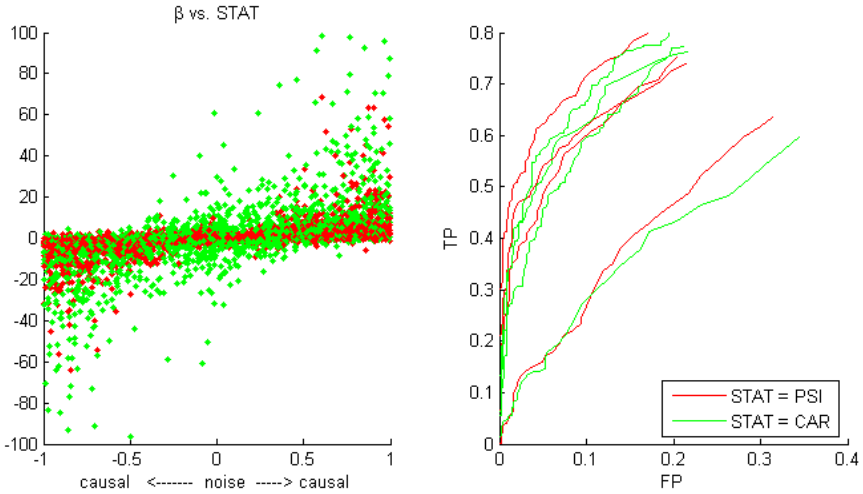


(*b*) Mixed colored noise

Figure 2: PSI vs. DTF Scatter plots of β vs. STAT (to the left of each panel), TP vs. FP curves for different time series lengths (100, 500,1000 and 500) (right). a) colored unmixed additive noise. b) colored mixed additive noise. DTF is equivalent to Granger Causality for linear systems. All STAT values are jackknife mean normalized by standard deviation.



(*a*) Unmixed colored noise



(*b*) Mixed colored noise

Figure 3: PSI vs. CSI Scatter plots of β vs. STAT (to the left of each panel), TP vs. FP curves for different time series lengths (right). a) unmixed additive noise. b) mixed additive noise

Table 2: Mixed colored noise PAIRS

	Max. Accuracy				TP , FP< 0.10			
	N=100	**500**	**1000**	**5000**	**N=100**	**500**	**1000**	**5000**
Ψ_C	0.64	0.74	0.81	0.83	0.31	0.49	0.64	0.73
Ψ	0.66	0.76	0.78	0.81	0.25	0.59	0.61	0.71
CSI	0.63	0.77	0.79	0.80	0.27	0.62	0.59	0.66
PDC	0.64	0.71	0.69	0.66	0.24	0.30	0.29	0.24
DTF	0.55	0.61	0.66	0.66	0.11	0.10	0.09	0.12

method was also robust to the addition of colored mixed noise, not showing any significant difference with respect to PSI except a higher FP rate for longer time series (N=5000). The advantage of PSIcardinal was near PSI in overall accuracy. In conclusion, DTF (or weak Granger causality) and PDC are not robust with respect to additive mixed colored noise, although they perform similarly to PSI and CSI for independent colored noise. [1]

9. Conditional causality assignment

In multivariate time series analysis we are often concerned with inference of causal relationship among more than 2 variables, in which the role of a potential common cause must be accounted for, analogously to vanishing partial correlation in the static data case. For this reason the PAIRS data set was extended into a set called TRIPLES in which the degree of common driver influence versus direct coupling was controlled.

In effect, the TRIPLES DGP is similar to PAIRS, in that additive noise is mixed colored noise (in 3 dimensions) but in this case another variable x_3 may drive the pair x_1, x_2 independently of each other, also with random coefficients (but either one set to 1/10 of the other randomly). That is to say, the signal is itself a mixture of one where there is a direct one sided causal link among x_1, x_2 as in PAIRS and one where they are actually independent but commonly driven, according to a parameter χ which at 0 is commonly driven and at 1 is causal.

$$\beta < 0 \quad y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i}$$

1. Correlation and rank correlation analysis was performed (for N=5000) to shed light on the reason for the discrepancy between PSI and CSI. The linear correlation between rawSTAT and STAT was .87 and .89 for PSI and CSI. No influence of model order K of the simulated system was seen in the error of either PSI or CSI, where error is estimated as the difference in rank of $rankerr(STAT) = |rank(\beta) - rank(STAT)|$. There were however significant correlations between $rank(|\beta|)$ and $rankerr(STAT)$, -.13 for PSI and -.27 for CSI. Note that as expected, standard Granger causality (GC) performed the same as DTF (TP=0.116 for FP<.10). Using Akaike's Information Criterion (AIC) instead of BIC for VAR model order estimation did not significantly affect AR-based STAT values.

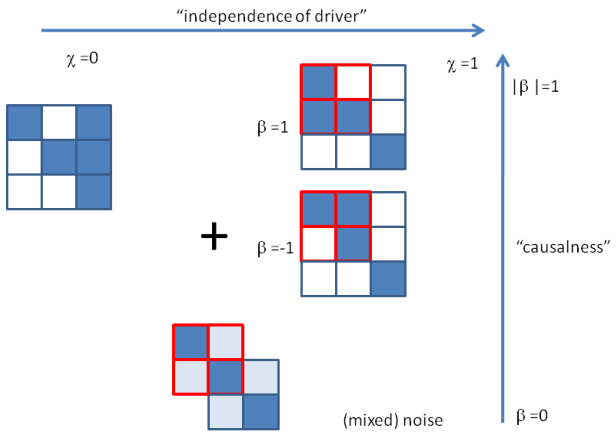


Figure 4: Diagram of TRIPLES dataset with common driver

$$\beta > 0 \quad y_{C,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & 0 \\ a_{12} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}_{C,k} y_{C,i-k} + w_{C,i} \tag{39}$$

$$x_{N,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{22} \end{bmatrix}_{N,k} x_{N,i-k} + w_{N,i}$$

$$y_{N,i} = Bx_{N,i} \tag{40}$$

$$x_{D,i} = \sum_{k=1}^{K} \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{22} \end{bmatrix}_{D,k} x_{D,i-k} + w_{D,i}$$

$$y_{MC} = (1 - |\beta|)y_N + |\beta| y_C \frac{\|y_N\|_F}{\|y_C\|_F} \tag{41}$$

$$y_{DC} = (1 - \chi\ )y_{MC} + \chi\ \ y_D \frac{\|y_{MC}\|_F}{\|y_D\|_F} \tag{42}$$

The Table 3, similar to the tables in the preceding section, shows results for all usual methods, except for PSIpartial which is PSI calculated on the partial coherence as defined above and calculated from Welch (cross-spectral) estimators in the case of mixed noise and a common driver.

Notice that the TP rates are lower for all methods with respect to Table 2 which represents the mixed noise situation without any common driver.

Table 3: TRIPLES: Commonly driven, additive mixed colored noise

	Max. Accuracy				TP , FP< 0.10			
	100	**500**	**1000**	**5000**	**100**	**500**	**1000**	**5000**
Ψ_p	0.53	0.61	0.71	0.75	0.12	0.31	0.49	0.56
Ψ	0.54	0.60	0.70	0.72	0.10	0.25	0.40	0.52
CSI	0.51	0.60	0.69	0.76	0.09	0.27	0.38	0.45
PDC	0.55	0.54	0.60	0.58	0.13	0.12	0.16	0.13
DTF	0.51	0.56	0.59	0.61	0.12	0.09	0.09	0.11

10. Discussion

In a recent talk, Emanuel Parzen (Parzen, 2004) proposed, both in hindsight and for future consideration, that aim of statistics consist in an 'answer machine', i.e. a more intelligent, automatic and comprehensive version of Fisher's almanac, which currently consists in a plenitude of chapters and sections related to different types of hypotheses and assumption sets meant to model, insofar as possible, the ever expanding variety of data available. These categories and sub-categories are not always distinct, and furthermore there are competing general approaches to the same problems (e.g. Bayesian vs. frequentist). Is an 'answer machine' realistic in terms of time-series causality, prerequisites for which are found throughout this almanac, and which has developed in parallel in different disciplines?

This work began by discussing Granger causality in abstract terms, pointing out the implausibility of finding a general method of causal discovery, since that depends on the general learning and time-series prediction problem, which are incomputable. However, if any consistent patterns that can be found mapping the history of one time series variable to the current state of another (using non-parametric tests), there is sufficient evidence of causal interaction and the null hypothesis is rejected. Such a determination still does not address direction of interaction and relative strength of causal influence, which may require a complete model of the DGP. This study - like many others - relied on the rather strong assumption of stationary linear Gaussian DGPs but otherwise made weak assumptions on model order, sampling and observation noise. Are there, instead, more general assumptions we can use? The following is a list of competing approaches in increasing order of (subjectively judged) strength of underlying assumption(s):

- Non-parametric tests of conditional probability for Granger non-causality rejection. These directly compare the probability distributions $P(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1})$ $P(y_{1,j} \mid y_{1,j-1..1}, u_{j-1..1})$ to detect a possible statistically significant difference. Proposed approaches (see chapter in this volume by (Moneta et al., 2011) for a detailed overview and tabulated robustness comparison) include product kernel density with kernel smoothing (Chlaß and Moneta, 2010), made robust by bootstrapping and with density distances such as the Hellinger (Su and White, 2008), Euclidean (Szekely and Rizzo, 2004), or completely nonparametric difference tests such Cramer-Von Mises or Kolmogorov-Smirnov. A potential pitfall of nonparametric approaches is their loss of power for higher dimensionality of the space over which the probabilities are esti-

mated - *aka* the curse of dimensionality (Yatchew, 1998). This can occur if the lag order K needed to be considered is high, if the system memory is long, or the number of other variables over which GC must be conditioned ($u_{j-1..1}$) is high. In the case of mixed noise, strong GC estimation would require accounting for all observed variables (which in neuroscience can number in the hundreds). While non-parametric non-causality rejection is a very useful tool (and could be valid even if the lag considered in analysis is much smaller than the true lag K), in practice we would require robust estimated of causal direction and relative strength of different factors, which implies a complete accounting of all relevant factors. As was already discussed, in many cases Granger non-causality is likely to be rejected in both directions: it is useful to find the dominant one.

- General parametric or semi-parametric (black-box) predictive modeling subject to GC interpretation which can provide directionality, factor analysis and interpretation of information flow. A large body of literature exists on neural network time series modeling (in this context see White (2006)), complemented in recent years by support vector regression and Bayesian processes. The major concern with black-box predictive approaches is model validation: does the fact that a given model features a high cross-validation score automatically imply the implausibility of another predictive model with equal CV-score that would lead to different conclusions about causal structure? A reasonable compromise between nonlinearity and DGP class restriction can be seen in (Chu and Glymour, 2008) and Ryali et al. (2010), in which the VAR model is augmented by additive non-linear functions of the regressed variable and exogenous input. Robustness to noise, sample size influence and accuracy of effect strength and direction determination are open questions.

- Linear dynamic models which incorporate (and often require) non-Gaussianity in the innovations process such as ICA and TDSEP (Ziehe and Mueller, 1998). See Moneta et al. (2011) in this volume for a description of causality inference using ICA and causal modeling of innovation processes (i.e. independent components). Robustness under permutation is necessary for a principled accounting of dynamic interaction and partition of innovations process entropy. Note that many ICA variants assume that at most one of the innovations processes is Gaussian, a strong assumption which requires a posteriori checks. To be elucidated is the robustness to filtering and additive noise.

- Non-stationary Gaussian linear models. In neuroscience non-stationarity is important (the brain may change state abruptly, respond to stimuli, have transient pathological episodes etc). Furthermore accounting for non-stochastic exogenous inputs needs further consideration. Encouragingly, the current study shows that even in the presence of complex confounders such as common driving signals and co-variate noise, segments as small as 100 points can yield accurate causality estimation, such that changes in longer time series can be adaptively tracked. Note that in establishing statistical significance we must take into account signal bandwidth: up-sampling the same process would arbitrarily increase the number of samples but not the information contained in the signal. See Appendix A for a proposal on non-parametric bandwidth estimation.

- Linear Gaussian dynamic models: in this work we have considered SVAR but not wider classes of linear DGPs such as VARMA and heteroskedastic (GARCH) models. In comparing PSI and CSI note that overall accuracy of directionality assignment was virtually identical, but PSI correlated slightly better with effect size. While CSI made slightly more errors at low strengths of 'causality', PSI made slightly more errors at high strengths. Nevertheless, PSI was most robust to (colored, mixed) noise and hidden driving/conditioning signal (tabulated significance results are provided in Appendix A). Jackknifed, normalized estimates can help establish causality at low strength levels, although a large raw PSI statistic value may also suffice. A potential problem with the jackknife (or bootstrap) procedure is the strong stationarity assumption which allows segmentation and rearrangement of the data.

Although AR modeling was commonly used to model interaction in time series and served as a basis for (linear) Granger causality modeling (Blinowska et al., 2004; Baccalá and Sameshima, 2001), robustness to mixed noise remained a problem, which the spectral method PSI was meant to resolve (Nolte et al., 2008). While 'phase', if structured, already implies prediction, precedence and mutual information among time series elements, it was not clear how SVAR methods would reconcile with PSI performance, until now. This prompted the introduction in this article of the causal AR method (CSI) which takes into account 'instantaneous' causality . A prior study had shown that strong Granger causality is preserved under addition of colored noise, as opposed to weak (i.e. strictly time ordered) causality Solo (2006). This is consistent with the results obtained herein. The CSI method, measuring strong Granger Causality, was in fact robust with respect to a type of noise not studied in (Solo, 2006), which is *mixed* colored noise; other VAR based methods and (weak) Granger causality measures were not. While and SVAR DGP observed under additive colored noise is a VARMA process (the case of the PAIRS and TRIPLES datasets), SVAR modeling did not result in a severe loss of accuracy. AR processes of longer lags can approximate VARMA processes by using higher orders and more parameters, even if doing so increases exposure to over-fit and may have resulted in a small number of outliers. Future work must be undertaken to ascertain what robustness and specificity advantages result from VARMA modeling, and if it is worth doing so considering the increased computational overload. One of the common 'defects' of real-life data are missing/outlier samples, or uneven sampling in time, or that the time stamps of two time series to be compared are unrelated though overlapping: it is for these case that the method PSIcardinal was developed and shown to be practically equal in numerical performance to the Welch estimate-based PSI method (though it is slower computationally). Both PSI estimates were robust to common driver influence even when not based on partial but direct coherency because it is the *asymmetry* in influence of the driver on phase that is measured rather than its overall strength. While 2-way interaction with conditioning was considered, future work must organize multivariate signals using directed graphs, as in DAG-type static causal inference. Although only 1 conditioning signal was analysed in this paper, the methods apply to higher numbers of background variables. Directed Transfer Function and Partial Directed Coherence did not perform as well under additive colored noise, but their formulation does address a theoretically important question, namely the partition of strength of influence among various candidate causes of an observation; CSI also proposes an index for this important purpose.

Whether the assumptions about stationarity or any other data properties discussed are warranted may be checked by performing appropriate *a posteriori* tests. If these tests justify prior assumptions and a correspondingly significant causal effect is observed, we can assign statistical confidence intervals to the causality structure of the system under study. The 'almanac' chapter on time series causality is rich and new alternatives are emerging. For the entire corpus of time-series causality statistics to become an 'answer machine', however, it is suggested that a principled bottom-up investigation be undertaken, beginning with the simple SVAR form studied in this paper and all proposed criteria be quantified: type I, II and III errors, accurate determination of causality strength and direction and robustness in the presence of conditioning variables and colored mixed noise.

Acknowledgments

I would like to thank Guido Nolte for his insightful feedback and informative discussions.

References

H. Akaike. On the use of a linear model for the identification of feedback systems. *Annals of the Institute of statistical mathematics*, 20(1):425–439, 1968.

L. A. Baccalá, M. A. Nicolelis, C. H. Yu, and M. Oshiro. Structural analysis of neural circuits using the theory of directed graphs. *Computers and Biomedical Research, an International Journal*, 24:7–28, Feb 1991. URL http://www.ncbi.nlm.nih.gov/pubmed/2004525.

L. A Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463â“474, 2001.

A. B. Barrett, L. Barnett, and A. K. Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, April 2010. doi: 10.1103/PhysRevE.81.041907. URL http://link.aps.org/doi/10.1103/PhysRevE.81.041907.

B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics*, 120(1):387–422, 2005.

K. J. Blinowska, R. Kuś, and M. Kamiński. Granger causality and information flow in multivariate processes. *Physical Review E*, 70(5):050902, November 2004. doi: 10.1103/PhysRevE.70.050902. URL http://link.aps.org/doi/10.1103/PhysRevE.70.050902.

P. E. Caines. Weak and strong feedback free processes. *IEEE. Trans. Autom . Contr*, 21: 737–739, 1976.

N. Chlaß and A. Moneta. Can Graphical Causal Inference Be Extended to Nonlinear Settings? *EPSA Epistemology and Methodology of Science*, pages 63–72, 2010.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *The Journal of Machine Learning Research*, 9:967–991, 2008.

R. A. Fisher. *Statistical Methods for Research Workers.* Macmillan Pub Co, 1925. ISBN 0028447301.

W. Gersch and G. V. Goddard. Epileptic focus location: spectral analysis method. *Science (New York, N.Y.)*, 169(946):701–702, August 1970. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/5429908. PMID: 5429908.

J. Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77:304–313, 1982.

G. Gigerenzer, Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Kruger. *The Empire of Chance: How Probability Changed Science and Everyday Life.* Cambridge University Press, October 1990. ISBN 052139838X.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, August 1969. ISSN 00129682. URL http://www.jstor.org/stable/1912791.

I. Guyon. Time series analysis with the causality workbench. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. wcci2008 workshop on causality, hong kong, june 3-4 2008. *Journal of Machine Learning Research Workshop and Conference Proceedings*, 3:1–33, 2008.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3: 1157–1182, March 2003.

I. Guyon, D. Janzing, and B. Schölkopf. Causality: Objectives and assessment. *JMLR W&CP*, 6:1–38, 2010.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. 2009.

Ding M. Truccolo W. A. & Bressler S. L. Kaminski, M. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biol Cybern*, 85:145–57, 2001.

M. J. Kaminski and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, 65(3):203–210, 1991. ISSN 0340-1200. doi: 10.1007/BF00198091. URL http://dblp.uni-trier.de/rec/bibtex/journals/bc/KaminskiB91.

A. N. Kolmogorov and A. N. Shiryayev. *Selected Works of A.N. Kolmogorov: Probability theory and mathematical statistics.* Springer, 1992. ISBN 9789027727978.

T. C. Koopmans. *Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph, No. 10.* New York: John Wiley & Sons, 1950.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering Cyclic Causal Models by Independent Component Analysis. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI-2008), Helsinki, Finland*, 2008.

A. D. Lanterman. Schwarz, wallace and rissanen: Intertwining themes in theories of model selection. *International Statistical Review*, 69(2):185–212, 2001.

M. Li and P. M. B. Vitanyi. *An introduction to Kolmogorov complexity and its applications, 2nd edition.* Springer-Verlag, 1997.

A. Moneta, N. Chlaß, D. Entner, and P.Hoyer. Causal search in structural vector autoregression models. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

A. Moneta, D. Entner, P.O. Hoyer, and A. Coad. Causal inference by independent component analysis with applications to micro-and macroeconomic data. *Jena Economic Research Papers*, 2010:031, 2010.

G. Nolte, A. Ziehe, N. Kraemer, F. Popescu, and K.-R. Müller. Comparison of granger causality and phase slope index. *Journal of Machine Learning Research Workshop & Conference Proceedings.*, Causality: Objectives and Assessment:267:276, 2010.

G. Nolte, A. Ziehe, V.V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Physical Review Letters*, 00(23):234101, 2008.

E. Parzen. Long memory of statistical time series modeling. presented at the 2004 nber/nsf time series conference at smu, dallas, usa. Technical report, Texas A&M University, http://www.stat.tamu.edu/ eparzen/LongSeries

J. Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, Cambridge, 2000.

K. Pearson. *Tables for statisticians and biometricians,.* University Press, University College, London, [Cambridge Eng., 1930.

Peter C.B. Phillips. The problem of identification in finite parameter continous time models. *Journal of Econometrics*, 1:351–362, 1973.

F. Popescu. Identification of sparse multivariate autoregressive models. *Proceedings of the European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland*, 2008.

T. Richardson and P. Spirtes. Automated discovery of linear feedback models. In *Computation, causation and discovery.* AAAI Press and MIT Press, Menlo Park, 1999.

A. Roebroeck, A. K. Seth, and P. Valdes-Sosa. Causal time series analysis of functional magnetic resonance imaging data. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, XX. Time Series Causality:XX–XX, 2011.

S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fmri. *Neuroimage*, 2010.

K. Sameshima and L. A Baccalá. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods*, 94(1):93:103, 1999.

R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, July 2000. doi: 10.1103/PhysRevLett.85.461. URL http://link.aps.org/doi/10.1103/PhysRevLett.85.461.

C. A. Sims. An autoregressive index model for the u.s. 1948-1975. In J. Kmenta and J.B. Ramsey, editors, *Large-scale macro-econometric models: theory and practice*, pages 283–327. North-Holland, 1981.

V. Solo. On causality i: Sampling and noise. *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 3634–3639, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, Cambridge MA, 2nd edition, 2000.

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, 2008.

G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–65, 1936.

J. M Valdes-Sosa, P. A aand Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garca, and E. Canales-Rodriguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Neuroinformatics*, 360(1457):969, 2005.

H White. *Approximate Nonlinear Forecasting Methods*, chapter Handbook of Economic Forecasting, pages 460–512. Elsevier, New York, 2006.

H. White and X. Lu. Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193, 2010.

N. Wiener. The theory of prediction. *Modern mathematics for engineers, Series*, 1:125–139, 1956.

H. O. Wold. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell., 1938.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998.

G. U. Yule. Why do we sometimes get nonsense correlations between time series? a study in sampling and the nature of time series. *Journal of the Royal Statistical Society*, 89: 1–64, 1926.

A. Ziehe and K.-R. Mueller. Tdsep- an efficient algorithm for blind separation using time structure. *ICANN Proceedings*, pages 675–680, 1998.

S. T. Ziliak and D. N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press, February 2008. ISBN 0472050079.

Appendix A. Statistical significance tables for Type I and Type III errors

In order to assist practitioners in evaluating the statistical significance of bivariate causality testing, tables were prepared for type I and type III error probabilities as defined in (1) for different values of the base statistic. Below tables are provided for both the jacknifed statistic $\Psi/std(\Psi)$ and for the raw statistic Ψ, which is needed in case the number of points is too low to allow a jackknife/cross-validation/bootstrap or computational speed is at a premium. The spectral evaluation method is Welch's method as described in Section 6. There were 2000 simulations for each condition. The tables in this Appendix differ in one important aspect with respect to those in the main text. In order to avoid non-informative comparison of datasets which are, for example, analyses of the same physical process sampled at different sampling rates, the number of points is scaled by the 'effective' number of points which is essentially the number of samples relative to a simple estimate of the observed signal bandwidth:

$$N^* = N\tau_S/\widehat{BW}$$

$$\widehat{BW} = \frac{\|X\|_F}{\|\Delta X/\Delta T\|_F}$$

The values marked with an asterisk have values of both α and γ which are less than 5%. Note also that Ψ is non-dimensional index.

Table 4: α vs. $\Psi/std(\Psi)$

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.82	0.83	0.86	0.87	0.88	0.90	0.89	0.89	0.89
0.25	0.67	0.69	0.73	0.77	0.76	0.78	0.78	0.78	0.77
0.5	0.41	0.44	0.50	0.54	0.55	0.56	0.56	0.59	0.57
0.75	0.26	0.26	0.32	0.36	0.36	0.37	0.38	0.40	0.39
1	0.15	0.15	0.20	0.23	0.23	0.25	0.25	0.26	0.26
1.25	0.09	0.09	0.11	0.13	0.14	0.16	0.14	0.15	0.16
1.5	0.06	0.05	0.06	0.07	0.08	0.09	0.08	0.10	0.10
1.75	0.04	0.03	0.04	0.05	0.05 *	0.05 *	0.04 *	0.06	0.06
2	0.03	0.02	0.03	0.02 *	0.02 *	0.03 *	0.02 *	0.03 *	0.03 *
2.5	0.01	0.01	0.01	0.01 *	0.01 *	0.01 *	0.00 *	0.01 *	0.01 *
3	0.01	0.00	0.00	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
4	0.00	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
8	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
16	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 5: γ vs. $\Psi/std(\Psi)$

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.46	0.41	0.35	0.26	0.23	0.21	0.19	0.18	0.16
0.25	0.46	0.39	0.33	0.24	0.21	0.19	0.18	0.16	0.14
0.5	0.44	0.36	0.29	0.21	0.18	0.15	0.14	0.13	0.12
0.75	0.43	0.35	0.28	0.17	0.14	0.12	0.11	0.10	0.09
1	0.43	0.31	0.23	0.13	0.12	0.09	0.08	0.07	0.06
1.25	0.42	0.31	0.20	0.09	0.08	0.06	0.05	0.05	0.04
1.5	0.40	0.26	0.20	0.06	0.05	0.04	0.04	0.04	0.02
1.75	0.42	0.26	0.16	0.05	0.04 *	0.02 *	0.02 *	0.03	0.01
2	0.41	0.23	0.11	0.03 *	0.03 *	0.02 *	0.02 *	0.02 *	0.01 *
2.5	0.41	0.20	0.06	0.02 *	0.02 *	0.01 *	0.01 *	0.01 *	0.01 *
3	0.33	0.09	0.06	0.03 *	0.01 *	0.00 *	0.00 *	0.00 *	0.00 *
4	0.33	0.00 *	0.00 *	0.02 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
8	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
16	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 6: α vs. Ψ

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.25	0.22	0.25	0.24	0.24	0.24	0.21	0.21	0.15
0.25	0.09	0.09	0.09	0.09	0.10	0.09	0.09	0.08	0.05 *
0.5	0.02 *	0.01 *	0.01 *	0.02 *	0.02 *	0.02 *	0.01 *	0.01 *	0.01 *
0.75	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
1	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *

Table 7: γ vs. Ψ

$N^* \rightarrow$	**50**	**100**	**200**	**500**	**750**	**1000**	**1500**	**2000**	**5000**
0.125	0.21	0.17	0.14	0.10	0.08	0.07	0.06	0.05	0.03
0.25	0.11	0.08	0.06	0.04	0.03	0.03	0.02	0.02	0.01 *
0.5	0.03 *	0.01 *	0.01 *	0.01 *	0.01 *	0.00 *	0.00 *	0.01 *	0.00 *
0.75	0.02 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *
1	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *	0.00 *