Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

On Causality and Mutual Information

Victor Solo

Abstract— **We provide for the first time a formulation of nonlinear, nonstationary causality in terms of mutual information. We provide two fundamental mutual information identities; the first relating Granger type causality to Sims type causality. The second providing a decomposition of mutual information into a sum of Granger and Sims type terms. We also develop asymptotic relations that emerge under strict stationarity and generalise earlier work of Geweke. We relate our work in general with earlier developments.**

I. Introduction

Following Granger's development of an operational definition of temporal causality [1],[2], the attempt at empirically unraveling complex driving relations amongst macroeconomic variables from observational time series has played an important role in Econometrics. Granger causality has also been applied in bioengineering (aided by early work of [3]) e.g. [4],[5],[6],[7]. and more recently neuroimaging.

The theory underwent major development in the 1970s and 80s in Econometrics [8],[9],[10],[11],[12] and Systems and Control e.g.[13],[14],[15],[16]. But a number of significant issues remained and important work followed on nonlinearity and non-stationarity [17],and omitted variables [18]. More recently some new issues have arisen which have motivated the current work.

For completeness we mention the large statistical literature on empirical testing of causality in a static setting. The important recent book [19] has a guide to earlier literature. From that large literature we draw attention to the insightful article of [20].

In this work we are concerned to develop the basis of a theory of non-linear observational causality using mutual information. Based on the linear theory, there are several things such a nonlinear theory must accomplish;

(i) Provide definitions extending Granger and Sims causality and a result showing their equivalence.
(ii) Provide a means of measuring strength of causality by extending the linear measures provided by [10].
(ii) Provide a framework sufficient to support system identification.

While the connexion to mutual information has not previously been made in this generality, sub-pieces of this agenda have otherwise been previously accomplished. For (i) [12] Granger gave a definition of nonlinear Granger causality (GC).[17] gave a different defintion of nonlinear Granger causality as well as an extension of Sims causality and proved equivalence. For (ii) [21] and independently [22],[23] have given definitions of so-called directed information which are closely related to measures of strength of causality. Also [24] (see also [25][p631]) drew attention to a mutual information interpretation of the linear results of [10]. Finally as this paper was about to be submitted the author became aware of important work of [26] who develop a nonlinear causality theory in a Markov context that covers (i),(ii),(iii) to some extent.

In this work we formulate a theory of nonstationary nonlinear causality properties in terms of mutual information apparently for the first time. We then develop new results covering (i),(ii),(iii) and show how our work includes and extends the previous contributions.

The remainder of the paper is organized as follows. In section 2 we give a brief review of mutual information from a slightly unusual viewpoint. In section 3 we review previous work on nonlinear causality . In section 4 we develop our mutual information framework. In section 5 we obtain some asymptotics under a strict stationarity assumption. Conclusions are in section 6.

A. Notation and Acronyms

For $n > a$, X_a^n denotes $(x_a, x_{a+1}, \cdots, x_n)$; and then $X_n^n \equiv x_n$. We also use a shorthand notation: X^0 = current value =x_{n+1}; X^- = past values = X_1^n; X^+ = future values = X_{n+2}^∞. However we will need to define this last notation more carefully below.

AOD denotes analysis of deviation; AOI denotes analysis of information; ARMA denotes autoregressive moving average; ChRu denotes chain rule; GC denotes Granger causality; HMM denotes hidden Markov model; LR denotes likelihood ratio; LRT denotes likelihood ratio test; LHS denotes left hand side; MI denotes mutual information; MP denotes Markov process; RHS denots right hand side; VAR denotes vector autoregression; SC denotes Sims cauality; SSY denotes strict stationarity.

II. Mutual Information Review

We first review some basic aspects of mutual information and then assemble some properties that are crucial for our subsequent discussion. We refer to [27] for complete definitions and some proofs.

For a continuous valued random vector X with probability density function $p(x)$ the (differential) entropy is h(X)=-$E(\ln p(X))$ and the conditional entropy is $h(Y|X) = -E[\ln p(Y|X)]$. A fundamental property is the (conditional) entropy chain rule (ChRu)

$$h(X, Y|Z) = h(X|Z) + h(Y|X, Z)$$

V Solo is with School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, AUSTRALIA. v.solo@unsw.edu.au

Continuing we define the mutual information (MI)

$$\begin{aligned} I(X;Y) &= E[ln\frac{p(X,Y)}{p(X)q(Y)}] \\ &= h(X)+h(Y)-h(X,Y) \end{aligned}$$

and similarly the conditional MI

$$\begin{aligned} I(X;Y|Z) &= E[ln\frac{p(X,Y|Z)}{p(X|Z)q(Y|Z)}] \\ &= h(X|Z)+h(Y|Z)-h(X,Y|Z) \end{aligned}$$

By the entropy chain rule this is

$$\begin{aligned} &= h(X|Z)-h(X|Y,Z) \\ &= h(Y|Z)-h(Y|X,Z) \end{aligned}$$

Note the symmetry

$$I(Y;X|Z)=I(X;Y|Z)$$

And importantly MI inherits,

Property: ChRu. Chain Rule

$$I(X;(W,Y)|Z)=I(X;Y|Z)+I(X;W|Y,Z)$$

Finally Jensen's inequality shows (conditional)MI is non-negative

$$I(X;Y|Z)\geq 0$$

This leads immediately to a fundamental property of entropy:

Property CRE Conditioning reduces entropy.

$$h(X|Z)\geq h(X|Y,Z)$$

Now we can deduce some special properties of (conditional) MI needed in our subsequent discussion.

Property: RED Redundancy

$$I(Y;(X,Z)|Z)=I(Y;X|Z)$$

Property: DRI Discard reduces information

$$I(X;(Y,W)|Z)\geq I(X;Y|Z)$$

Property: DCRI Discard Conditioning reduces information

$$I(X;(Y,W)|Z)\geq I(X;W|Y,Z)$$

RED follows by substitution into the definition, while **DRI** ,**DCRI** follow from the chain rule .

It proves useful for our subsequent development to re-express MI in terms of LRs,

$$\begin{aligned} I(X;Y) &= E[ln\frac{q(Y|X)}{q(Y)}] \\ &= E[ln\frac{p(X|Y)}{p(X)}] \end{aligned}$$

This exhibits MI in two different ways as an expected LR. The expressions are interesting since at first sight the symmetry is not apparent. Similarly we can express conditional MI

$$\begin{aligned} I(X;Y|Z) &= E[ln\frac{q(Y|X,Z)}{q(Y|Z)}] \\ &= E[ln\frac{p(X|Y,Z)}{p(X|Z)}] \end{aligned}$$

We now use these alternative expressions for MI to derive the chain rule from an elementary nested decomposition of LR. We have

$$\begin{aligned} & ln\frac{p(X|W,Y,Z)}{p(X|Z)} \\ = \; & ln\frac{p(X|W,Y,Z)}{p(X|Y,Z)}+ln\frac{p(X|Y,Z)}{p(X|Z)} \end{aligned}$$

and taking expectations delivers the conditional chain rule .

If we introduce the sample MI

$$\begin{aligned} i(X;Y) &= ln\frac{p(X,Y)}{p(X)q(Y)} \\ &= ln\frac{q(Y|X)}{q(Y)}=ln\frac{p(X|Y)}{p(X)} \end{aligned}$$

and then the correspondingly defined sample conditional MI $i(X;Y|Z)$ then the nested LR decomposition is in fact a sample conditional ChRu.

$$i(X;(W,Y)|Z)=i(X;Y|Z)+i(X;W|Y,Z)$$

We note that sample conditional MI is not new, having been defined in [28].

III. Nonlinear Causality

In [12] Granger extended his notion of causality to the nonlinear case.

Definition GC. Y does not cause X, if

$p(x_{n+1}|X_1^n,Y_1^n)=p(x_{n+1}|X_1^n), n\geq 1$

In the Gaussian case this is equivalent to the standard linear definition

$var(x_{n+1}|X_1^n,Y_1^n)=var(x_{n+1}|X_1^n), n\geq 1.$

Actually this is what [9],[29] call weak causality. Strong causality entials additional conditioning on y_{n+1} on the LHS. We develop results here for weak causality only due to lack of space. However corresponding results for strong causality can be developed in a straightforward way.

[17] provided a significant development of nonlinear causality via the framework of conditional independence.

Definition. Given random vectors X,Y,Z we say X,Y are conditionally independent given Z and written

$X\perp Y|Z$ if $p(X,Y|Z)=p(X|Z)q(Y|Z)$.

[17] establish some basic properties of conditional indpendence and then define nonlinear versions of GC and SC. Their definitions involve infinite sequences. We modify them to apply to finite sequences as follows.

Definition GC$'$.

Y does not cause X if $x_{n+1}\perp Y_1^n|X_1^n, n\geq 1$.

Definition SFMC. Y does not cause X if

$X_{n+1}^{n+m}\perp y_n|X_1^n,Y_1^{n-1}, n\geq 1, m\geq 1$.

It is straightforward to see that GC agrees with GC$'$. Indeed by GC$'$, $p(x_{n+1},Y_1^n|X_1^n)=p(x_{n+1}|X_1^n)p(Y_1^n|X_1^n)$ But the LHS is also $p(x_{n+1}|Y_1^n,X_1^n)p(Y_1^n|X_1^n)$ and so $p(x_{n+1}|Y_1^n,X_1^n)=p(x_{n+1}|X_1^n)$ which is just GC.

In a similar way we find that SFMC is equivalent to

$p(y_{n+1}|X_{n+1}^{n+m},X_1^n,Y_1^{n-1})=p(y_{n+1}|X_1^n,Y_1^{n-1})$

Compared to the linear case [8] this has extra conditioning on Y_1^{n-1}. The significance of this will be clear shortly but

it explains why we have added the initials of [17] to the acronym.

Using an intermediate result of Bahadur [30], [17] establish the fundamental equivalence result.

Theorem 1. GC$'$ ≡SFMC.

This result again involves infinite sequences. Below we will prove a finite sequence version. Some comments on an infinite sequence version are made in section 5.

The situation with measures of strength of causality is somewhat tangled and we postpone a discussion until later.

IV. Mutual Information and Causality

To develop our theory and relate it to earlier work we begin with a well known result which expresses conditional independence in terms of MI,

Theorem 2. $X \perp Y|Z$ iff $I(X;Y|Z) = 0$

Proof. The result is well known and in any case easily proved by using elementary properties of Kullback-Liebler information [27].

The result is closely associated with the so-called data processing inequality which says that if $X \perp Y|Z$ then $I(X;Y) \geq I(X;Z)$. This result follows from Theorem 2 and the chain rule .

With Theorem 2 we can re-express the earlier definitions as follows,

Definition GC$'$. Y does not cause X iff
$I(x_{n+1}; Y_1^n|X_1^n) = 0, n \geq 1$.

Definition SFMC. Y does not cause X iff
$I(y_n; X_{n+1}^{n+m}|X_1^n, Y_1^{n-1}) = 0, n \geq 1, m \geq 1$

To provide a MI proof of Theorem 1 we develop a new and fundamental identity.

Theorem 3. GSFM Identity.

$$\Sigma_1^{N-1} I(y_n; X_{n+1}^N|X_1^n, Y_1^{n-1}) = \Sigma_1^{N-1} I(x_{n+1}; Y_1^n|X_1^n)$$

for $N \geq 2$, and where Y_1^0 is omitted from the term where it appears.

Proof. Denote $z_n = (x_n, y_n)$ etc. We expand $h(Z_1^N)$ in two different ways. By the chain rule

$$\begin{aligned} h(Z_1^N) &= \Sigma_2^N h(z_n|Z_1^{n-1}) + h(z_1) \\ &= \Sigma_2^N h(x_n, y_n|X_1^{n-1}, Y_1^{n-1}) + h(x_1, y_1) \end{aligned}$$

Now apply the chain rule again to find

$$\begin{aligned} h(Z_1^N) &= \Sigma_2^N h(x_n|X_1^{n-1}, Y_1^{n-1}) + h(x_1) \\ &+ \Sigma_2^N h(y_n|X_1^n, Y_1^{n-1}) + h(y_1|x_1) \end{aligned}$$

On the other hand we can also use the ChRu to write

$$h(Z_1^N) = h(X_1^N) + h(Y_1^N|X_1^N)$$

Applying the ChRu yet again gives

$$\begin{aligned} h(Z_1^N) &= \Sigma_2^N h(x_n|X_1^{n-1}) + h(x_1) \\ &+ \Sigma_2^N h(y_n|Y_1^{n-1}, X_1^N) + h(y_1|X_1^N) \end{aligned}$$

Now rewrite the third term to get

$$\begin{aligned} & h(Z_1^N) = \Sigma_2^N h(x_n|X_1^{n-1}) + h(x_1) \\ + \; & \Sigma_2^{N-1} h(y_n|Y_1^{n-1}, X_1^n, X_{n+1}^N) + h(y_N|Y_1^{N-1}, X_1^N) \\ + \; & h(y_1|X_1^N) \end{aligned}$$

Now equate the two expressions for $h(Z_1^N)$ and reorganize to obtain

$$\begin{aligned} & \Sigma_2^N I(x_n; Y_1^{n-1}|X_1^{n-1}) \\ = \; & \Sigma_2^N (h(x_n|X_1^{n-1}) - h(x_n|X_1^{n-1}, Y_1^{n-1})) \\ = \; & \Sigma_2^{N-1} h(y_n|Y_1^{n-1}, X_1^n, X_{n+1}^N) + h(y_N|Y_1^{N_1}, X_1^N) \\ - \; & \Sigma_2^N h(y_n|X_1^n, Y_1^{n-1}) + h(y_1|X_1^N) - h(y_1|x_1) \\ = \; & \Sigma_2^{N-1} (h(y_n|Y_1^{n-1}, X_1^n, X_{n+1}^N) - h(y_n|X_1^n, Y_1^{n-1}) \\ + \; & h(y_1|X_1^N) - h(y_1|x_1) \\ = \; & \Sigma_2^{N-1} I(y_n; X_{n+1}^N|Y_1^{n-1}, X_1^n) + I(y_1; X_2^N|x_1) \\ = \; & \Sigma_1^{N-1} I(y_n; X_{n+1}^N|Y_1^{n-1}, X_1^n) \end{aligned}$$

and a change of summation variables on the LHS delivers the result.

Remark. Naturally the identity can be written with y, x interchanged.

Using the GSFM identity we can now establish a finite sequence version of Theorem 1 with a MI proof,

Theorem 1$'$. GC$'$ ≡SFMC.

Proof. Given $N \geq 1$, if GC$'$ holds then each term in the sum on the RHS of the GSFM identity vanishes. Then the LHS vanishes but again since MI is non-negative each term in the LHS sum vanishes. Since N is arbitary this means SFMC holds. The reverse argument is the same.

With this result in hand it is natural to introduce measures of strength of causality.

Definition. Strength of causality measures.

$$\begin{aligned} F_{Y \to X,N}^G &= \frac{1}{N} \Sigma_1^{N-1} I(x_{n+1}; Y_1^n|X_1^n), N \geq 2 \\ F_{Y \to X,N}^{SFM} &= \frac{1}{N} \Sigma_1^{N-1} I(y_n; X_{n+1}^N|X_1^n Y_1^{n-1}), N \geq 2 \end{aligned}$$

Of course the GSFM identity ensures they are equal but the separate definitions will prove useful below. There are analagous definitions of $F_{X \to Y,N}^G, F_{X \to Y,N}^{SFM}$.

To develop system identification methods we now develop a fundamental nested decomposition of LR.

Theorem 4. Analysis of Deviance (AOD).

$$\begin{aligned} & ln \frac{p(Y_1^N, X_1^N)}{q(Y_1^N) p(X_1^N)} = ln \frac{p(Y_1^N|X_1^N)}{p(Y_1^N)} \\ = \; & \Sigma_1^{N-1} ln \frac{p(y_n|X_1^n, X_{n+1}^N, Y_1^{n-1})}{p(y_n|X_1^n, Y_1^{n-1})} \\ + \; & \Sigma_1^N ln \frac{p(y_n|X_1^n, Y_1^{n-1})}{p(y_n|X_1^{n-1}, Y_1^{n-1})} \\ + \; & \Sigma_2^N ln \frac{p(y_n|X_1^{n-1}, Y_1^{n-1})}{p(y_n|Y_1^{n-1})} \end{aligned}$$

where, on the RHS Y_1^0 is omitted where it occurs so e.g. $p(y_1|Y_1^0) \equiv p(y_1)$.

Proof. By iterated conditional probability we have

$$\begin{aligned} ln\frac{p(Y_1^N,X_1^N)}{q(Y_1^N)p(X_1^N)} &= ln\frac{p(Y_1^N|X_1^N)}{p(Y_1^N)} \\ = \Sigma_1^N ln\frac{p(y_n|X_1^N,Y_1^{n-1})}{p(y_n|Y_1^{n-1})} \\ = \Sigma_1^N ln\frac{p(y_n|X_1^n,X_{n+1}^N,Y_1^{n-1})}{p(y_n|X_1^{n-1},Y_1^{n-1})} \\ + \Sigma_1^N ln\frac{p(y_n|X_1^{n-1},Y_1^{n-1})}{p(y_n|Y_1^{n-1})} \\ = \Sigma_1^{N-1} ln\frac{p(y_n|X_1^n,X_{n+1}^N,Y_1^{n-1})}{p(y_n|X_1^n,Y_1^{n-1})} \\ + \Sigma_1^N ln\frac{p(y_n|X_1^n,Y_1^{n-1})}{p(y_n|X_1^{n-1},Y_1^{n-1})} \\ + \Sigma_2^N ln\frac{p(y_n|X_1^{n-1},Y_1^{n-1})}{p(y_n|Y_1^{n-1})} \end{aligned}$$

where in the first term the Nth term vanishes and in the third term the first term vanishes. The result is thus established.

Remarks. We can interpret the LHS as a LRT for testing the hypothesis: Y_1^N depends on X_1^N, versus the alternative that it does not. The AOD gives a decomposition of this LRT into a sum of LRTs corresponding to a decomposition of the overall hypotheses into a sequence of nested hypotheses:

Y^0 depends on past y and current x, past x and future x versus Y^0 depends on past y and past x and current x.

Y^0 depends on past y and past x and current x versus Y^0 depends on past y and past x.

Y^0 depends on past y and past x versus Y^0 depends on past y.

We have given the decomposition in terms of y regressed on x. We can of course write it the other way round.

Each of the LRs is called a deviance in the statistics literature e.g. [31]. And this kind of nested decomposition is well known in statistics as a generalization of analysis of variance; hence the name [31]. However, the particular AOD in Theorem 4 appears to be new. It is this kind of nested decomposition that is required for a system identification procedure for testing causality. This is made clear by the next result.

Theorem 5. Analysis of Information (AOI).

$$\begin{aligned} I(Y_1^N;X_1^N) &= \Sigma_1^{N-1} I(y_n;X_{n+1}^N|Y_1^{n-1},X_1^n) \\ &+ \Sigma_1^N I(y_n;x_n|X_1^{n-1},Y_1^{n-1}) \\ &+ \Sigma_1^{N-1} I(y_{n+1};X_1^n|Y_1^n) \end{aligned}$$

Proof. Take expectations in the AOD and then change the summation index in the third term.

If we now introduce a mesure of instantaneous causality.
Definition. Instantaneous Causality Measure.

$$F_{XoY,N} = \frac{1}{N}\Sigma_1^N I(y_n;x_n|X_1^{n-1},Y_1^{n-1})$$

Then we have the following result.

Corollary. Decomposition of Measures of Causality.

$$\begin{aligned} \frac{1}{N}I(Y_1^N;X_1^N) &= F_{Y\to X,N}^{SFM} + F_{XoY,N} + F_{X\to Y,N}^{G} \\ &= F_{Y\to X,N}^{G} + F_{XoY,N} + F_{X\to Y,N}^{SFM} \\ &= F_{Y\to X,N}^{G} + F_{XoY,N} + F_{X\to Y,N}^{G} \\ &= F_{Y\to X,N}^{SFM} + F_{XoY,N} + F_{X\to Y,N}^{SFM} \end{aligned}$$

Proof. The first line is just AOI. The second line is the AOI written with X, Y interchanged. The subsequent lines (and also line 2) follow by using the GSFM identity.

Remarks. We can now make connexions with previous work.

(i) In [21] there is a result which we may interpret as being line 4 of the corollary. [21] do not discuss causality, do not obtain the GSFM identity and do not obtain the AOD. No derivation of their result is given and there is some vagueness in the notation (their basic result has a sum over n on the RHS but n also appears unsummed on the LHS!). However it seems clear that it is line 4 that is being discussed.

(ii) In [22] there is a definition of directed information (DMI) that equals our $F_{Y\to X,N}^G + F_{XoY,N}$. And in [23] is a result (the conservation law) which gives essentially line 3 but with the first two terms amalgamated as above. Neither reference discusses causality, nor obtains the GSFM identity nor obtains the AOD. However [22] does have a result showing that DMI equals the LHS iff 'there is no feedback from Y to X'. This is very close to a strong version of GC.

(iii) The important results of [26] are closest to ours. As noted earlier we came on this paper after a draft of our paper was completed. Our AOD was inspired by comments of [32] following the paper [10]. Once one has the AOD then the AOI is trivial. [26] obtain a GSFM identity, an AOD and an AOI (not under these names). However their results are under a Markov process assumption on $z_n = (x_n, y_n)^T$. This is not a good assumption because even if x_n, y_n are jointly Markov then neither x_n nor y_n are marginally Markov. This is true even in the linear case. e.g. if z_n is VAR(1) then x_n, y_n are each ARMA(1,1) which is not Markov. This fact makes the results ungainly. Thus the GSFM identity does not have the nice form of our result; rather there is a Markov term that must be added to $F_{Y\to X,N}^G$ to get $F_{Y\to X,N}^{SFM}$. Also [26] are unaware of MI and so all results are in terms of Kullback-Liebler information. This means that the underlying simplicity of the expressions is lost. Also they do not obtain Theorem 1 as we have.

V. **Stationarity**

In this section we discuss an asymptotic version of the above results that occurs under strict stationarity (SSY). This allows a precise connexion to the results of [10].

But first we need a preliminary result. Below we will be introducing various variations of entropy rates [27]. Such rates are well defined for discrete valued SSY processes, since proofs rely heavily on the fact that entropy is non-negative. But the situation for analog processes is much harder since

(differential) entropy is no longer necessarily positive (e.g. for a Gaussian white noise just take the noise variance to be arbitarily small) and there is no other elementary lower bound [1]. For analog Gaussian processes results are well known [27] but we need to guarantee something beyond that to justify our discussion. Since there do not seem to be any general results we develop the following result applicable under hidden Markov assumptions.

Theorem 6. If z_n obeys a SSY hidden Markov model (HMM) i.e. $z_n = \phi(\xi_n)$ where $\phi(\cdot)$ is continuous and ξ_n is a SSY Markov process (MP) then $h(z_n|Z_1^{n-1})$ converges to a finite limit which we denote $h(Z^0|Z^-)$.

Remarks.

(i) This result establishes the existence of an analog entropy rate for a reasonably general class of nonlinear processes.

(ii) A simple and well known example of such a HMM is given by the SS model; $z_n = \psi(\zeta_n) + v_n; \zeta_{n+1} = \chi(\zeta_n, w_n)$ with given initial condition ζ_o and where w_n, v_n are independent identically distributed sequences and ψ, χ are continuous functions and where all appropriate finite order conditional entropies exist.

Proof. We adapt a standard argument developed for different purposes in [27](section 4.4). Firstly consider that

$$\begin{aligned} h_{n+1} &= h(z_{n+1}|Z_1^n) \\ &\leq h(z_{n+1}|Z_2^n), \text{ by CRE} \\ &= h(z_n|Z_1^{n-1}), \text{ by SSY} \end{aligned}$$

So h_n is a non-increasing sequence. Also by CRE $h_n \leq h(z_n) = h(Z^0)$. So h_n is upper bounded. We now need to develop a lower bound. But we now just repeat a standard argument outlined clearly in [27](p69) to find that for any fixed $n > 3$ and any $k > 0$

$$h(z_n|Z_2^{n-1}, \xi_1) \leq h_{n+k}$$

Thus as a sequence in k, h_{n+k} is bounded below. Thus since it is non-increasing and bounded below and above it must have a limit which we denote as $h(Z^0|Z^-)$. The result is thus established.

Further Remark.

If we partition $z_n = (x_n, y_n)$ then by reworking the argument of [27](p69) we can similarly establish the existence of limits for quantities such as $h(y_n|X_1^n, Y_1^{n-1}), h(x_n|X_1^{n-1})$ needed below.

Let us return to the GSFM identity and note the four entropy terms.

$$\begin{aligned} A_N = h(X_1^N) &= \Sigma_1^N h(x_n|X_1^{n-1}) \\ B_N = h(Y_1^N|X_1^N) &= \Sigma_1^N h(y_n|Y_1^{n-1}, X_1^n, X_{n+1}^N) \\ &= \Sigma_1^N h(y_n|Y_1^{n-1}, X_1^N) \\ C_N &= \Sigma_1^N h(x_n|X_1^{n-1}, Y_1^{n-1}) \\ D_N &= \Sigma_1^N h(y_n|X_1^n, Y_1^{n-1}) \\ A_N + B_N &= C_N + D_N \end{aligned}$$

[1]There is an erroneous claim in [27](section 11.5) suggesting the discrete valued theory carries over.

Consider the fourth term D_N. Then by CRE

$$\begin{aligned} h(y_n|X_1^n, Y_1^{n-1}) &\leq h(y_n|X_2^n, Y_2^{n-1}) \\ &= h(y_{n-1}|X_1^{n-1}, Y_1^{n-2}) \end{aligned}$$

by SSY. So $h(y_n|X_1^n, Y_1^{n-1})$ is a non-increasing sequence. It is bounded above by $h(y_n) = h(Y^0)$. So it either converges or diverges to $-\infty$. In view of theorem 6 we assume the limit is finite. We denote the limit $h(Y^0|X^-, X^0, Y^-)$. Thus $\frac{1}{N}D_N \to h(Y^0|X^-, X^0, Y^-)$.

Similarly $\frac{1}{N}C_N \to h(X^0|X^-, Y^-)$ and $\frac{1}{N}A_N \to h(X^0|X^-)$. The B_N term needs more careful treatment. We show B_N is subadditive i.e. $B_{N+J} \leq B_N + B_J$. It then follows by the subadditive limit theorem [33],[34] that $\frac{1}{N}B_N \to inf_n(B_n/n)$ which we denote as $h(Y^0|Y^-, X^-, X^0, X^+)$.

With these results we obtain.

Theorem 7. Under the assumptions of Theorem 6 we have the: Stationary GSFM identity

$$\begin{aligned} lim\frac{1}{N}F_{Y\to X,N}^G &= F_{Y\to X}^G \\ &= I(X^0; Y^-|X^-) \\ &= I(Y^0; X^+|X^-, X^0, Y^-) \\ &= F_{Y\to X}^{SFM} = lim\frac{1}{N}F_{Y\to X,N}^{SFM} \end{aligned}$$

Proof. Since $A_N + B_N = C_N + D_N$ we have of course the GSFM identity,

$$F_{Y\to X,N}^G = \frac{1}{N}(A_N - C_N) = \frac{1}{N}(D_N - B_N) = F_{Y\to X,N}^{SFM}$$

from which it follows on taking limits as above, that

$$\begin{aligned} & F_{Y\to X}^G = I(X^0; Y^-|X^-) \\ = \; & h(X^0|X^-) - h(X^0|X^-, Y^-) \\ = \; & h(Y^0|X^-, X^0, Y^-) - h(Y^0|Y^-, X^-, X^0, X^+) \\ = \; & I(Y^0; X^+|X^-, X^0, Y^-) = F_{Y\to X}^{SFM} \end{aligned}$$

We have only now to prove subbadditivity of B_N. We have

$$\begin{aligned} B_{N+J} &= \Sigma_1^{N+J} h(y_n|Y_1^{n-1}, X_1^{N+J}) \\ &= \Sigma_1^N h(y_n|Y_1^{n-1}, X_1^{N+J}) \\ &+ \Sigma_{N+1}^{N+J} h(y_n|Y_1^{n-1}, X_1^{N+J}) \end{aligned}$$

Using CRE on the first term shows it is bounded by

$$\Sigma_1^N h(y_n|Y_1^{n-1}, X_1^N) = B_N$$

For the second term change summation indices to $l = n - N$ to obtain

$$= \Sigma_1^J h(y_{l+N}|Y_1^{l+N-1}, X_1^{N+J})$$

Now use CRE to see this is

$$\leq \Sigma_1^J h(y_{l+N}|Y_{N+1}^{l+N-1}, X_{N+1}^{N+J})$$

and by SSY this is

$$= \Sigma_1^J h(y_l|Y_1^{l-1}, X_1^J) = B_J$$

and subbaditivity is established and so the result.

Using these results we have.

Theorem 8. Stationary AOI Asymptotics.

$$\begin{aligned} lim\frac{1}{N}I(Y_1^N;X_1^N) &= I(Y;X) \\ &= F_{Y\to X}^{SFM}+F_{X\to Y}^{0}+F_{X\to Y}^{G} \\ F_{Y\to X}^{0} &= I(X^0;Y^0|X^-,Y^-) \end{aligned}$$

Proof. Following the same argument as used in the first part of Theorem 6 we find $F_{XoY,N}$ converges to a limit which we denote $I(X^0;Y^0|X^-,Y^-)=F_{Y\to X}^0$. This establishes that $\frac{1}{N}$RHS of the finite data AOI in Theorem 5 converges to the RHS quoted here. Thus the $\frac{1}{N}$LHS also converges and we have denoted the limit $I(Y;X)$. One can also establish that the limit exists directly by a subbadditivity argument.

Remarks.

(i) In the Gaussian case the limiting MI expressions are given by well known formulae e.g.[27]. Using these formulae in Theorems 6,7 give the results of [10]. This gives a MI interpretation of [10] results which is already implicit from [24],[25].

(ii) In [26], their Theorem 5, provides some stationarity asymptotics (under a Markov assumption on z_n) related to our Theorem 8. But their result is incomplete because they have an assumption A which essentially assumes a certain entropy rate limit exists. Such a result cannot be obtained without a result such as our Theorem 6 which they do not have. We note also their proof does not use subadditivity.

(iii) The infinite sequence version of Theorem 1 established by [17] can be established under SSY using the results of the last two sections. Space limits preclude the details which will be developed elsewhere.

VI. Conclusion

In this paper we have provided, for the first time, a formulation of nonlinear, nonstationary causality in terms of mutual information.

We have provided a fundamental finite data identity relating SFMC to GC thereby providing a new proof of earlier results of [17].

We have also provided a fundamental AOD and its corresponding AOI. Earlier workers had special cases of these results. The AOD is important for system identification.

Finally we proved asymptotic versions of these results valid under strict stationarity. This reproduces results of [10] as well as providing a nonlinear generalization of them.

Our formulation in terms of mutual information is also important since it provides an operational definition suitable for empirical use e.g. system identification.

Acknowledgement. To the referees for useful comments.

References

[1] CWJ Granger, "Economic processes involving feedback", *Information and Control*, vol. 6, pp. 28–48, 1963.

[2] C.W.J. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica*, vol. 37, 1969.

[3] W Gersch, "Causality or driving in electrophysiological signal analysis", *Math Biosciences*, vol. 14, pp. 177–196, 1972.

[4] SM Schnider, RH Kwong, FA Lenz, and HC Kwan, "Detection of feedback in the central nervous system using system identification techniques", *Biol. Cyb.*, vol. 60, pp. 203–212, 1989.

[5] MJ Kaminski and KJ Blinowska, "A new method of the description of the information flow in the brain structures", *Biol. Cyb.*, vol. 65, pp. 203–210, 1991.

[6] C Bernasconi and P Konig, "On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings", *Biol. Cyb.*, vol. 81, pp. 199–210, 1999.

[7] Y Chen, G Rangarajan, J Feng, and M Ding, "Analyzing multiple nonlinear time series with extended granger causality", *Phys Lett A*, vol. 324, pp. 26–35, 2003.

[8] C A Sims, "Money, income and causality", *American Economic Review*, vol. 62, pp. 540–552, 1972.

[9] DA Pierce and LD Haugh, "Causality in temporal systems characterizations and a survey", *Journal of Econometrics*, vol. 5, pp. 265–293, 1977.

[10] J Geweke, "The measurement of linear dependence and feedback between multiple time series", *Jl Amer Stat Assoc*, vol. 77, pp. 304–313, 1982.

[11] C Hsiao, "Time series modeling and causal ordering of canadian money income and interest rates", in *Time Series Analysis: Theory and Practice I*. North Holland, 1982, pp. 671–699.

[12] CWJ GRANGER, "Testing for causality", *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, 1980.

[13] P E Caines and W W Chan, "Feedback between stationary stochastic processes", *IEEE Trans Autom Contr*, vol. 20, pp. 498–508, 1975.

[14] PE Caines, "Weak and strong feedback free processes", *IEEE. Trans Auto Contr*, vol. 21, pp. 737–739, 1976.

[15] MR Gevers and BDO Anderson, "On jointly stationary feedback-free stochastic processes", *IEEE Trans Autom Ctr*, vol. AC-27, pp. 431–436, 1982.

[16] V Solo, "Topics in advanced time series analysis", in *Lectures in Probability and Statistics*. 1986, pp. 165–328, G del Pino and R Rebolledo , eds; Springer-Verlag.

[17] JP Florens and M Mouchart, "A note on non-causality", *Econometrica*, vol. 50, pp. 583–592, 1982.

[18] Y Hosoya, "Elimination of third series effect and defining partial measures of causality", *Jl Time Series Analysis*, vol. 22, pp. 537–554, 2001.

[19] J Pearl, *Causality: Models Reasoning and Inference*, CUP, Cambridge, UK, 2000.

[20] D R Cox, "Causality: Some statistical aspects", *Jl Royal Stat Soc A*, vol. 155, pp. 291–301, 1992.

[21] T Kamitake, H Harashima, and H Miyakawa, "A time-series analysis method based on the directed trans-information", *Electronics and Communication in Japan*, vol. 67-A, pp. 103–110, 1984.

[22] JL Massey, "Causality, feedback and directed information", in *Proc. 1990 Intl. Symp. on Info. Th. and its Applications, Waikiki, Hawaii, Nov. 27-30, 1990*, 1990.

[23] JL Massey and PC Massey, "Conservation of mutual and directed information", 2004.

[24] J Rissannen and M Wax, "Measures of mutual and causal dependence between two time series", *IEEE TRANSACTIONS ON INFORMATION THEORY*, vol. 33, pp. 598–601, 1987.

[25] P.E. Caines, *Linear Stochastic Systems*, J Wiley, New York, 1988.

[26] C Gourieroux, A Monfort, and E Renault, "Kullback causality measures", *Annales d'Economie et de Statistique*, vol. 6/7, pp. 369–410, 1987.

[27] T Cover and J Thomas, *Elements of Information Theory*, J Wiley, New York, 1991.

[28] MS Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day London, New York, 1964.

[29] PE Caines and CW Chan, "Estimation, identification and feedback", in *System Identification:Advances and case Studies*. Academic Press, 1976, pp. 349–405.

[30] RR Bahadur, "Sufficiency and statistical decision functions", *Ann Math Stat*, vol. 25, pp. 423–462, 1954.

[31] P McCullagh and J Nelder, *Generalized Linear Models*, Chapman and Hall 1983, London, 1983.

[32] E Parzen, "Comment on: Measurement of linear dependence and feedback between multiple time series", *Jl Amer Stat Assoc*, vol. 77, pp. 320–322, 1982.

[33] E HILLE and RS PHILLIPS (1957), *Functional Analysis and Semi-Groups*, American Mathematical Society, Providence, RI,USA, 1957.

[34] JFC Kingman, "Subadditive ergodic theory", *Ann Prob*, vol. 1, pp. 883–899, 1973.