

CAUSALITY CHARACTERIZATIONS: BIVARIATE,
TRIVARIATE, AND MULTIVARIATE PROPOSITIONS

Gary R. Skoog

November 1976

Federal Reserve Bank of Minneapolis

Staff Report #14

The views expressed herein are solely those of the author and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Causality Characterizations: Bivariate, Trivariate,
and Multivariate Propositions

I. Introduction

In this paper, several propositions are proved which relate to the concept of causality or exogeneity in multivariate weakly stationary stochastic processes. From a mathematical viewpoint, the results concern certain projections in Hilbert space, a fact which suggests standards for proofs--standards which are, lamentably, lacking in much of this literature. ([24] is the obvious exception.) Theorems about these projections would be of limited interest, however, were it not for a natural interpretation of these concepts to causality in multiple time series, due originally to Wiener [29]. This interpretation and the propositions of Granger [4] and Sims [24] have given rise to a flourishing empirical literature, including, but not limited to, [3], [5], [18], [21]. It is hoped that the results proved here, which, for the most part, give necessary and sufficient conditions for causality in terms of structural aspects of the time series involved, will further the understanding of this concept and these empirical results.

Propositions 1 and 2 strengthen and generalize theorems of Granger [4] and Sims [24]: specialists may find these proofs

of independent interest because they differ in character from their predecessors. Proposition 3 was proved when the author was unaware of three unpublished works (Haugh, [8], Haugh and Box [9], and Pierce and Haugh [15]). These papers independently arrive, via "operational methods" at what is perhaps a special case of the present result. The difference in emphasis between our papers--they stress the identification (in both the econometrician's sense and the time series analyst's sense) of models from data, whereas we stress intrinsic mathematical-logical properties of the wide sense stationary process--is reflected in a difference in the languages employed; this complicates a comparison of our papers, but some comments on this subject are made at the end of Section III.

Propositions 4 and 5 continue to deal with bivariate systems, but set out in a new direction. Any definition of a concept as loaded with philosophical connotations as is causality must be able to withstand severe scrutiny. Here we inquire about the behavior of this definition under time reversal; equivalently, is time treated symmetrically with regard to the past and future? Specifically, assume according to the usual definition that Y does not cause X, by which we mean that past Y is of no additional help, given past X, in predicting current X; alternatively, X is said to be exogeneous. Is it now true that, again trying to predict the current X but now given future X, that future Y is of no marginal value? Were this the case, causality might be said to be neutral with respect to the

flow of time. Whether this latter property would enhance the definition is academic, because we give (restrictive) necessary and sufficient conditions for time neutrality to occur in terms of the Wold (bivariate moving average) representation, a Wold-like representation, and the population regression of current Y on past, current, and future X. A corollary notes that, with X exogenous, to predict current X, in general the prognosticator will prefer the future X,Y data to the past X,Y data--intuitively because in the latter situation he will find Y of no marginal use.

Next, in hopes of shedding some light on a common criticism of this methodology, we add a third series and consider the trivariate system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t)$. The sensitivity of a "Y causes X" finding to the underlying data set available for prediction has been appreciated from the start. But beyond the presumption that conclusions in lower order systems will be overturned in higher order systems and a suggestion by Granger that partial cross spectra be considered ([4], p. 437), little attention has been given to the analysis of systems of higher order than two. Certain natural definitions are made and a straightforward generalization to bloc-bivariate systems noted in Proposition 6. Then a more substantive result is proved for trivariate and bloc-trivariate systems. Propositions 7 and 8 closely examine the relationship between the events "X is exogeneous with respect to Y in the trivariate system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ "

and "X is exogeneous with respect to Y in the implied bivariate $\begin{pmatrix} X \\ Y \end{pmatrix}$ system." Using previous propositions, some relations are indicated so that the researcher is provided with a systematic way of using any information about a third process Z which may be available. Indeed, the result may be interpreted as an infinite dimensional omitted variables theorem. We stress that the word "finding" pertains to a condition about theoretical regressions or projections in the "population" (Hilbert space) which would be attained by consistent estimators; many thorny issues involving statistical estimation procedures are not discussed here.

Finally, some remarks on the economic significance of causality-exogeneity relationships are offered, followed by a conclusion and indication of some directions for further research.

II. Mathematical and Statistical Framework; Background and Definitions; Normalization-Identification Issues

In this section the definitions and notation employed in the rest of the paper are presented. Several very useful facts relating these notions are stated for ready reference. A few theorems in the prediction theory of multivariate stochastic processes which are important for our purposes are explicitly mentioned. For a comprehensive treatment of this entire topic, including proofs, the reader is referred to any or all parts of these excellent references: Rozanov [17], Hannan [7], and Wiener-Masani [30]. To make this part more readable and to offer documentation for some of these

assertions, extensive use of technical footnotes is made; these may be skimmed on a first reading.

Because the first part of this paper and most of the related econometrics and statistics literature deal with bivariate processes, we adopt this tack as an expository device here. Since the major complications introduced by the general n -variate mathematical theory are already present when $n=2$, there results neither a loss of generality nor a need for excessive repetition when multivariate situations are encountered.

On an underlying probability space Ω with accompanying σ -algebra of subsets F and probability measure P is defined a vector family of random variables (measurable functions), indexed by the discrete parameter t , $t \in I \equiv \text{integers}$, $\begin{pmatrix} X \\ Y \end{pmatrix}(t) \equiv \begin{pmatrix} X(t) \\ Y(t) \end{pmatrix}$,¹ which is the subject of our study. Following tradition, we have already suppressed the dependence of $\begin{pmatrix} X \\ Y \end{pmatrix}$ on $\omega \in \Omega$: $\begin{pmatrix} X \\ Y \end{pmatrix}(t, \omega)$ might have appeared more appropriate. Our notation reflects the fact that we will never investigate the behavior of sample paths (a sequence $\{\begin{pmatrix} X \\ Y \end{pmatrix}(t, \bar{\omega}), t = \dots -1, 0, 1, \dots \text{ for fixed } \bar{\omega}\}$) in the sequel, so there is no need to keep track of a second argument. We may require that $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ be a weakly stationary stochastic process (w.s.s.p.),² which means: (i) $E \begin{pmatrix} X \\ Y \end{pmatrix}(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, all $t \in I$; and (ii) the Gramian or auto-covariance matrix

$$\begin{aligned} \begin{pmatrix} X \\ Y \end{pmatrix}(t), \begin{pmatrix} X \\ Y \end{pmatrix}(t-k) &\equiv \Gamma_{X,Y}(t,k) \equiv E \begin{pmatrix} X(t)X(t-k) & X(t)Y(t-k) \\ Y(t)X(t-k) & Y(t)Y(t-k) \end{pmatrix} \\ &\equiv \begin{pmatrix} R_X(k) & R_{XY}(k) \\ R_{YX}(k) & R_Y(k) \end{pmatrix} \end{aligned}$$

does not depend on t , and so may be written $\Gamma_{X,Y}(k) (= \Gamma_{X,Y}^T(-k))$, where T denotes transpose). Here E denotes mathematical expectation, so that one effect of (ii) is that $X(t)$ and $Y(t)$ are in $L_2(\Omega, F, P)$, the space of all complex random variables $Z(\cdot)$ such that $\int |Z(w)|^2 dP(w) < \infty$. This latter space is a Hilbert space, H , with the inner product given by $\langle Z_1(w), Z_2(w) \rangle \equiv \int Z_1(w) \overline{Z_2(w)} dP(w)$ ³ and norm $\|Z_1\| \equiv (\int |Z_1(w)|^2 dP(w))^{1/2}$; these classical facts may be found in any analysis text, e.g., [10], p. 235. If $\langle Z_1, Z_2 \rangle = 0$, we write $Z_1 \perp Z_2$, call Z_1 and Z_2 orthogonal elements in H , and understand these symbols to say the random variables are uncorrelated (if EZ_1 or $EZ_2 = 0$, as will always be assumed).

More relevant for our purposes is a subspace⁴ (closed linear manifold) of H , the space of values of the process $\begin{pmatrix} X \\ Y \end{pmatrix}$, to be denoted $H_{X,Y}$ or $H_{X,Y}(-\infty, \infty)$. For any sets of integers $s_1, \dots, s_m; t_1, \dots, t_n$ and any sets of real or complex constants $a_1, \dots, a_m; b_1, \dots, b_n$ the finite linear combination

$$\sum_{i=1}^m a_i X(s_i) + \sum_{j=1}^n b_j Y(t_j)$$

is also a random variable in H . The

set of all such random variables will be indicated by

$$\left(\bigcup_{i \in I} X(i) \right) \cup \left(\bigcup_{j \in I} Y(j) \right) \text{ or } S(X(i), Y(j), i, j \in I);$$
⁵

this is by

definition a linear manifold of H , which is in general not closed in the topology of the norm. The closure of this set is defined to be $H_{X,Y}$ (or $H_{X,Y}(-\infty, \infty)$ to emphasize the set of times which may be used in forming combinations). $H_{X,Y}$ is also referred to as the past, present, and future of the $\binom{X}{Y}$ process and for our purposes may be regarded as the underlying Hilbert space, several of whose subspaces will command particular attention.

Let us regard the present as time t , and imagine that we possess a long data series extending into the remote past, $D(t) \equiv \{\binom{X}{Y}(s), s=t, t-1, \dots\}$ generated by the w.s.s.p. $\binom{X}{Y}$, a series sufficiently representative to yield perfect knowledge of the covariance sequence $\{\Gamma_s, s = \dots -1, 0, 1, \dots\}$. It is natural to pose the question: What is the "best" predictor of $\binom{X}{Y}(t+1)$ and what is the meaning of "best"? Since we do not know which elementary event ω has occurred, the meaning of "best" will have to involve some statistical or averaging criterion; by predictor, we mean Borel function measurable with respect to the σ -algebra generated by D . If our statistical criterion is now to minimize mean square error, the best predictor will be a conditional expectation; proceeding to give an effective formulation for the solution will be quite difficult and will involve hard analysis in stochastic process theory. If, however, we restrict ourselves to linear predictors (those in $H_{X,Y}(-\infty, t)$; this subspace may hereafter be abbreviated as $H_{X,Y}(t)$ when no confusion will arise) and

if we maintain the criterion of minimizing mean square error, then finding the optimal predictor for $X(t+1)$ involves projecting⁶ $X(t+1)$ onto $H_{X,Y}(t)$, and similarly for $Y(t+1)$. (Since so much use is made of the concept of projection, footnote 6 provides an extensive discussion of this and related topics.)

These (orthogonal) projections always exist and will be denoted $(X(t+1)|H_{X,Y}(t))$ and $(Y(t+1)|H_{X,Y}(t))$, respectively. Consequently, there result the orthogonal decompositions $X(t+1) = (X(t+1)|H_{X,Y}(t)) + u(t+1)$ and $Y(t+1) = (Y(t+1)|H_{X,Y}(t)) + w(t+1)$, where all four of the R.H.S. terms are unique. $u(t+1)$ and $w(t+1)$ are called the bivariate innovations of $X(t+1)$ and $Y(t+1)$, respectively; they are the errors associated with the optimal one-step-ahead predictors for the process. Letting t vary through the integers, the corresponding errors $\begin{pmatrix} u \\ w \end{pmatrix} (t)$ form a new s.p., the innovations process (i.p.), corresponding to the original $\begin{pmatrix} X \\ Y \end{pmatrix} (t)$ process: stationarity in the latter can be shown to induce stationarity in the former with the aid of a family of unitary operators on $H_{X,Y}$ familiar to economists as L^t , where L is the backward lag operator.⁷ More evident is the uncorrelatedness of $\begin{pmatrix} u \\ w \end{pmatrix} (t)$ over time. Since $\begin{pmatrix} u \\ w \end{pmatrix} (t) \perp H_{X,Y}(t-1)$ ⁸ and $\begin{pmatrix} u \\ w \end{pmatrix} (t-k) \in H_{X,Y}(t-k) \subseteq H_{X,Y}(t-1)$, $\begin{pmatrix} u \\ w \end{pmatrix} (t) \perp \begin{pmatrix} u \\ w \end{pmatrix} (t-k)$ (by which is meant that the autocovariance matrix formed from the two vectors,

$$\Gamma_{u,w}^*(k) = \left(\begin{pmatrix} u \\ w \end{pmatrix}(t), \begin{pmatrix} u \\ w \end{pmatrix}(t-k) \right) = \begin{pmatrix} \langle u(t), u(t-k) \rangle & \langle u(t), w(t-k) \rangle \\ \langle w(t), u(t-k) \rangle & \langle w(t), w(t-k) \rangle \end{pmatrix},$$

vanishes for $k \neq 0$; this will happen precisely when all of the components of one vector are \perp to all of the components of the other). If, as is the case here, $\Gamma_{u,w}^*(k) = \delta_{0,k} \cdot \Sigma$ where $\delta_{0,k} \equiv \begin{cases} 1 & k=0 \\ 0 & k \neq 0 \end{cases}$ and $\Sigma = \Gamma_{uw}^*(0)$, the process $\begin{pmatrix} u \\ w \end{pmatrix}(t)$ will be said to be vector white noise (v.w.n.); this said, we will emphasize that being v.w.n. is a characteristic but not characterizing feature of the innovations process.

The rank of Σ , $\rho(\Sigma)$, is known as the rank⁹ of the $\begin{pmatrix} X \\ Y \end{pmatrix}$ process and indicates an important structural characteristic of the system. Some taxonomy regarding system rank follows:

(a) $\begin{pmatrix} X \\ Y \end{pmatrix}$ may be perfectly predicted from its past only if $\begin{pmatrix} u \\ w \end{pmatrix}(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, all t ; in this case Σ is the null matrix, $\rho(\Sigma) = 0$, and the process is said to be deterministic. (b) $\rho(\Sigma) \geq 1$, the process is nondeterministic (n.d.): it possesses at least one "component" which cannot be perfectly predicted from the past. The subcases are: (i) $\rho(\Sigma) = 1 < 2$, a degenerate case in which the bivariate shock $\begin{pmatrix} u \\ w \end{pmatrix}(t)$ is essentially univariate. We will not study this case here; however, the description suggests an alternative modeling for k -index models [22] in which a few aggregate shocks impinge on several sectors of the economy.¹⁰ (ii) $\rho(\Sigma) = 2$, the full rank case, is surely the object of most interest. From now on we deal exclusively with this case: Σ^{-1} exists, $|\Sigma| \neq 0$, two genuine (linearly independent) shocks perturb the system

each period.

This last development suggests the decomposition

$H_{X,Y}(t) \equiv H_{X,Y}(t-1) \oplus D_{X,Y}(t)$, where the space $D_{X,Y}(t)$ is the two-dimensional orthogonal complement of $H_{X,Y}(t-1)$ in $H_{X,Y}(t)$

(\oplus was defined in footnote 7). It is not hard to show that

$D_{X,Y}(t) = S(u(t), w(t))$. This construction is canonical:

$H_{u,w}(t) \equiv H_{u,w}(t-1) \oplus D_{u,w}(t)$. The v.w.n. property of

$\binom{u}{w}(t)$ immediately gives $H_{u,w}(t) \equiv \overline{S(u(s), w(s), s \leq t)} = \sum_{s=-\infty}^{s=t}$

$\oplus S(u(s), w(s)) = \sum_{s=-\infty}^{s=t} \oplus D_{X,Y}(s)$,¹¹ since $D_{u,w}(t) = D_{X,Y}(t)$,

$t \in I$. At the other extreme, the space $\bigcap_{t \in I} H_{X,Y}(t) \equiv H_{X,Y}(-\infty)$

is called the remote past of $\binom{X}{Y}$; we could forecast a variable

in it, $Z(t+1)$ say, perfectly from $H_{X,Y}(t)$, and just as well from

$H_{X,Y}(t-k)$ for any $k \in I$. Stationarity guarantees, of course,

that perfect forecasts are available arbitrarily far into the future for such random variables.

By combining these subspaces, an important orthogonal decomposition of the present and past of $\binom{X}{Y}_t$ is obtained:

$$H_{X,Y}(t) = H_{X,Y}(-\infty) \oplus H_{u,w}(t) = H_{X,Y}(-\infty) \oplus \sum_{s=-\infty}^t \oplus D_{X,Y}(s).^{12}$$

The ground work has now been laid for the most important

result in the time domain analysis of wide sense stationary

stochastic processes.

Wold Decomposition Theorem. For the w.s.s.p. $\binom{X}{Y}(t)$ with

innovations process $\binom{u}{w}(t)$, and where the associated spaces

are as defined above,

$$(i) \quad \begin{pmatrix} X \\ Y \end{pmatrix} (t) = \begin{pmatrix} X \\ Y \end{pmatrix}_{l.r.} (t) + \begin{pmatrix} X \\ Y \end{pmatrix}_{d.} (t), \text{ where}$$

$$\begin{pmatrix} X \\ Y \end{pmatrix}_{l.r.} (t) \equiv \left(\begin{pmatrix} X \\ Y \end{pmatrix} (t) \mid H_{u,w}(t) \right) \perp \begin{pmatrix} X \\ Y \end{pmatrix}_{d.} (t) \equiv \left(\begin{pmatrix} X \\ Y \end{pmatrix} (t) \mid H_{X,Y}(-\infty) \right);$$

(ii) $\begin{pmatrix} X \\ Y \end{pmatrix}_{l.r.} (t)$ has the (one-sided) moving average representation

$$\sum_{k=0}^{\infty} A(k) \cdot \begin{pmatrix} u \\ w \end{pmatrix} (t-k) \equiv A * \begin{pmatrix} u \\ w \end{pmatrix} (t) \equiv \sum_{k=0}^{\infty} \left(\begin{pmatrix} X \\ Y \end{pmatrix} (0), \begin{pmatrix} u \\ w \end{pmatrix} (-k) \right) \Sigma^{-1} \begin{pmatrix} u \\ w \end{pmatrix} (t-k),$$

and

$$\| |X_{l.r.}(t)| \|^2 + \| |Y_{l.r.}(t)| \|^2 = \text{trace} \sum_{k=0}^{\infty} A(k) \Sigma A'(-k)$$

$$\equiv \text{tr} A \Sigma * A'(0) < \infty;$$

(iii) $\begin{pmatrix} X \\ Y \end{pmatrix}_{d.} (t)$ is deterministic, and, for all $t \in I$,

$$\overline{S \left(\begin{pmatrix} X \\ Y \end{pmatrix}_{d.} (j), -\infty < j \leq t \right)} = H_{X,Y}(-\infty)$$

The mnemonics l.r. and d. stand for linearly regular and deterministic, respectively. The latter term has already been defined; concerning the former, a s.p. $\begin{pmatrix} X \\ Y \end{pmatrix} (t)$ is said to be linearly regular (purely nondeterministic is also used) if $\left(\begin{pmatrix} X \\ Y \end{pmatrix} (t) \mid H_{X,Y}(s) \right) \rightarrow 0$ as $s \rightarrow -\infty$; intuitively, if the effect of the past diminishes as the "past becomes more remote," or equivalently by stationarity, if the distant future can be predicted no better than by solely using the process mean

(here, zero). Equivalent characterizations of linear regularity¹³ are each of (a) the ability to express the entire process as a m.a. involving its innovations; and (b) $H_{X,Y}^{(-\infty)} = \{0\}$.

We may now paraphrase the Wold theorem to say that an arbitrary w.s.s.p. may be decomposed into two parts, uncorrelated with each other, of which one is purely deterministic and the other purely nondeterministic. Since the purely deterministic part may be perfectly predicted arbitrarily far into the future (with no effect on the linearly regular part because of the orthogonality), we can without loss of generality subtract it from $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ and assume that the process we are analyzing is linearly regular.¹⁴ This assumption will often be employed throughout the remainder of the paper: $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ is a l.r. w.s.s.p.

Consequently, the process we study is characterized by (ii), which requires further discussion for reasons other than the notation implicitly introduced.

The matrices $A(k)$, sometimes referred to as A_k , are unique, from the second identity in (ii). The convolution definition is given generically by the first identity, the interpretation of this infinite sum, of course, being convergence in quadratic mean of each of the random variable-partial-sum components. The condition that the indicated trace be finite is necessary and sufficient for this convergence; it is succinctly expressed in terms of the (now, nonrandom) matrix

convolution, and amounts simply to the requirement of finite variance for $X(t)$ and $Y(t)$ because the autocovariance sequence $\Gamma_{X,Y}(\cdot)$ of $\begin{pmatrix} X \\ Y \end{pmatrix} = A * \begin{pmatrix} u \\ v \end{pmatrix}$ may easily be expressed in terms of the autocovariance sequence $\Gamma_{u,v}(\cdot)$ as $A * \Gamma_{u,v} * A'(k)$,

where $A'(k) \equiv A^T(-k)$, T indicates ordinary matrix transpose and $*$ is the appropriate notion of convolution transpose, and $A * B(m) \equiv \sum_{j=-\infty}^{\infty} A(m-j)B(j) \equiv \sum_{j=-\infty}^{\infty} A(j)B(m-j)$. When $A(\cdot)$ and $B(\cdot)$ are one-sided, i.e., $A(s) = B(s) = 0, s < 0$, then these sums are both finite and the lower limit may be replaced by 0.

For the case of v.w.n. the double summation implied by the double convolution reduces to $A\Sigma * A'(k)$ ¹⁵ or $A * \Sigma A'(k)$;

these last formulae suggest the desirability of a representation in which $\Sigma = I$ so that $R_{X,Y} = A * A'(k)$. This may be done by tucking $\Sigma^{1/2} * \Sigma^{-1/2}$ into the convolution, to arrive at

$$\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A\Sigma^{1/2} * \Sigma^{-1/2} \begin{pmatrix} u \\ v \end{pmatrix}(t) \equiv B * \begin{pmatrix} e \\ n \end{pmatrix}(t), \text{ say (see pp. 18-19}$$

for an elaboration of this procedure). In the new representation $R_{X,Y}(k) = B * B'(k)$, since $\text{Cov } \Sigma^{-1/2} \begin{pmatrix} u \\ v \end{pmatrix}(t) \equiv E \begin{pmatrix} e \\ n \end{pmatrix}(t) \begin{pmatrix} e \\ n \end{pmatrix}^T(t) = I$; the requirement of finite variances of $X(t)$ and $Y(t)$

becomes $\text{tr } B * B'(0) < \infty$, which will occur precisely when $\sum_{i=0}^{\infty} b_{11}^2(i) + \sum_{i=0}^{\infty} b_{12}^2(i) < \infty$ and $\sum_{i=0}^{\infty} b_{21}^2(i) + \sum_{i=0}^{\infty} b_{22}^2(i) < \infty$, where

$$B(\cdot) = \begin{pmatrix} b_{11}(\cdot) & b_{12}(\cdot) \\ b_{21}(\cdot) & b_{22}(\cdot) \end{pmatrix}, X(t) = b_{11} * e(t) + b_{12} * n(t) \text{ and } Y(t)$$

$$= b_{21} * e(t) + b_{22} * n(t).$$

The last remarks show that, if we form $B(z)$ with typical element $(j,k=1,2) b_{jk}(z)$, $|z| < 1$ where z is now a complex number, then $b_{jk}(z) \equiv \sum_{s=0}^{\infty} b_{jk}(s)z^s$ converges pointwise in the unit circle, and so defines an analytic function there. On the unit circle, square summability of the sequence and classical methods yield the representation $b_{jk}(e^{i\lambda}) = \sum_{s=0}^{\infty} b_{jk}(s)e^{i\lambda s}$, where the convergence is not pointwise but in $L_2[0, 2\pi]$. The latter function can be shown to be a radial limit of the former; analogous results hold on $|z| > 1$ for $b_{jk}(z^{-1})$. The close connection between these representations is the study of functions of Hardy class H_2 : those square integrable functions with Fourier series involving only positive powers of $z = e^{i\lambda}$.

These considerations suggest use, at least for placeholder purposes, of the method of "z-transforms," a principal result of which is: $R_{X,Y}(z) \equiv \sum_{k=-\infty}^{\infty} ((\frac{X}{Y})(t), (\frac{X}{Y})(t-k))z^k = B(z) B^T(z^{-1})$, where the equality means "equality of the coefficients of z^k in the formal expansion of." In other words, the coefficients of the convolution $B * B'(s)$ may be ascertained by multiplication in $B(z) B^T(z^{-1})$ and checking the coefficient of z^s ; nothing more is involved here than the familiar notion that "convolution in the time domain corresponds to multiplication in the frequency domain." More significantly, however, the theoretical importance of these analytic $B(z)$ matrices has been hinted at ([17], pp. 58-63).

The discussion of the last several pages has given a sketch of an existence proof of a very important way of looking at the process under study. The guaranteed representation has not been constructed, however; the problem in practice is, given $R_{X,Y}(s)$, how to "factor" it into the $B * B'(s)$ of the past paragraph? There are several layers of difficulties involved: (1) When a $B(\cdot)$ is found which performs the factorization, there is the further requirement that, in $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = B * \begin{pmatrix} e \\ n \end{pmatrix}(t)$, the $\begin{pmatrix} e \\ n \end{pmatrix}$ process must "be in the right space," by which is meant, $H_{X,Y}(t) = H_{e,n}(t)$, all t . (This latter notion will be abbreviated (m.s.) and taken up in the sequel.) In other words, not just any v.w.n. process will do; and not only must the $\begin{pmatrix} X \\ Y \end{pmatrix}$ and $\begin{pmatrix} e \\ n \end{pmatrix}$ processes be defined on the same probability space, they must each essentially be linear combinations of each other's past and present, or in another (perhaps more economic) context, they must carry the same information. The interplay between analytic properties of $B(z)$ and the associated stochastic properties (of the corresponding $\begin{pmatrix} e \\ n \end{pmatrix}(t)$) is treated in [17], Ch. 2. These remarks will be expanded momentarily. The second difficulty is:

(2) There is an identification problem which, when (1) is understood, is naturally solved by restricting attention to those $B(\cdot)$ associated with "errors in the right space" and imposing a normalization rule to distinguish between the observationally equivalent structures within this appropriate class. (3) Finally, when a theoretical understanding of the

first two points is in hand, and even in an ideal case where the observable data, $R_{X,Y}(s)$, are generated by elements which are ratios of polynomials, a procedure for obtaining the desired factorization is not trivial. A method which terminates in a finite number of steps is presented in [17], p. 44-47; since many square roots and polar decompositions may be needed, it is not an easy exercise to generate examples with pencil and paper. Actually, what this algorithm generates is a $B(z)$ matrix such that: (i) all of its elements are rational and analytic in $|z| < 1$; and (ii) $\det B(z)$ has all of its (finitely many) zeros in $|z| \geq 1$. Only after much more machinery is developed (p. 88) is Rozanov able to show that this $B(z)$ has an associated errors process which is in the right space, thereby correctly stating and proving for the first time a result which had often been assumed true, in various forms, and even to the present is often not adequately appreciated.¹⁶

A concept closely allied to m.a.r. is that of autoregressive representation (a.r.)¹⁷ When the l.r. w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ with associated i.p. $\begin{pmatrix} e \\ n \end{pmatrix}(t)$ permits the representation $B * \begin{pmatrix} X \\ Y \end{pmatrix}(t) \equiv \sum_{s=0}^{\infty} B(s) \begin{pmatrix} X \\ Y \end{pmatrix}(t-s) = \begin{pmatrix} e \\ n \end{pmatrix}(t)$, with $B(0) = I$, where the

partial sums converge in quadratic mean, then that representation is known as the a.r. We stress that it is important that the "errors" be the innovations, in which case the force of the a.r. is that $(\begin{pmatrix} X \\ Y \end{pmatrix}(t) | H_{X,Y}(t-1))$ has the convenient representation

- $\sum_{s=1}^{\infty} B(s) \begin{pmatrix} X \\ Y \end{pmatrix} (t-s)$ rather than the more generally necessary

representation as the limit of a sequence of finite sums, with possibly changing weights.

The Wold decomposition has given us a coordinate free representation in terms of the physically meaningful innovation vectors for the l.r.w.s.s.p. under study: $\begin{pmatrix} X \\ Y \end{pmatrix} (t) = \sum_{k=0}^{\infty} A_k \begin{pmatrix} U \\ W \end{pmatrix} (t-k) \equiv A * \begin{pmatrix} U \\ W \end{pmatrix} (t)$, $(\begin{pmatrix} U \\ W \end{pmatrix} (t), \begin{pmatrix} U \\ W \end{pmatrix} (t-k)) \Sigma^{-1} = A_k$ and $\Sigma = (\begin{pmatrix} U \\ W \end{pmatrix} (t), \begin{pmatrix} U \\ W \end{pmatrix} (t))$ (so that $A_0 = I$). The parameters A_k and Σ are of course unique, since the formulae give them in terms of observables or well-defined operations (projections) involving observables. We will refer to this as parameterization (I) or the natural parameterization (n.p.); it is especially convenient when the autoregressive representation exists, since its coefficients will be those in the power series expansion of $A^{-1}(z)$, where $A(z) \equiv \sum_{k=0}^{\infty} A(k) z^k$. We emphasize that the representation itself implies $H_{X,Y}(t) \subseteq H_{u,w}(t)$, and the construction of $\begin{pmatrix} U \\ W \end{pmatrix} (t)$ implies its subordination to $\begin{pmatrix} X \\ Y \end{pmatrix} (t)$; thus $\begin{pmatrix} U \\ W \end{pmatrix} (t)$ are mutually subordinate (m.s.): $H_{X,Y}(t) \equiv H_{u,v}(t)$. It is this last fact which is crucial for prediction theory generally and our decompositions in particular: $H_{u,v}(t)$ must represent a reasonably convenient description of current and past $\begin{pmatrix} X \\ Y \end{pmatrix}$. Other convolution representations, say $\begin{pmatrix} X \\ Y \end{pmatrix} = \underline{A} * \underline{\begin{pmatrix} U \\ W \end{pmatrix}}$ which, like $\begin{pmatrix} U \\ W \end{pmatrix}$, are vector white noise, also exist; as will be clearer from the ensuing paragraphs, there will be no difficulty in normalizing these

$\begin{pmatrix} u \\ w \end{pmatrix}$ so that their contemporaneous covariance matrix is the identity. Yet only for $\begin{pmatrix} u \\ w \end{pmatrix} = T \begin{pmatrix} u \\ w \end{pmatrix}$, $|T| \neq 0$ will $H_{\begin{pmatrix} u \\ w \end{pmatrix}}(t) \subseteq H_{X,Y}(t)$; those $\begin{pmatrix} u \\ w \end{pmatrix}$ which are not subordinate to $\begin{pmatrix} X \\ Y \end{pmatrix}$ yield lower prediction variances for $\begin{pmatrix} X \\ Y \end{pmatrix}$. They are not suitable for prediction purposes because they carry more information about the future of the $\begin{pmatrix} X \\ Y \end{pmatrix}$ process than is available from the present and past.

For some purposes we may wish a representation $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = D * \begin{pmatrix} e \\ n \end{pmatrix}(t)$, where the v.w.n. $\begin{pmatrix} e \\ n \end{pmatrix}(t)$ is mutually subordinate to $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ and $\begin{pmatrix} e \\ n \end{pmatrix}(t), \begin{pmatrix} e \\ n \end{pmatrix}(t) = I$. Rozanov terms $\begin{pmatrix} e \\ n \end{pmatrix}$ a fundamental process (f.p.) ([17], p. 56), and makes it a part of his definition of the moving average representation of a process. Obtaining a f.p. from the i.p. amounts to choosing an orthonormal basis in $D_{u,w}(t)$; since there are as many of these as there are orthonormal matrices, a f.p. retains this nonuniqueness, which may either be accepted or eliminated by imposing additional restrictions.

Starting with the n.p. $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A * \begin{pmatrix} u \\ w \end{pmatrix}(t)$, $\text{Cov} \begin{pmatrix} u \\ w \end{pmatrix} = \Sigma$, we arrive at a f.p. by: diagonalizing Σ , $P\Sigma P' = \Lambda$; writing $\Sigma = P'\Lambda P = (P'\Lambda^{1/2}P)(P'\Lambda^{1/2}P) \equiv \Sigma^{1/2} \Sigma^{1/2}$, $\Sigma^{-1/2} \equiv P'\Lambda^{-1/2}P$, $\Lambda \equiv \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ and $\Lambda^{1/2} \equiv \begin{pmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{pmatrix}$; and finally

taking $\begin{pmatrix} e \\ n \end{pmatrix}(t) \equiv \Sigma^{-1/2} \begin{pmatrix} u \\ w \end{pmatrix}(t)$. $\begin{pmatrix} e \\ n \end{pmatrix}(t)$ has covariance matrix the identity, and $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A\Sigma^{1/2} * \Sigma^{-1/2} \begin{pmatrix} u \\ w \end{pmatrix}(t) \equiv A\Sigma^{1/2} * \begin{pmatrix} e \\ n \end{pmatrix}(t)$ is a m.a.r. with $\begin{pmatrix} e \\ n \end{pmatrix}(t)$ as f.p. Any $Q, Q'Q = I$ results in

$\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A \Sigma^{1/2} Q' * Q \begin{pmatrix} e \\ n \end{pmatrix}(t) \equiv A \Sigma^{1/2} Q' * \begin{pmatrix} \bar{e} \\ \bar{n} \end{pmatrix}(t)$, so that

$\begin{pmatrix} \bar{e} \\ \bar{n} \end{pmatrix}(t) \equiv Q \begin{pmatrix} e \\ n \end{pmatrix}(t)$ is also fundamental. Since $A(0) = I$, the

nonuniqueness of the f.p. may be expressed by noting that the zero-order coefficient in the last convolution is $\Sigma^{1/2} Q'$

where $\Sigma^{1/2}$ is unique, but Q' may be any orthonormal matrix.

It is a well-known result in matrix theory¹⁸ ([16], p. 191-192)

that any symmetric matrix ($\Sigma^{-1/2}$) may be lower (respectively,

upper) triangularized by postmultiplication by an orthonormal

matrix; the triangularization is unique if the diagonal elements

are required to be positive. We call these normalizations

II-L and II-U, and will actually produce them below in a way

which does not lose track, as this argument has, of the

innovations.

Let us return to the n.p.: $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A * \begin{pmatrix} u \\ w \end{pmatrix}(t)$,

$$\text{Cov} \begin{pmatrix} u \\ w \end{pmatrix}(t) = \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{wu} & \sigma_w^2 \end{pmatrix} \text{ and } A(0) = I. \text{ We retain } u(t), \text{ but}$$

replace $w(t)$ in our basis by $v(t)$, where $v(t) \equiv w(t)$

$$- (w(t)|u(t)) = w(t) - \frac{\sigma_{uw}}{\sigma_u} u(t). \text{ In other words,}$$

$$\begin{pmatrix} u \\ v \end{pmatrix}(t) = \begin{pmatrix} 1 & 0 \\ \frac{-\sigma_{uw}}{\sigma_u} & 1 \end{pmatrix} \begin{pmatrix} u \\ w \end{pmatrix}(t), \text{ or } \begin{pmatrix} u \\ w \end{pmatrix}(t) = \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u} & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}(t).$$

Consequently, $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u} & 1 \end{pmatrix} * \begin{pmatrix} u \\ v \end{pmatrix}(t)$ in which the m.s.

process $\begin{pmatrix} u \\ v \end{pmatrix}(t)$ is serially and contemporaneously uncorrelated,

with the first element the X innovation and the second element that part of the Y innovation \perp to the X innovation. Not only

are the convolution matrix coefficients $A(k) \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{2} & 1 \\ \sigma_u & \end{pmatrix}$, $k \in I$,

well defined in terms of this physically specified basis, they are econometrically identified under the following (classical) pair of "zero restrictions":¹⁹ (i) That the zero order convolution coefficient be lower triangular with ones on the diagonal, and (ii) that the covariance matrix be diagonal. To see this, consider the zero-order coefficient and covariance

matrix of any other m.s. representation: $A(0) \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{2} & 1 \\ \sigma_u & \end{pmatrix} T^{-1}$,
 $T \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} T'$. Only lower triangular T with units on the

diagonal will satisfy (i) (recall that $A(0)=I$); if $t_{21} \neq 0$, then the off-diagonal terms ($t_{21}\sigma_u^2$) fail to vanish in the transformed covariance matrix, causing (ii) to fail. Of course, had the Y innovation been retained in the basis (as the initial step of a Gram-Schmidt orthogonalization) the result would have been an upper triangular zero-order matrix with diagonal covariance matrix. We call these normalization conventions III-L,D and III-U,D, respectively (the mnemonics L,U are for lower, upper triangular and D is for diagonal covariance matrix).

We may now deduce the exactly identifying nature of the normalizations II-U and II-L from III-L,D and III-U,D (which have just been shown exactly identified, via I). This procedure involves: algebraically, normalizing the diagonals of the covariance matrix at unity and making the corresponding adjustment to the zero-order coefficient matrix; or, geometrically, scaling the orthogonal innovations in $\begin{pmatrix} u \\ v \end{pmatrix}(t)$ so that they have unit variance. In other words, $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$

$$= A \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{2} & 1 \\ \sigma_u & \end{pmatrix} * \begin{pmatrix} u \\ v \end{pmatrix}(t) = A \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{2} & 1 \\ \sigma_u & \end{pmatrix} \begin{pmatrix} \sigma_u & 0 \\ 0 & \sigma_v \end{pmatrix} * \begin{pmatrix} \frac{1}{\sigma_u} & 0 \\ 0 & \frac{1}{\sigma_v} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}(t);$$

or again $A \begin{pmatrix} \sigma_u & 0 \\ \sigma_{uw} & \sigma_v \\ \sigma_u & \end{pmatrix} * \begin{pmatrix} u \\ \sigma_u \\ v \\ \sigma_v \end{pmatrix}$, which is in the form II-L.²⁰

The point is that, starting from III-L, only lower-triangular T will preserve the lower triangularity required by II-L; the diagonal elements must be as above to produce unit diagonal in the covariance matrix, and, if $t_{21} \neq 0$, as before, the diagonality of the resulting covariance matrix would be spoiled. Consequently, we have again arrived at the II-normalizations, but in a constructive way which has not lost sight of the innovations.

Of the three normalization variants, only the II^S are fundamental; how does this square with the rightful emphasis

given fundamental representations? Since the important part of fundamentalness is mutual subordination, and all the normalizations were among m.s. processes, it is not imperative to adopt the further, underidentifying normalization convention that fundamentalness carries with it. We have found it helpful in organizing thought to adopt some normalization and stick with it, interpreting, if necessary, the results for other normalizations at the last stage; the alternative, stating results valid for some unspecified or all possible normalizations, is likely to leave the reader (if not the author) frustrated and confused.

Finally, we come to define the notion of causality for l.r. w.s.s.p. $(\begin{smallmatrix} X \\ Y \end{smallmatrix})(t)$. Y is said to cause X if, given past X, past Y aids in the prediction of current X (notation: $Y \rightarrow X$). In symbols, Y will cause X when $(X(t) | H_{X,Y}(t-1)) \neq (X(t) | H_X(t-1))$. Y is said to cause X instantaneously if adding current Y helps predict X, given past X and past Y (notation: $Y \rightarrow X$). In symbols, $Y \rightarrow X$ whenever $(X(t) | \overline{H_{X,Y}(t-1) \cup Y(t)}) \neq (X(t) | H_{X,Y}(t-1))$. Since part of Proposition 1 shows that $Y \rightarrow X$ if and only if $X \rightarrow Y$, the above notion of (Wiener-Granger) causality does not permit any distinction as regards instantaneous causality; consequently, the definition is only meaningful over time. Whether these are the more interesting causality events depends on one's philosophical bent. On the necessity of a stochastic notion of causality, see [4], p. 430; for a comparison of this definition with other notions of

causality, see the first section of [26].

If we now adopt the normalization II-L (so that $\begin{pmatrix} u \\ w \end{pmatrix}(t)$ is fundamental--v.w.n. with covariance matrix the identity and a linear combination of the innovations--and $b(0) = 0$) then we may state:

Sims' Theorem 1. In the l.r. w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$, Y does not cause X if and only if the Wold (m.a.) representation subject to II-L is $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} * \begin{pmatrix} u \\ w \end{pmatrix}(t)$.

The method of proof employed in [24] is direct, uses the characterizing features of the m.a.r., and cannot be improved upon. A careful reader of the proof might wonder why the u process, which by construction is part of the bivariate innovation, is also the univariate innovation for the X process. Since we use both this theorem and this fact, and the reason illustrates the theoretical importance of viewing $B(z) \equiv \begin{pmatrix} a(z) & 0 \\ c(z) & d(z) \end{pmatrix}$ as an analytic function, we will supply an explanation. The crucial notion is that among matrices $\bar{B}(z)$ which factorize the autocovariance function, or equivalently the spectral density matrix, that matrix which corresponds to the desired fundamental representation, $B(z)$, has the maximality property $B(0)B^T(0) \geq \bar{B}(0)\bar{B}^T(0)$ where \geq here means "LHS minus RHS is positive semidefinite." ([17], p. 60, 61). This maximality notion applies both to univariate and bivariate factorizations, so if u weren't fundamental for X, there would be $\tilde{a}(z) = g(z) a(z)$, with $g(z) g(z^{-1}) = 1$ on

$|z| = 1$, such that $|\underline{a}(0)|^2 > |a(0)|^2$ (here $a(z)$ is a scalar).

Then if we consider a competing $\underline{B}(z) \equiv \begin{pmatrix} \underline{a}(z) & 0 \\ c(z) & g(z) & d(z) \end{pmatrix}$,

$\underline{B}(z)$ still factors the spectral density and $\det|\underline{B}(0)|^2$
 $= |\underline{a}(0)|^2 |d(0)|^2 > |a(0)|^2 |d(0)|^2$, contradicting the

maximality of $B(z)$ and hence the joint fundamentalness of

$\begin{pmatrix} u \\ w \end{pmatrix}$.

This concludes our background survey and introduction of notation. It is time to begin work.

III. Bivariate Characterizations

We begin this section with a lemma which will find immediate use. Paraphrased, it says if $Y(t)$ is not perfectly predictable from the past of X and Y , it must remain so when current and a finite number of future X are included among the predictors, or else $X(t)$ will be perfectly predictable.

Lemma. Assume the l.r.w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}$ has an autoregressive representation with innovations process $\begin{pmatrix} u \\ w \end{pmatrix}$, and $w(t) \neq 0$. Then for any $n \geq 0$, $Y(t-n)$ in $\overline{H_X(t-1) \cup H_Y(t-n-1)} \equiv A(n)$ implies $u(t-1) = 0$, in which case $|\Sigma| \equiv |\text{Cov} \begin{pmatrix} u \\ w \end{pmatrix}(0)| = 0$.

Proof: The case $n = 0$ is self-evident, so we may choose $n \geq 1$. Since $Y(t-n) \in A(n)$ is equivalent to the existence of $c \equiv (c_1, \dots, c_n)$ such that (*) $Y(t-n) = c_1 X(t-n) + \dots + c_n X(t-1) + z$, $z \in H_{X,Y}(t-n-1)$, $c \neq 0$ or else $w(t) = 0$. Relabeling if necessary, we may take $c_n \neq 0$. Substituting for $X(t-1)$ from the a.r. into (*), dividing by c_n , and transposing all the terms on the RHS except $u(t-1)$ shows

$u(t-1)$ to be in $H_{X,Y}(t-2)$. But as an innovation,

$u(t-1) \perp H_{X,Y}(t-2)$; hence $u(t-1) = 0$. Q.E.D.

The time indices were chosen to accord with their use below; setting $t = s+n$ yields a form consistent with the paraphrase above.

We waste no time in putting the machinery developed in the last section and above to work in the proof of

Proposition 1. Assume that the l.r.w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}$ has the autoregressive representation (a.r.) $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} * \begin{pmatrix} X \\ Y \end{pmatrix}(t)$

$$+ \begin{pmatrix} u \\ w \end{pmatrix}(t) \equiv \begin{pmatrix} \sum_{i=1}^{\infty} a(i)X(t-1) + \sum_{i=1}^{\infty} b(i)Y(t-i) \\ \sum_{i=1}^{\infty} c(i)X(t-i) + \sum_{i=1}^{\infty} d(i)Y(t-i) \end{pmatrix}$$

$$+ \begin{pmatrix} u \\ w \end{pmatrix}(t) \text{ with } E_{\begin{pmatrix} u \\ w \end{pmatrix}(t)}(u(t)w(t))$$

$$\equiv \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{wu} & \sigma_w^2 \end{pmatrix} \equiv \Sigma.$$

Then: (i) Y does not cause X if and only if $b(\cdot) \equiv 0$;

(ii) whether or not $b(\cdot) \equiv 0$, instantaneous feedback (or instantaneous causality) is present if and only if $\sigma_{uw} \neq 0$, this last result holding even if no a.r. exists, where then $\begin{pmatrix} u \\ w \end{pmatrix}(t)$ remains the innovations process.

Proof: (i) Assume first that $b(\cdot) \equiv 0$. Then by the definition of the a.r. $\begin{pmatrix} u \\ w \end{pmatrix}(t) \perp H_{(X,Y)}(t-1)$, so that $(X(t) | H_{X,Y}(t-1)) = a * X(t) + 0 * Y(t)$. In general, we may form

$(X(t) | H_X(t-1))$ by projecting $(X(t) | H_{X,Y}(t-1))$ onto $H_X(t-1)$, a step which is not necessary here. Rather trivially the two projections are equal, and sufficiency is established.

Now assume $(X(t) | H_{X,Y}(t-1)) = a*X(t) + b*Y(t)$, so that

$(X(t) | H_X(t-1)) = a*X(t) + (b*Y(t) | H_X(t-1))$; by hypothesis these projections are equal. This entails $\sum_{i=1}^{\infty} b(i)\{Y(t-i)$

$- (Y(t-i) | H_X(t-1))\} = 0$. If $b(\cdot) \not\equiv 0$, we will have a contradiction. Let n be the first index i such that $b(i) \neq 0$.

Dividing by $b(n)$ gives $Y(t-n) = - \sum_{i=1}^{\infty} \frac{b(i)}{b(n)} Y(t-n-i)$
 $+ \sum_{i=1}^{\infty} \frac{b(i)}{b(n)} (Y(t-i) | H_X(t-1))$. The first term is in $H_{X,Y}(t-n-1)$

from the a.r. for $X(t)$, and the second term is in $H_X(t-1)$.

Application of the lemma above entails $u(t-1) = 0$, which would contradict the full rank assumption of $|\Sigma| \neq 0$. (ii) To

check for instantaneous feedback-causality, we compare

$(X(t) | \overline{H_X(t-1) \cup H_Y(t)})$ with $(X(t) | H_{X,Y}(t-1))$ and

$(Y(t) | \overline{H_X(t) \cup H_Y(t-1)})$ with $(Y(t) | H_{X,Y}(t-1))$. To compute the

former, we first regress $u(t) = \frac{\langle u(t), w(t) \rangle}{\langle w(t), w(t) \rangle} w(t) + v(t)$ so

that $\langle u(t) - \frac{\sigma_{uw}}{\sigma_w} w(t), w(t) \rangle = 0$, or $w(t) \perp v(t)$.

$(X(t) | \overline{H_X(t-1) \cup H_Y(t)}) = (X(t) | H_{u,w}(t-1) \oplus w(t))$

$= (X(t) | H_{X,Y}(t-1)) + (X(t) | w(t))$. Thus, the marginal effect

of current Y is to change our forecast of $X(t)$ by $(X(t) | w(t))$

$= (u(t) | w(t)) = \frac{\sigma_{uw}}{\sigma_w} w(t) + (v(t) | w(t)) = \frac{\sigma_{uw}}{\sigma_w} w(t)$; the

predictive variance is correspondingly lowered by $\frac{\sigma_{uw}^2}{\sigma_w^2}$. These

effects are zero if and only if $\sigma_{uw} = 0$, as asserted. The computation for predicting $Y(t)$ shows that the predictors

differ by $\frac{\sigma_{uw}}{\sigma_u} u(t)$ and the variances differ by $\frac{\sigma_{uw}^2}{\sigma_u^2}$, so that

current X is of additional help in predicting current Y if and only if $\sigma_{uw} \neq 0$. The results for X and Y together establish that instantaneous effects are present either together or not at all, justifying the term instantaneous feedback. Finally, no use was made of the a.r. representation in proving (ii), as was promised. Q.E.D.

A special case occurs when the order of the longest lag necessary in the a.r. is finite, m , say, and $\sigma_{uw} = 0$. Granger called this the "simple causal model" and proved (i) of Proposition 1 ([4], p. 436), establishing the first theorem in the subject with no fanfare (the result modestly bears no label whatever). We wish both to emphasize the importance of his result and to make the following observations.

Remark 1. Although he asserts his result for $m=\infty$ as well, it appears that Granger's method of proof, which involves examination of Kolmogorov's expression for the predictive error variance, will not be easily adapted to this case, because some statements which are "clearly" true in his proof are not so clear when infinite products are involved (although

perhaps the introduction of Blaschke products along the lines of [7], p. 142-3, may be used to advantage.)

Remark 2. The fundamentalness of the $\binom{u}{w}$ process and the defining properties of the a.r. were not stressed by Granger, although uniqueness of the a.r. certainly must be present for his result to hold. It is useful to be clear that the result is not specific to the simple causal model, as is also evidenced in:

Remark 3. By "inverting" the m.a. and taking $b(\cdot) \equiv 0$ in precisely those cases where Y does not cause X, the a.r. is found, when it exists, to have all coefficients of lagged Y equal to zero. Thus, application of Sims' Theorem 1 yields another proof of the Granger result. Of course, as our normalization discussion has shown, the a.r. so obtained is not necessarily associated with a simple causal model (even when the a.r. is finite). A justification of this "inverting" procedure may be found in the second of the Wiener-Masani references [30].

With the characterization of instantaneous causality in hand, let us return to the Sims theorem and observe that instantaneous causality obtains if and only if $c(0) \neq 0$. This is apparent, since the identification-normalization discussion shows $c(0) = \frac{\sigma_{uw}}{\sigma_u}$; in other words, the presence of i.c. is thrust entirely into $c(\cdot)$, and the force of $b(s) = 0$, all s consists not in $b(0) = 0$, which holds by

normalization and so is always possible, but in the ability to take $b(s) = 0$, $s=1, 2, \dots$.

In a comparison of these theorems, the Sims result has the advantage of mathematical generality, in that its hypotheses are met for any l.r.w.s.s.p.; the Granger result, while requiring in addition an a.r., yields an immediate statistical test. We refer the reader to [31] for a discussion of the estimation of multivariate autoregression; of course, ordinary least squares, comparing X on lagged X with X on lagged X and lagged Y may be used as well.

Our next result presents another characterization of the exogeneity of X , which, like the earlier result in terms of the Wold representation, has the advantage of requiring no additional assumptions. Indeed, the remark below shows that the l.r.w.s.s.p. assumption may even be relaxed.

Proposition 2. In the l.r.w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$, Y does not cause X if and only if $(Y(t) | H_X(t)) = (Y(t) | H_X(-\infty, \infty))$.

Proof: We prove sufficiency first, assuming equality of the two projections. By the characterizing property of $(Y(t) | H_X(-\infty, \infty))$, $Y(t) - (Y(t) | H_X(-\infty, \infty)) \perp X(t+s)$, all t and s , and particularly for $s > 0$. Hence, by the assumed projection equality for all t , we substitute, shift, and define to arrive at $n_j \equiv Y(t-j) - (Y(t-j) | H_X(t-1)) \perp X(t+k)$, $j=1, 2, \dots$ and all $k \in I$. If $N_1 \equiv \bigcup_{j=1}^{\infty} \{n_j\}$, we have $N_1 \perp H_X$, and à fortiori $N_1 \perp H_X(t)$ and $N_1 \perp H_X(t-1)$, since by construction each vector

in N_1 has these properties. Taking a closure and using the continuity of the inner product yields $\bar{N}_1 \perp H_X(t)$ and $H_X(t-1)$ as well. This gives us license to write

$$(X(t) | H_X(t-1) \cup \bar{N}_1) = (X(t) | H_X(t-1)) + (X(t) | \bar{N}_1),$$

since clearly $S(X(t-j), Y(t-j), j=1, 2, \dots) = S(\bigcup_{j=1}^{\infty} X(t-j), N_1)$

which by definition is $H_{X,Y}(t-1)$. Thus,

$$(*) \quad (X(t) | H_{X,Y}(t-1)) = (X(t) | H_X(t-1)) + (X(t) | \bar{N}_1).$$

But $(X(t) | \bar{N}_1) = 0$, since $X(t) \in H_X(t)$ and we have seen that $H_X(t) \perp \bar{N}_1$.²¹ In other words, Y does not cause X . Conversely, assume Y does not cause X . Since $(*)$ is always a valid decomposition, we have $(X(t) | \bar{N}_1) = 0$, or the countable statements $(j) X(t) \perp Y(t-j) - (Y(t-j) | H_X(t-1))$, $j=1, 2, \dots$. We will use only the fact that orthogonality relations are valid when shifted over time, a weak implication of covariance stationarity (cf. the Remark below), in establishing by induction that $(Y(t) | H_X(t+k)) = (Y(t) | H_X(t))$, $k=1, 2, \dots$. Setting $j=1$ and shifting forward one unit yields $X(t+1) \perp Y(t) - (Y(t) | H_X(t))$; consequently by definition $(Y(t) | H_X(t)) = (Y(t) | H_X(t+1))$, anchoring the induction. Now setting $j=n$ yields $X(t) \perp Y(t-n) - (Y(t-n) | H_X(t-1))$, so that $(Y(t-n) | H_X(t-1)) = (Y(t-n) | H_X(t))$. Shifting this last equality forward n units and employing the induction hypothesis, $(Y(t) | H_X(t)) = (Y(t) | H_X(t+n-1))$ yields $(Y(t) | H_X(t+n)) = (Y(t) | H_X(t))$. Consequently $(Y(t) | H_X(-\infty, \infty))$

$$= (Y(t) | \overline{\bigcup_{n=1}^{\infty} H_X(t+n)}) = (Y(t) | H_X(t)), \text{ using the continuity of}$$

the projection operator. Q.E.D.

Remark. An inspection of the proof reveals that the hypothesis $(Y(t) | H_X(-\infty, \infty)) = (Y(t) | H_X(t))$, all $t \in I$, would have sufficed, along with the hypothesis that the variances of $X(t)$ and $Y(t)$ exist for all t . This latter requirement provides the Hilbert space $H_{X,Y}(-\infty, \infty)$ with respect to which the necessary projections are defined. Thus, the linear regularity and stationarity of the covariance structure of the l.r.w.s.s.p. hypothesis may be dismissed. However, most applications will be either to stationary processes or very special departures from stationarity.

Corollary 1. If in the l.r.w.s.s.p. $(\begin{smallmatrix} X \\ Y \end{smallmatrix})$, $X(t)$ has an autoregressive representation, $f * X(t) = e(t)$, then Y does not cause X implies that $Y(t)$ can be expressed as a distributed lag on current and past X with a residual which is not correlated with $X(s)$, past ($s < t$), present ($s=t$), or future ($s > t$).

Proof: Define $w(t) = Y(t) - (Y(t) | H_X)$. Then $w(t) \perp X(s)$, all integer t and s , by construction. Using the proposition, $w(t) = Y(t) - (Y(t) | H_X(t))$. Since $(Y(t) | H_X(t)) \in H_X(t) = H_e(t)$ and $\{e(s)\}_{s=-\infty}^{s=t}$ is a complete orthonormal set, we have the Fourier representation

$$(Y(t) | H_X(t)) = \sum_{s=-\infty}^{s=t} \langle (Y(t) | H_X(t)), e(s) \rangle e(s) = q * e(t),$$

say. So $(Y(t) | H_X(t)) = q * f * X(t)$, and it follows that $Y(t) = q * f * X(t) + w(t)$. $q * f$ as the convolution of two, one-sided convolutions, is clearly one-sided, and $w(t)$ has the desired orthogonality property. Q.E.D.

Corollary 2. The converse of Corollary 1 holds, even if X has no a.r.

Proof: By assumption we have $Y(t) = h * X(t) + w(t)$, say, with $w(t) \perp X(s)$, all t and s . Consequently, $(Y(t) | H_X(t)) = h * X(t)$ and $(Y(t) | H_X(-\infty, \infty)) = h * X(t)$. Application of the proposition shows that Y does not cause X . Q.E.D.

The corollaries taken together provide a strengthening²² (and alternate proof) of Sims' Theorem 2. Thus, the Sims test for exogeneity--testing whether "future" coefficients of $h(\cdot)$ vanish--is an implication of "Y does not cause X" under the milder assumption that only X (and not $\begin{pmatrix} X \\ Y \end{pmatrix}$ jointly) possesses an a.r. On the other hand, the presence of a one-sided $f(\cdot)$, always referring to a population or theoretical regression, guarantees that "Y does not cause X" without any further qualifications.

Despite the corollaries and the appealing interpretation of this result which is given in Section VIII, the main interest in Proposition 2 lies in its usefulness in proving:

Proposition 3. For the l.r.w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}$, let the univariate innovations processes for X and Y be e and v , respectively. Then Y does not cause X if and only if $v(t), v(t-1), \dots$ are

uncorrelated with $e(t+1), e(t+2), \dots$; equivalently,

$$\sum_{s=-\infty}^t \oplus D_v(s) \perp \sum_{s=t+1}^{\infty} \oplus D_e(s).$$

Proof: Using the notation of the proof of the previous sections wherever possible,

$$(Y(t) | H_X(t)) = (Y(t) | \sum_{s=-\infty}^{s=t} \oplus D_e(s)) = \sum_{s=-\infty}^{s=t} (Y(t) | D_e(s))$$

and

$$(Y(t) | H_X) = (Y(t) | \sum_{s=-\infty}^{s=\infty} \oplus D_e(s)) = \sum_{s=-\infty}^{s=\infty} (Y(t) | D_e(s)).$$

Thus

$$|| (Y(t) | H_X(t)) - (Y(t) | H_X) ||^2 = \sum_{s=t+1}^{\infty} || (Y(t) | D_e(s)) ||^2$$

by the Pythagorean Theorem. But both directions of the proposition may now be proved with the aid of

$$(*) \quad Y(t) \perp e(t+j) \text{ all } j > 0 \Leftrightarrow Y(t) \perp D_e(t+j), \text{ all } j > 0 \Leftrightarrow$$

$$\begin{aligned} & (Y(t) | D_e(t+j)) = 0 \text{ all } j > 0 \Leftrightarrow || (Y(t) | D_e(t+j)) ||^2 \\ & = 0 \text{ all } j > 0 \Leftrightarrow \sum_{s=t+1}^{\infty} || (Y(t) | D_e(s)) ||^2 = 0 \Leftrightarrow (Y(t) | H_X(t)) \end{aligned}$$

$$= (Y(t) | H_X) \Leftrightarrow Y \text{ does not cause } X, \text{ using Proposition 4}$$

at the last step. By joint stationarity, (*) is equivalent

to $Y(t-k) \perp e(t+j), \text{ all } j > 0 \text{ and all } k \leq 0$. Since

$$H_V(t) = H_Y(t) = \sum_{s=-\infty}^{s=t} \oplus D_v(s) \text{ and } \sum_{s=t+1}^{\infty} \oplus D_e(s) = \overline{S(e(t+j), j > 0)},$$

the result now follows. Q.E.D.

The scalar process $X(t)$ whose autocovariance function is

$$R_X(\tau) = \begin{cases} 2 & \tau = 0 \\ -1 & \tau = +1, -1 \\ 0 & \text{elsewhere} \end{cases}$$

may be used to illustrate a case in which the assumptions of Propositions 2 and 3 are met while the theorems they generalize do not apply. The reason, of course, is that a process with moving average representation (m.a.r.) $X(t) = u(t) - u(t-1) = (1-L) u(t)$ is "widely known" not to permit an autoregressive representation (a.r.). ([28], p. 27; [14], p. 137). The usual evidence supporting this assertion is that the natural candidate for an inverse, $(1-L)^{-1} = 1 + L + L^2 + \dots$ results in the unpleasant $\lim_{n \rightarrow \infty} \sum_{i=0}^n X(t-i) = \lim_{n \rightarrow \infty} \{u(t) - u(t-n-1)\}$, which does not converge. This argument, of course, only proves that one attempt at finding an autoregressive representation has failed; to show that all candidates must fail is more difficult but instructive because the precise meanings of the commonly used terms m.a. and a.r. must be confronted. This is our excuse for proving in detail the following:

Lemma. The process $X(t) = u(t) - u(t-1)$, $u(t)$ white noise with unit variance, does not possess an autoregressive representation.

Proof: An a.r. is by definition a decomposition of the form $X(t) = \sum_{i=1}^{\infty} a(i) X(t-i) + e(t)$, where $e(t)$ is the innovation

in the $X(t)$ process; also by definition, an m.a.r. is a

decomposition of the form $X(t) = \sum_{i=0}^{\infty} b(i) e(t-i)$ where, if $b(0)$ is normalized to unity, $e(t)$ is again the one-step-ahead prediction error for $X(t)$, or innovation. To be more specific about the a.r. (to substitute $u(t)$ for $e(t)$) thus requires showing that $X(t) = u(t) - u(t-1)$ is indeed the m.a.r. Evidently, $H_X(t) \subseteq H_u(t)$; we need to prove that $u(t)$ is not just any driving white noise process, but that it is in the space from which predictors may be drawn, that it is in the linear manifold generated by current and past X . To show $H_u(t) \subseteq H_X(t)$ it suffices to get $u(t) \in H_X(t)$. We do this directly by producing a sequence of vectors in $H_X(t)$, $\{X(t) - \hat{X}_n\}$, which converge to $u(t)$ in the norm of $H_X(t)$; completeness of $H_X(t)$ then ensures that $u(t) \in H_X(t)$. We take for \hat{X}_n the projection of $X(t)$ onto $\langle X(t-1), X(t-2), \dots, X(t-n) \rangle$.²³ Writing out the normal equations yields

$$\hat{X}_n = \sum_{i=1}^n c(i) X(t-i) \text{ where the } c(i) \text{ satisfy}$$

$$\begin{pmatrix} -1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & & & \circ & \\ & & & & 2 & -1 \\ & & \circ & & -1 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ \cdot \\ \cdot \\ \cdot \\ c_n \end{pmatrix}, \text{ or } d = Ac.$$

Since A is symmetric, so is A^{-1} ; the first row or column may be verified to be $(\frac{n}{n+1} \frac{n-1}{n+1} \dots \frac{1}{n+1})$ which allows the

determination of c . Hence, $\hat{X}_n = - \sum_{i=1}^n (\frac{n+1-i}{n+1}) X(t-i)$. Now

$$(n+1)[X(t) - \hat{X}_n] = (n+1)X(t) + nX(t-1) + \dots + X(t-n) = (n+1)u(t) - u(t-1) - u(t-2) - \dots - u(t-n) - u(t-n-1). \text{ Consequently,}$$

$$\| (X(t) - \hat{X}_n) - u(t) \|^2 = \frac{n+1}{(n+1)^2} = \frac{1}{n+1} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad u(t)$$

is thus fundamental for $X(t)$ and hence $X(t) = u(t) - u(t-1)$ is a m.a.r. But $\hat{X}_\infty \equiv (X(t) | H_u(t-1)) = -u(t-1)$ is the optimal predictor of $X(t)$ using the entire past, leaving prediction error $X(t) - \hat{X}_\infty = u(t)$ with unit variance: This implies

$$(\dagger) \quad \| d^*X(t) \|^2 \geq 1 \text{ for any } d^*X(t) \equiv \sum_{i=0}^{\infty} d(i)X(t-i)$$

with $d(0) = 1$ which converges. We use (\dagger) in concluding.

Because $\hat{X}_n + u(t) \rightarrow X(t)$ as $n \rightarrow \infty$, if an a.r. exists, say $X(t) = \sum_{i=1}^n a(i)X(t-i) + u(t)$, then $\lim_{n \rightarrow \infty} \sum_{i=1}^n a(i)X(t-i) = \lim_{n \rightarrow \infty} \hat{X}_n$, or $\lim_{n \rightarrow \infty} X_n, \sum_{i=1}^n (a(i) + \frac{n-i}{n+1})X(t-i) = 0$. If $a(i) \neq -1$ for some i ,

let i' be the first such i . We then have, after renormalizing via $d(i-i') \equiv (a(i) + \frac{n-i}{n+1}) / (a(i') + \frac{n-i'}{n+1})$, $\lim_{n \rightarrow \infty} \sum_{i=0}^n d(i)X(t-i'-i)$

$\equiv d^*X(t-i') = 0, d(0) = 1$. Stationarity implies $d^*X(t) = 0$, contradicting (\dagger) . Thus, the only possible candidate for an a.r. is $\sum_{i=1}^{\infty} X(t-i)$, but we have already seen that this does not

converge. Q.E.D. As a final tutorial comment, $u(t-1)$

illustrates two technical points: $u(t-1) \overline{\bigcup_{j=1}^{\infty} X(t-j)}$ but

$u(t-1) \notin \bigcup_{j=1}^{\infty} X(t-j)$, showing the need for closures; and $u(t-1),$

while a limit of finite linear combinations of past X , is not an infinite linear combination of past X .

By exogenously embedding this X process into a bivariate system, the desired example may be constructed. Thus, if we take

$$\begin{pmatrix} X \\ Y \end{pmatrix}(t) \equiv \begin{pmatrix} 1-L & 0 \\ 1 & 1-L \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}(t) \equiv B(L) \begin{pmatrix} u \\ v \end{pmatrix}(t),$$

$\begin{pmatrix} u \\ v \end{pmatrix}(t)$ v.w.n. with contemporaneous covariance the identity matrix, then by the cited reference in Rozanov ([17], p. 88) $B(z)$ is a maximal matrix and $\begin{pmatrix} u \\ v \end{pmatrix}(t)$ thus jointly fundamental for $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$. By Sims' Theorem 1, X is exogenous by the form of the m.a.r., and previous discussion has established that $u(t)$ is the univariate innovation for $X(t)$, although $v(t)$ is not the univariate innovation for $Y(t)$. Proposition 2 is illustrated by observing

$$\begin{aligned} (Y(t) | H_X(t)) &= (u(t) + v(t) - v(t-1) | H_X(t)) \\ &= u(t) = (Y(t) | H_X(-\infty, \infty)); \end{aligned}$$

$u(t)$ is not expressible by a distributed lag on current and past X , by the Lemma. The message of Proposition 3 is that current and past Y , which will be linear combinations of current and past u and v , will be orthogonal to all future innovations in the X process, i.e., all future u . Of course, the Y innovations would, if derived, enjoy this orthogonality property as well.

We conclude this section by commenting on what may

be the independent discovery of Proposition 3 in the unpublished works [8], [9], and [15]; for specificity, we will concentrate on Theorem 4.2.7 of [15], although the idea in one form or another undoubtedly goes back to [8]. In any event, the proof of 4.2.7 states that, in the moving average representation

$$(*) \quad \begin{pmatrix} u \\ v \end{pmatrix} (t) = \begin{pmatrix} \Theta_{11}(B) & \Theta_{12}(B) \\ \Theta_{21}(B) & \Theta_{22}(B) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} (t), \quad \Theta_{11}(0) = \Theta_2(0) = 1,$$

(B here is our L, the lag operator)

since $u(t)$, $a(t)$, and $b(t)$ are each white noise, it follows that $\Theta_{11}(B) = 1$ whenever $\Theta_{12}(B)$ is a constant or zero. This is a crucial step in their proof which, in this author's opinion, represents a lacuna. Considering the case where $\Theta_{12}(B) = 0$, let

$$\Theta_{11}(B) \equiv 2 \frac{1-2B}{2-B} = 1 - \frac{3}{2}B - \frac{3}{4}B^2 - \dots;$$

it is evident that $\Theta_{11}(B)$, while not the identity, nevertheless maps a white noise input $a(t)$ into a white noise output $u(t)$: that $|\Theta_{11}(e^{i\lambda})|^2 = 1$ for $\lambda \in [0, 2\pi]$ is the easiest way to see this, but it follows by direct computation as well. Of course, $\Theta_{11}(B)$ maps a nonfundamental "innovation" process into a fundamental innovation process, but since these concepts are not used in [15], the inference might be that their operational method of proof breaks down at this point.

If this particular $\Theta_{11}(B)$ is dismissed on the grounds of the "invertibility" assumption on $|\pi(z)|$ made earlier, the

question of the burden of proof still seems open. We question here the soundness, not the validity of the deduction; indeed, Proposition 3 shows that the result holds without any assumption of "invertibility," i.e., without the assumption that an a.r. exists. Actually, this point may be important in practice, if processes are known not to have autoregressive representations due, for example, to seasonal adjustment procedures.

IV. The Forward Flow of Time; Symmetries and Asymmetries; Time Reversal

In this section, we consider the effect of what may be termed time reversal on the Wiener-Granger-Sims notion of statistical causality. Our finding will be that, while all of the previous theorems have natural analogues, when time is reversed the property $Y \vdash X$ is itself not invariant, except in a special case.

Situating ourselves at time t and considering $X(t+1)$ the classical prediction problem involved projecting into $H_{X,Y}(t)$, because this space represented the past, the data at hand. If time were "flowing backwards" or "reversed," we can imagine knowing, instead, only the future,

$(X(t+i), Y(t+j), i, j=1, 2, \dots)$, a family of random variables the closure of whose span is $H_{X,Y}(t+1, \infty)$, and trying to "predict" $X(t)$ by projecting onto $H_{X,Y}(t+1, \infty)$. Denoting the latter space by $\underline{H}_{X,Y}(t+1)$, we define "Y does not cause X under time reversal" (notation: $Y \not\vdash^{\overleftarrow{t}} X$) whenever future Y does not help in predicting current X, given future X; in

symbols, $(X(t) | \underline{H}_{X,Y}(t+1)) = (X(t) | \underline{H}_X(t+1))$. As with the usual definition, the LHS has in general a lower predictive variance because the projection is onto a larger subspace; as before, time reversal exogeneity of X with respect to Y (synonymous with $Y \stackrel{t, X}{\dashv} X$) indeed represents a testable hypothesis. To avoid the use of an awkward phrase, we will throughout this section describe "predicting the present from the future" as "backcasting."

The counterparts to the ordinary, Section II constructs of prediction theory will be indicated by an underline, to emphasize symmetry, continuing the precedent of the preceding paragraph. Thus, the crucial decomposition $\underline{H}_{X,Y}(t) = \underline{H}_{X,Y}(t+1) \oplus \underline{D}_{X,Y}(t)$ leads, as before, to $\underline{H}_{X,Y}(t) = \sum_{s=t}^{\infty} \underline{D}_{X,Y}(s) \oplus \underline{H}_{X,Y}(\infty)$, the latter term representing the infinite future, $\bigcup_{s=-\infty}^{\infty} \underline{H}_{X,Y}(s)$, which we define as $\underline{H}_{X,Y}(\infty)$. A random variable contained in $\underline{H}_{X,Y}(\infty)$ can be backcast arbitrarily distantly, given any stretch of the future, $\underline{H}_{X,Y}(s, \infty)$, no matter how far removed (how large s). From its description, it may be thought that $\underline{H}_{X,Y}(\infty) = \{0\}$ on physical grounds in most applications; such processes we define as linearly regular on the future (l.r.f.). Processes for which $\underline{H}_{X,Y}(-\infty)$ and $\underline{H}_{X,Y}(\infty)$ both are $\{0\}$ will be called totally linearly regular (t.l.r.), and might be considered the rule rather than the exception.

We recall from the discussion in Section II that,

l.r. or not, it was the l.r. part of a w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ which had a moving average representation; when it is understood that the structural theorems concern this part, it is a matter of aesthetics whether the deterministic part exists or is the zero vector. In other words, the assumption of l.r. could be made without loss of generality. So, too, it is here, with both the concepts of l.r.f. and t.l.r.: we state results for totally regular processes to economize on words, fully cognizant of the fact that the result applies to the regular parts of non-t.l.r. processes as well.

One of the reasons so much background was presented earlier, and the particular version of the Wold decomposition was given, occurs at this juncture. Once the orthogonal decomposition of the space $H_{X,Y}(t, \infty) = \sum_{s=t}^{\infty} \oplus D_{u,w}(s) \oplus H_{X,Y}(\infty)$ is available, a reversed version of a moving average representation falls out, just as before, by projecting $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ onto an orthogonally decomposed subspace of which it is an element. To do this, it is a matter of collecting Fourier coefficients, remembering that convolutions now extend forward in time, and noting that innovations now refer to optimal, one-step-behind backcast errors and that mutually subordinate means $H_{X,Y}(t, \infty) = H_{u,v}(t, \infty)$. Incorporating into our definition of l.r.f. the notion of full rank of the matrix of backcast errors, we have the, now underlined,

Wold Decomposition Theorem. The l.r.f. w.s.s.p. $\begin{pmatrix} X \\ Y \end{pmatrix}$ has the moving average representation forward (m.a.r.f.)

$$\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \sum_{k=0}^{\infty} \underline{A}(k) \begin{pmatrix} u \\ w \end{pmatrix}(t+k) \equiv \underline{A} * \begin{pmatrix} u \\ w \end{pmatrix}(t), \text{ where } \underline{A}(0)$$

$$= \underline{I}, \text{ Cov } \begin{pmatrix} u \\ w \end{pmatrix} = \underline{\Sigma}, \underline{A}(k) = \left(\begin{pmatrix} X \\ Y \end{pmatrix}(0), \begin{pmatrix} u \\ w \end{pmatrix}(k) \right) \underline{\Sigma}^{-1},$$

$$\text{and trace } \sum_{k=0}^{\infty} \underline{A}(k) \underline{\Sigma} \underline{A}'(-k) < \infty. \underline{H}_{X,Y}(t) = \underline{H}_{u,v}(t),$$

so that $\begin{pmatrix} u \\ w \end{pmatrix}$ is the innovations process.

Another analogous notion is the autoregressive representation forward (a.r.f.), which has the form

$$\underline{B} * \begin{pmatrix} X \\ Y \end{pmatrix}(t) \equiv \sum_{k=0}^{\infty} \underline{B}(k) \begin{pmatrix} X \\ Y \end{pmatrix}(t+k) = \begin{pmatrix} u \\ w \end{pmatrix}(t), \underline{B}(0) = \underline{I}, \text{ Cov } \begin{pmatrix} u \\ w \end{pmatrix} = \underline{\Sigma},$$

where again $\begin{pmatrix} u \\ w \end{pmatrix}$ and $\begin{pmatrix} X \\ Y \end{pmatrix}$ are mutually subordinate into the future.

Of course, the same normalization questions and answers arise, and the previous use of analytic function theory can be carried over to distinguish a fundamental m.a.r.f. from a nonfundamental one. An immediate consequence, for scalar processes, is a symmetry between past and future (which does not extend to vector processes).

Lemma. For the l.r.w.s.s.p. $X(t)$ the one-step-ahead and one-step-behind prediction errors have the same variance. Also, $H_X(-\infty) = \{0\} \Rightarrow H_X(+\infty) = \{0\}$, so that l.r. implies t.l.r.

Proof: Since $X(t)$ is l.r., the Wold decomposition yields $X(t) = b*u(t)$, where we may take $\sigma_u^2=1$. Furthermore,

([17], p. 60), $b(z)$ is maximal among analytic "matrices" with components in H_2 which factor the autocovariance function: $R_X(z) = b(z)b(z^{-1})$, and $|b(0)|^2 \geq |\tilde{b}(0)|^2$ for any other factorizing $\tilde{b}(z)$. But $X(t) = \underline{b} * \underline{u}(t)$ the same $b(\cdot)$ sequence again represents $X(t)$, because $R_X(\cdot)$ is a symmetric function. Consequently, there is no nonzero element in $H_X(+\infty)$ either. And the maximality condition which $b(\cdot)$ is known to satisfy is precisely that which guarantees that $\underline{u}(t)$ is future fundamental. Thus, $X(t) - (X(t) | H_X(t+1)) = b(0)\underline{u}(t)$, $\text{Var } b(0)\underline{u}(t) = |b(0)|^2 = \text{Var } b(0)u(t)$ where $b(0)u(t) = X(t) - (X(t) | H_X(t-1))$. One-step-ahead and backward forecast errors have thus been shown to have the same variance. Q.E.D.

We remark that the reason this result does not carry over to vector processes is because the matrix analogue of $b(\cdot)$, $B(\cdot)$, does not continue to factor $R_{X,Y}(\cdot) = B*B'(\cdot)$, since the latter is not symmetric in the multivariate case. Although it would take us off the track to prove it, we claim that a multivariate quantity which is invariant under time reversal is $|\Sigma|$, that is, $|\Sigma| = |\underline{\Sigma}|$: generalized variance is preserved.

It is now a question of substituting analogous concepts in the straightforward and obvious way to prove the next proposition. We begin the task where it is instructive, mimicking the proof of Sims' Theorem 1, from which the Granger result may be quickly derived.

Proposition 4. Let $\begin{pmatrix} X \\ Y \end{pmatrix}(t)$ be a l.r.f. w.s.s.p. Then X is time-reversed exogenous with respect to Y ($Y \xrightarrow{t, F} X$) if and only if:

(i) The m.a.r.f. is given by $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \begin{pmatrix} \underline{a} & 0 \\ \underline{c} & \underline{d} \end{pmatrix} * \begin{pmatrix} u \\ w \end{pmatrix}(t),$

i.e., $\underline{b}(\cdot) \equiv 0$; where the normalization is

$$\underline{a}(0) = \underline{d}(0) = 1, \underline{b}(0) = 0; \text{Cov} \begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix};$$

(the analogue of III-L).

(ii) When an a.r.f. exists, it is given by $\begin{pmatrix} \alpha & 0 \\ \gamma & \delta \end{pmatrix} *$

$$\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \begin{pmatrix} u \\ w \end{pmatrix}(t), \text{ where the normalization is as}$$

in (i), III-L.

(iii) $(Y(t) | H_X(t, \infty)) = (Y(t) | H_X(-\infty, \infty)).$

(iv) $\underline{e}(t), \underline{e}(t-1), \dots$ are uncorrelated with $\underline{v}(t+1), \underline{v}(t+2), \dots$, where $Y(t) = \underline{\mu} * \underline{v}(t)$ and

$$X(t) = \underline{a} * \underline{e}(t), \underline{\mu}(0) = \underline{a}(0) = 1, \text{ are univariate m.a.r.f.'s.}$$

Proof:

(i) \Leftarrow : With the given m.a.r.f., $X(t)$ lies in $H_u(t)$.

By definition of m.a.r.f., $H_{X,Y}(t) = H_{u,v}(t),$

and the earlier remarks in Section II show that

$$H_X(t) = H_u(t), \text{ or else, by an analagous maximality}$$

argument, the mutual subordination of the $\begin{pmatrix} X \\ Y \end{pmatrix}$ and

$\begin{pmatrix} u \\ v \end{pmatrix}$ processes would be contradicted. Now forming

$$\text{the projection } (X(t) | H_{X,Y}(t+1)) = \sum_{i=1}^{\infty} \underline{a}(t) \underline{u}(t+i),$$

we note that this is in $\underline{H}_X(t+1)$, and hence equals $(X(t) | \underline{H}_X(t+1))$. Consequently, future Y do not help predict current X. \Rightarrow : Assuming now the equality of these projections, and our definition of $Y \xrightarrow{t, r} X$, we may now write

$$(4.1) \quad X(t) - (X(t) | \underline{H}_{X,Y}(t+1)) = X(t) - (X(t) | \underline{H}_X(t+1)) \equiv \underline{u}(t)$$

$$(4.2) \quad Y(t) - (Y(t) | \underline{H}_{X,Y}(t+1)) \equiv \underline{w}(t).$$

$$\text{Now we define } \underline{v}(t) \equiv \underline{w}(t) - \frac{\langle \underline{w}(t), \underline{u}(t) \rangle}{\langle \underline{u}(t), \underline{u}(t) \rangle} \underline{u}(t)$$

so that $\underline{v}(t) \perp \underline{u}(t)$, and, of course, $\underline{v}(t) \perp \underline{u}(s)$, all t and s by the construction of the projections (cf. Section II, and remarks around the Wold decomposition). Thus, $(\underline{u}(s), \underline{v}(s), s=t, t+1, \dots)$ form a complete orthonormal system, $\underline{H}_{X,Y}(t) = \underline{H}_{\underline{u}, \underline{v}}(t)$, and taking Fourier representations of X and Y yields a representation of the lower triangular form. Q.E.D.

- (ii) By inverting the m.a.r.f., the a.r.f. is obtained, when it exists. Since lower triangularity is preserved, the result follows. Another proof is available by mimicking that of Proposition 1.
- (iii) and (iv) follow from Propositions 2 and 3, again by making the obvious replacements. Q.E.D.

A symmetry carries over to time reversed processes in

the sense that, if $Y \xrightarrow{t, r} X$, then the analogous results hold, and conversely. However, it quickly becomes apparent that $Y \dashrightarrow X$ does not in general hold up under time reversal, except in the special case where $(Y(t) | H_X(-\infty, \infty))$ is in $S(X(t))$, as we prove for totally regular processes in

Proposition 5. Let $\begin{pmatrix} X \\ Y \end{pmatrix}$ be a t.l.r.w.s.s.p. Then if and only if $(Y(t) | H_X(-\infty, \infty)) = k \cdot X(t)$ does $Y \dashrightarrow X$ imply $Y \xrightarrow{t, r} X$. The result remains valid when $Y \dashrightarrow X$ and $Y \xrightarrow{t, r} X$ are interchanged. In this case, when $a^{-1}(\cdot)$ exists, the m.a.r.f. and the m.a.r. may be expressed with the same coefficients:

$$(4.3) \quad \begin{pmatrix} X \\ Y \end{pmatrix} (t) = \begin{pmatrix} a & 0 \\ ka & d \end{pmatrix} * \begin{pmatrix} u \\ w \end{pmatrix} (t) = \begin{pmatrix} a & 0 \\ ka & d \end{pmatrix} * \underline{\begin{pmatrix} u \\ w \end{pmatrix}} (t),$$

$$\text{where } a(0) = d(0) = 1, \text{ Cov} \begin{pmatrix} u \\ w \end{pmatrix} (t) = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix} =$$

$$\text{Cov} \begin{pmatrix} u \\ w \end{pmatrix} (t), \text{ and } k=0 \text{ if and only if } X \xrightarrow{i.c.} Y.$$

Proof: While the first two statements follow immediately from Proposition 4 (iii) and Proposition 2, they will also follow from the proof of 4.3. Indeed, since $Y \dashrightarrow X$, a III-L normalized lower triangular representation exists:

$$\begin{pmatrix} X \\ Y \end{pmatrix} (t) = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} * \begin{pmatrix} u \\ w \end{pmatrix} (t).$$

For $Y \xrightarrow{t, r} X$ to hold, there must also be such a lower triangular representation on the future, $\begin{pmatrix} X \\ Y \end{pmatrix} (t) = \begin{pmatrix} a & 0 \\ e & f \end{pmatrix} * \underline{\begin{pmatrix} u \\ w \end{pmatrix}} (t)$, where we have used a fact encountered in the lemma: the same $a(\cdot)$

must be present in both. But writing out the cross-autocovariance $R_{YX}(\cdot)$ for each of the representations, we see that,

$$c*a' = e'*a, \text{ or, in terms of lag operators, } \frac{c(L^{-1})a(L)}{a(L^{-1})} = e(L),$$

using the assumption that $a^{-1}(\cdot)$ exists under convolution.

Now in this last equality, the LHS must contain no terms in L^{-1} (since $a(0)=1$); consequently, $c(L^{-1}) = k.a(L^{-1})$, or $c(\cdot)$

$= ka(\cdot)$. The statement about i.c. is part of Proposition 1 (ii).

The regression convolution coefficients of $(Y(t) | H_X(-\infty, \infty))$

may be computed in this case as $(a*a')^{-1}*ka*a' = k.\delta(\cdot)$.

$\therefore e(\cdot) = c(\cdot)$, and $R_Y = c*c' + d*d' = e*e' + f*f'$ entails

$d*d' = f*f'$. Maximality ensures $d=f$. Q.E.D.

Corollary. When $Y \dashv X$ and $\begin{pmatrix} X \\ Y \end{pmatrix}$ is a t.l.r. w.s.s.p., a prognosticator desiring to predict $X(t)$ and, given the choice between the future $H_{X,Y}(t+1)$ and the past $H_{X,Y}(t-1)$ will always choose the future, although he may be indifferent.

Proof: Since X is exogenous, if $X(t) = a*u(t)$, $\sigma_u^2=1$, by using only past X and by using only future X , the predictive variance has been shown to be same: $|a(0)|^2$. But this is also the mean square error when using past Y as well. Thus, future X allows as accurate a forecast as past X and Y , so future X and Y can do no worse than past X and Y . In the case of Proposition 5, it does only as well; in all other cases, the m.a.r.f. is not lower triangular, and the future will in general dominate in these cases. Q.E.D.

V. Multivariate Propositions

To perform the extension of the concepts of the previous sections to n dimensions, we write (the only occasion in this paper where X represents a vector) the l.r.w.s.s.p. $X(t)$

$$= A * e(t), H_e(t) = H_X(t) \text{ for the m.a. and } B * X(t) = e(t),$$

$$H_e(t) = H_X(t) \text{ for the a.r. (where it exists). Since the}$$

underlying mathematics (prediction theory) is available in the sources mentioned in Section II, the previous bivariate proofs may be adapted to prove results where block-triangularity

replaces triangularity. From a technical point of view it is the fact that $\det \begin{pmatrix} A(0) & 0 \\ C(0) & D(0) \end{pmatrix} = \det A(0) \det D(0)$ which allows

the theory of maximal matrix factorizations to be again used with the same advantage that was explained on p. 14. Since

block-triangularity is preserved under convolution and matrix inverse, we expect and find the same qualitative results

present in the bivariate case. A special case occurs when one of the blocks on the "diagonal" is a scalar: we indicate

this by writing x_1 for a scalar and X_1 for a vector. If $X = \begin{pmatrix} x_1 \\ X_2 \end{pmatrix}$ where x_1 is 1×1 and X_2 is $(n-1) \times 1$, then we say, as

before, that x_1 is exogenous w.r.t. X_2 if

$(x_1(t) | \overline{H_{x_1}(t-1) \cup H_{X_2}(t-1)})$ and $(x_1(t) | H_{x_1}(t-1))$ agree, or, X_2 does not cause (help predict) x_1 (notation: $X_2 \nrightarrow x_1$).

Now a new concept emerges: it may be that x_1 doesn't help in the prediction of some or all elements of X_2 . In the

latter case, when $(X_2(t) | \overline{H_{X_2}(t-1) \cup H_{x_1}(t-1)}) = (X_2(t) | H_{X_2}(t-1))$

we write $x_1 \not\rightarrow X_2$.

Generally, the notion of "does not cause" is very dependent on the conditioning set, as the next section will show. It will be seen that for systems with the block triangularity of the next proposition, much freedom is available in specifying the conditioning set. Nevertheless, we understand $x_i \not\rightarrow x_j$ to mean that x_i does not help predict x_j when x_i is added to the information set consisting of the past of all the other variables in the system (including x_j).

If $x_k \not\rightarrow x_j$ for all $k \in K$, $j \in J \equiv \{1, \dots, n\} \setminus K$, then we write $X_2 \not\rightarrow X_1$, where X_2 contains the x_k and X_1 contains the x_j . Thus $X_2 \not\rightarrow X_1$ means that no component of X_2 helps predict any component of X_1 ; the symbol $\not\rightarrow$ refers to $n_1 \cdot n_2$

elementary causality events of the form $x_i \not\rightarrow x_j$. To characterize these latter events, the a.r. is the most convenient, as Proposition 6 (i) shows. However, in describing results involving one component, say x_1 , and $(x_2, \dots, x_n)^T \equiv X_2$, the rest of the system, the a.r. and m.a. again have the same qualitative appearance, if a relation $\not\rightarrow$ is present, as in the bivariate case; here, however, both upper and lower triangular representations have an obvious interpretation.

We choose the natural parameterization, in which $A(0) = B(0) = I$ below, deferring any discussion of instantaneous causality until the next section. We record:

Proposition 6. For the l.r.w.s.s.p. with m.a. $X(t) = A * e(t)$, a.r. $B * X(t) = e(t)$, and $X^T = \begin{pmatrix} x_1^T & x_2^T \end{pmatrix}$,

- (i) $x_1 \dashv\vdash x_j$ if and only if $b_{ji}(\cdot) \equiv 0$ in the a.r.
- (ii) $x_2 \dashv\vdash x_1$, or x_1 is exogenous, if and only if either of the equivalent conditions hold:
 - (a) $(b_{12}(\cdot) \dots b_{1n}(\cdot)) \equiv 0$ in the a.r.
 - (b) $(a_{12}(\cdot) \dots a_{1n}(\cdot)) \equiv 0$ in the m.a.
- (iii) $x_1 \dashv\vdash x_2$, or x_1 does not cause any other variable in the system, if and only if either of the equivalent conditions hold:

$$(a) \begin{pmatrix} b_{21}(\cdot) \\ \vdots \\ b_{n1}(\cdot) \end{pmatrix} \equiv 0 \qquad (b) \begin{pmatrix} a_{21}(\cdot) \\ \vdots \\ a_{n1}(\cdot) \end{pmatrix} \equiv 0$$

- (iv) In the results (ii) and (iii), x_1 may be replaced by $\begin{matrix} n_1 \times 1 \\ X_1 \end{matrix}$, $\begin{matrix} (n-1) \times 1 \\ X_2 \end{matrix}$ by $\begin{matrix} n_2 \times 1 \\ X_2 \end{matrix}$, ($n_1 + n_2 = n$) and the conditions (a) and (b) by the upper right and lower left matrices in the conformably partitioned a.r. and m.a. representations.
- (v) Propositions 2 and 3, on one-sided projections and zero correlation of future X_1 innovations with past and present X_2 innovations, remain valid when interpreted as in parts (ii)-(iv) of this theorem.

Proof: All parts may be tediously demonstrated by repeating previous arguments with scalars replaced by vectors. Part (i) is proved in exactly the same manner as part (i) of Proposition 1.

The only new features are: recognition of the supremacy of the a.r. for the characterization of basic causality events in terms of zero lag distributions; the observation of an interpretation for zeros in the lower left blocks; and the choice of the particular parameterization to simultaneously allow the statements (ii) and (iii).

We will make use of this proposition in interpreting the results of the next proposition.

VI. Trivariate Systems and Bivariate Causality; Notions of Instantaneous Causality-Feedback

In Section II we remarked that all of the mathematical complexities of general, n-variate prediction theory are present for $n=2$. This does not mean that statements made as if the universe were bivariate will necessarily retain their validity when embedded in the natural way in a higher dimensional setting. Indeed, the presumption has been that findings of bivariate systems will generally be found spurious, and consequently overturned, when referred to the properly specified, larger system. Here, we propose to venture beyond the safety of the truism that "in general, everything depends on everything else" and to investigate what can go wrong (and right) in the simplest system of dimension higher than two.

Notions of instantaneous causality are first discussed. Although they don't enter into the main results, they do provide an understanding of a useful triangular normalization, which is then discussed. Next, a new (different from the specialization of Proposition 6 to $n=3$) characterization of "Y does not cause X in the trivariate system" is given along with the evident statistical test for its implementation.

The remaining major issue discussed is, if Y does not cause X in the trivariate system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ but the investigator instead examines the exogeneity of X in the bivariate system $\begin{pmatrix} X \\ Y \end{pmatrix}$, when will this implied bivariate system inherit the exogeneity of X? And conversely, what is the meaning of a finding of bivariate exogeneity relative to the higher order system of which it is a part? We first present a decomposition which provides both insight into the general case and immediate proof of some parts of our last proposition. A lemma pertinent to the subspaces emphasized by this decomposition is given. Then the last proposition, which gives two conditions, either of which bring bivariate and trivariate exogeneity into concordance, is proved. A discussion and interpretation of this result, and a suggestion of cases where the proposition is not likely to be helpful, conclude this section.

Logically prior to issues of normalization are notions of instantaneous causality; intimately related to any particular parameterization is the manner in which instantaneous

causality, if present, will manifest itself.

Perhaps the most natural definition of instantaneous (trivariate) causality is to say $Z(t)$ causes $X(t)$ instantaneously (notation: $Z \xrightarrow{i_1} X$) if and only if the error in predicting $X(t)$ given $Y(t)$ and all past $X, Y,$ and Z declines when $Z(t)$ is added; equivalently, in symbols, if

$$(6.1) \quad (X(t) | \overline{H_{X,Y,Z}(t-1) \cup Y(t)}) \neq (X(t) | \overline{H_{X,Y,Z}(t-1) \cup Y(t) \cup Z(t)}).$$

Alternatively, we may delete $Y(t)$ from the previous definition, and define $Z \xrightarrow{i_2} X$ as occurring when the addition of $Z(t)$ to $\overline{H_{X,Y,Z}(t-1)}$ helps lower the predictive variance: in symbols, if

$$(6.2) \quad (X(t) | \overline{H_{X,Y,Z}(t-1)}) \neq (X(t) | \overline{H_{X,Y,Z}(t-1) \cup Z(t)}).$$

As in the bivariate case, both relations are symmetric (that

is, $X \xrightarrow{i_1 \text{ (or } i_2)} Z$ if and only if $Z \xrightarrow{i_1 \text{ (or } i_2)} X$). Consequently,

the notation $X \leftrightarrow_{i_1 \text{ (or } i_2)} Z$ will be adopted when symmetry is proved. And, as in the bivariate case, the covariances between X and Z provide a handy criterion:

$$\sigma_{XZ}^{i_2} \neq 0 \Leftrightarrow X \leftrightarrow_{i_2} Z \text{ and } \sigma_{XZ.Y}^{i_1} \neq 0 \Leftrightarrow X \leftrightarrow_{i_1} Z.$$

The technique of proof is the same as was used in

Proposition 1, (ii), so the treatment here is terse. We take a m.a. to be $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) = A * e(t)$, $A(0) = I$, $\text{Cov } e(t) = \Sigma$. Since

the forecast error of $X(t)$ given the joint past is $e_1(t)$, and

$e_i(t) \perp H_{X,Y,Z}(t-1)$, $i=1,2,3$, adding $Y(t)$ is equivalent to adding $e_2(t)$. The second moment of the error from forecasting with the LHS of (6.1) is the variance of $e_1(t) - (e_1(t) | e_2(t))$; from the RHS, the mean square error is the variance of $e_1(t) - (e_1(t) | e_2(t) \cup e_3(t))$. Solving, the two

forecast errors are $e_1(t) - \frac{\sigma_{12}}{\sigma_{22}} e_2(t)$ and $e_1(t)$

$$- (e_2(t) \ e_3(t)) \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{12} \\ \sigma_{13} \end{pmatrix} = e_1(t)$$

$$- (e_2(t) \ e_3(t)) \frac{1}{\sigma_{22}\sigma_{33} - \sigma_{23}^2} \begin{pmatrix} \sigma_{33}\sigma_{12} - \sigma_{32}\sigma_{13} \\ -\sigma_{23}\sigma_{12} + \sigma_{22}\sigma_{13} \end{pmatrix}. \text{ Thus, it is}$$

clear that $Z \xrightarrow{i_1} X$ iff $\text{Cof } \sigma_{31} \equiv \begin{vmatrix} \sigma_{12} & \sigma_{13} \\ \sigma_{22} & \sigma_{23} \end{vmatrix} \neq 0$. Analogously,

$X \xrightarrow{i_1} Z$ iff, since the relevant forecast error is $e_3(t)$

$$- (e_1(t) \ e_2(t)) \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix} \begin{pmatrix} \sigma_{31} \\ \sigma_{32} \end{pmatrix}, \text{ Cof } \sigma_{13} \neq 0.$$

The symmetry of Σ thus entails symmetry of $\xleftrightarrow{i_1}$. As in the

bivariate case, $Z \xleftrightarrow{i_2} X$ if and only if $\sigma_{12} \neq 0$, as a computation above shows; symmetry of Σ thus extends to $\xleftrightarrow{i_2}$. Only in the

case where $\sigma_{12}\sigma_{23} = 0$ will there necessarily be agreement

between the within period notions $\xleftrightarrow{i_1}$ and $\xleftrightarrow{i_2}$ for X and Z ;

simple and multiple correlations needn't be in agreement,

unless additional conditions are in force.

We now proceed to derive a normalization of the a.r.,

which is analogous to III-L,D, in order to understand the meaning of certain zero-order coefficients being zero or nonzero; information about both variants of i.c. will be seen to be present. If the a.r. in the natural parameterization (n.p.) is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} (t)$$

where $\text{Cov} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} (t) = \Sigma$, and all terms in the convolution start

at $i=1$ (e.g., $\sum_{i=1}^{\infty} a_1(i)X(t-i)$, etc.), then we: retain the first

equation; replace the second equation, $(Y(t) | H_{X,Y,Z}(t-1))$

+ $e_2(t)$ by $(Y(t) | \overline{H_{X,Y,Z}(t-1) \cup X(t)}) + \underline{e}_2(t)$ where

$\underline{e}_2(t) \perp \overline{H_{X,Y,Z}(t-1) \cup X(t)}$; and, replace the third equation by

$(Z(t) | \overline{H_{X,Y,Z}(t-1) \cup X(t) \cup Y(t)}) + \underline{e}_3(t)$, where

$\underline{e}_3(t) \perp \overline{H_{X,Y,Z}(t-1) \cup X(t) \cup Y(t)}$. The reader who has pursued

matters to this point should not be confused by the presence

of the same lower bars that denoted backwards innovations in

Section IV; further, he will have no trouble showing that:

$$(6.3) \quad \begin{pmatrix} X(t) \\ Y(t) \\ Z(t) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\sigma_{21}}{\sigma_{11}} X(t) \\ r_1 X(t) + r_2 Y(t) \end{pmatrix} +$$

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 - \frac{\sigma_{21}}{\sigma_{11}} a_1 & b_2 - \frac{\sigma_{21}}{\sigma_{11}} b_1 & c_2 - \frac{\sigma_{21}}{\sigma_{11}} c_1 \\ a_3 - r_1 a_1 - r_2 a_2 & b_3 - r_1 b_1 - r_2 b_2 & c_3 - r_1 c_1 - r_2 c_2 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) + \begin{pmatrix} e_1(t) \\ e_2(t) \\ e_3(t) \end{pmatrix}$$

where $r_1 \equiv \frac{\text{Cof } \sigma_{13}}{\text{Cof } \sigma_{33}}$ and $r_2 \equiv \frac{-\text{Cof } \sigma_{23}}{\text{Cof } \sigma_{33}}$.

It seems reasonable to name this III-L,D, although it is autoregressive rather than moving average in nature. By bringing the contemporaneous vector on the RHS into the matrix convolution, new convolutions are naturally defined; e.g., the coefficient on X(t) in the second equation might be named $\underline{a}_2(s)$, $s=0,1 \dots$ and $\underline{a}_2(s)$ would be $\frac{\sigma_{21}}{\sigma_{11}}$ for $s=0$, $a_2(s) - \frac{\sigma_{21}}{\sigma_{11}} a_1(s)$ for $s=1,2, \dots$, etc. No confusion will arise if we drop the $\underline{\quad}$ and reuse the previous notation: a_1, a_2, \dots . Hence, what is important to notice in this representation,

$$(6.4) \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} (t) + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} (t),$$

$$\text{Cov} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}(t) = \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}, b_1(0) = c_1(0) = c_2(0) = 0,$$

$$\text{is that } a_2(0) = \frac{\sigma_{21}}{\sigma_{11}}, a_3(0) = \frac{\text{Cof } \sigma_{13}}{\text{Cof } \sigma_{33}}, b_3(0) = \frac{-\text{Cof } \sigma_{23}}{\text{Cof } \sigma_{33}}.$$

Consequently, the assumption $a_2(\cdot) \equiv 0$ rules out $X \leftrightarrow Y$, while the assumption $b_1(\cdot) \equiv 0$ does not have any implication for instantaneous causality. However, for $a_3(\cdot) \equiv 0$ in the new parameterization to correspond to X not causing Z , as in the natural parameterization, we must have both r_1 and $r_2 = 0$, a severe instantaneous causality assumption.

As in earlier sections, we have provided the algebraic normalization with a physical, innovations-related interpretation. And, as before, the natural parameterization is best suited to proofs exploiting the geometric nature of the spaces under consideration, as the next development shows.

The extension of the bivariate exogeneity results of the first three sections to the block-triangular systems of the previous section has been treated by investigators as sufficiently obvious that proofs are unwarranted; indeed, the statement of the proposition is itself often implicit. The natural analogue of Proposition 2 for part (iii) of Proposition 6 is that when $x_1 \vdash X_2$ the (two-sided) projection of the scalar $x_1(t)$ on the past, present, and future of the vector

$X_2(\cdot)$ will in fact be one-sided on the current and past. We prove this as a corollary of the next proposition which provides a multivariate characterization which would not be formally deduced by such playing with blocks. The earlier remark after Proposition 2 concerning the relaxation of the assumption of stationarity remains in force and will not be repeated. The dimensions of X, Y, and Z below may be taken as 1, 1, and n-2, although obviously block generalizations of this result hold also.

Proposition 7 In the vector process $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$, $Y \vdash X$ if and only

if $(Y(t) | \overline{H_{X,Z}(-\infty, t+k-1) \cup X(t+k)}) = (Y(t) | H_{X,Z}(-\infty, t+k-1))$,

all integer $k \geq 1$. When the projections are expressible as distributed lags, the result may be paraphrased as: if $Y(t)$ is regressed on the past and any finite number k of future X and Z , then the next X will, if permitted, never enter the regression.

Proof $Y \vdash X$ is equivalent to $(X(t) | H_{X,Y,Z}(t-1) \ominus H_{X,Z}(t-1)) = 0$,

or $(X(t) | \bigcup_{j=1}^{\infty} \{Y(t-j) - (Y(t-j) | H_{X,Z}(t-1))\}) = 0$: the notion of \ominus

as an inverse to \oplus is discussed on p.126. This in turn is equivalent to $X(t) \perp Y(t-k) - (Y(t-k) | H_{X,Z}(t-1))$ for any

integer $k \geq 1$. Because orthogonality relations (although not necessarily the covariances) are assumed invariant to time

shifts, shifting the last relation forward k units yields the

equivalent $(*) X(t+k) \perp Y(t) - (Y(t) | H_{X,Z}(-\infty, t+k-1))$, all $k \geq 0$.

But by definition $(Y(t) | \overline{H_{X,Z}(-\infty, t+k-1) \cup X(t+k)})$ is the vector whose residual is orthogonal to $\overline{H_{X,Z}(-\infty, t+k-1) \cup X(t-k)}$, and (*) indicates that $(Y(t) | H_{X,Z}(-\infty, t+k-1))$ meets the necessary qualifications; so the two projections are equal. Conversely, assuming the two projections to be equal leads immediately back to (*), which is equivalent to $Y \vdash X$. Q.E.D.

Corollary If $Y \vdash X$ and $Y \vdash Z$, then $(Y(t) | H_{X,Z}(-\infty, \infty)) = (Y(t) | H_{X,Z}(-\infty, t))$.

Proof Since $Y \vdash X$, $(Y(t) | H_{X,Z}(-\infty, t) \cup X(t+1)) = (Y(t) | H_{X,Z}(-\infty, t))$. Using the obvious multivariate analogue of the proposition for $Y \vdash Z$, Z now a vector, yields $(Y(t) | \overline{H_{X,Z}(-\infty, t) \cup Z(t+1)}) = (Y(t) | H_{X,Z}(-\infty, t))$. Hence, $(Y(t) | \overline{H_{X,Z}(-\infty, t) \cup Z(t+1) \cup X(t+1)}) = (Y(t) | H_{X,Z}(-\infty, t+1)) = (Y(t) | H_{X,Z}(-\infty, t))$, and the bootstrap method of proof is so clear that the formal induction argument is omitted. Continuity of the projection operation ensures that establishing the result for every finite k allows the passage to $H_{X,Z}(-\infty, \infty)$. Q.E.D.

The corollary indicates that a test of "Y not causing any other variable in the system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ " follows from regressing

Y on the past, current, and future of all other variables, $\begin{pmatrix} X \\ Z \end{pmatrix}$ together, and checking for one-sidedness of this joint regression on the past and present. Pairwise Sims tests of Y on the variables in $\begin{pmatrix} X \\ Z \end{pmatrix}$ are inappropriate to test the hypothesis $Y \vdash \begin{pmatrix} X \\ Z \end{pmatrix}$ and rightly so, since they weren't designed for this

purpose. As the proposition indicates, the appropriate test of $Y \overset{t}{\vdash} X$ involving regressions with future values asserts vanishing single coefficients in a sequence of regressions rather than a vanishing sequence of coefficients in a single regression. Finally, it is the former rather than the latter concept which corresponds to the notion of econometric exogeneity in multivariate time series regressions.

We now begin the investigation of the relationship between bivariate and trivariate exogeneity. As the last paragraph showed, trivariate exogeneity, is a misnomer, but nevertheless preferable to the ponderous trivariate lack of feedback. The symbol $Y \overset{t}{\vdash} X$ (respectively, $Y \overset{b}{\vdash} X$) indicates Y does not cause X in the trivariate (respectively, implied bivariate) system. Of course, the reader may interpret X, Y, and Z as vectors or scalars, as before.

It is instructive to develop a direct sum decomposition of the principal Hilbert spaces under consideration which gives one perspective from which to view the problem. By projecting $X(t)$ onto this decomposition, vectors result which, if zero, are equivalent to the desired elementary bivariate and trivariate exogeneity (or causality) events.

The decomposition is

$$(6.5) \quad [H_{X,Y,Z}(t-1) \ominus H_{X,Y}(t-1)] \oplus [H_{X,Y}(t-1) \ominus H_X(t-1)] \\ = [H_{X,Y,Z}(t-1) \ominus H_{X,Z}(t-1)] \oplus [H_{X,Z}(t-1) \ominus H_X(t-1)],$$

or

$$(6.6) \quad S_3 \oplus S_4 = S_2 \oplus S_1.$$

The spaces S_i in (6.6) appear below the terms which define them in (6.5), and the symbol for \oplus is defined in the standard way from \oplus . Thus, for example, $S_3 \oplus H_{X,Y}(t-1) \equiv H_{X,Y,Z}(t-1)$. Hence $(X(t)|S_3) + (X(t)|H_{X,Y}(t-1)) = (X(t)|H_{X,Y,Z}(t-1))$, so that $(X(t)|S_3) = 0$ if and only if $Z \stackrel{t}{\vdash} X$. Since similar remarks apply to the other three subspaces, the equivalence in the previous sentence makes clear the importance of the S_i .

By projecting $X(t)$ onto (6.6) and interpreting the results, we arrive at

$$(6.7) \quad (X(t)|S_3) + (X(t)|S_4) = (X(t)|S_2) + (X(t)|S_1)$$

$$\begin{array}{cccc} \begin{array}{c} t \\ Z \vdash X \end{array} & \begin{array}{c} b \\ Y \vdash X \end{array} & \begin{array}{c} t \\ Y \vdash X \end{array} & \begin{array}{c} b \\ Z \vdash X \end{array} \end{array}$$

An event is true if and only if the projection immediately above it is zero.

First, assume $Y \stackrel{b}{\vdash} X$, so that $(X(t)|S_4) = 0$. Since the remaining projection is the orthogonal direct sum of the other two, the Pythagorean theorem yields

$$(6.8) \quad ||(X(t)|S_3)||^2 = ||(X(t)|S_2)||^2 + ||(X(t)|S_1)||^2.$$

Then if $Z \stackrel{t}{\vdash} X$, both terms on the RHS of (6.8) must be zero, so both projections are zero. This one auxiliary hypothesis thus implies $Y \stackrel{t}{\vdash} X$ and $Z \stackrel{b}{\vdash} X$, of which the first is likely of most

interest. We thus have one situation in which bivariate exogeneity implies trivariate exogeneity.

Second, assume $Y \overset{t}{\vdash} X$ so that $(X(t) | S_2) = 0$. If we make the same auxiliary assumption, that $(X(t) | S_3) = 0$, we learn

$$(6.9) \quad (X(t) | H_{X,Y}(t-1) \oplus H_X(t-1)) = (X(t) | H_{X,Z}(t-1) \oplus H_X(t-1)).$$

But $S_4 \cap S_1 = \{0\}$, since a non-empty non-trivial intersection would make the system degenerate by arguments analogous to those used previously. Hence the only way for equality in (6.9) to obtain is if both terms are zero, or if $Y \overset{b}{\vdash} X$ and $Z \overset{b}{\vdash} X$.

Combining the last two paragraphs, under $Z \overset{t}{\vdash} X$, the relations $Y \overset{b}{\vdash} X$ and $Y \overset{t}{\vdash} X$ each imply the other. Yet the indispensability of this additional assumption indicates that the bivariate and trivariate relations aren't equivalent in general. The natural question is, are there other assumptions which bring the exogeneity concepts into concordance?

The last question, coupled with the fact that projections behave like conditional expectations with regard to nested subspaces, suggests the study of S_4 and S_2 . When $S_4 \subseteq S_2$, $Y \overset{t}{\vdash} X \Rightarrow Y \overset{b}{\vdash} X$.

Some insight into the relation between $S_4 \equiv H_{X,Y}(t-1) \oplus H_X(t-1)$ and $S_2 \equiv H_{X,Y,Z}(t-1) \oplus H_{X,Z}(t-1)$ is provided by the lemma below. Results for the relation between $S_3 \equiv H_{X,Y,Z}(t-1) \oplus H_{X,Y}(t-1)$ and $S_1 \equiv H_{X,Z}(t-1) \oplus H_X(t-1)$ follow immediately, and the decomposition $S_3 \oplus S_4 = S_2 \oplus S_1$ is responsible for

the equivalences between the even- and odd-subscripted subspaces.

- Lemma (i) $H_Z(t-1) \perp H_{X,Y}(t-1) \Rightarrow S_4 = S_2 (\Leftrightarrow S_1 = S_3)$
- (ii) $H_Z(t-1) \perp S_4 \Leftrightarrow S_4 \subseteq S_2 (\Leftrightarrow S_1 \subseteq S_3)$
- (iii) $S_4 = S_2$ and $H_Z(t-1) \perp H_X(t-1) \Rightarrow H_Z(t-1) \perp H_{X,Y}(t-1)$
- (i)' $H_Y(t-1) \perp H_{X,Z}(t-1) \Rightarrow S_1 = S_3$
- (ii)' $H_Y(t-1) \perp H_{X,Z}(t-1) \oplus H_X(t-1) \Leftrightarrow S_1 \subseteq S_3 \Leftrightarrow S_4 \subseteq S_2$
- (iii)' $S_1 = S_3$ and $H_Z(t-1) \perp H_Y(t-1) \Rightarrow H_Z(t-1) \perp H_{X,Z}(t-1)$

Proof (i) The hypothesis is equivalent to $H_{X,Y,Z}(t-1) = H_Z(t-1) \oplus H_{X,Y}(t-1)$. Let $\phi \in S_4$, so $\phi \in H_{X,Y}(t-1)$ and $\phi \perp H_X(t-1)$. So $\phi \perp H_Z(t-1)$, hence $\phi \perp H_{X,Z}(t-1)$, so $\phi \in S_2$. If $\phi \in S_2$, $\phi \perp H_Z(t-1)$, so by hypothesis $\phi \in H_{X,Y}(t-1)$. Since $\phi \perp H_Z(t-1)$, $\phi \in S_4$. Hence $S_4 = S_2$. (ii) Let $\phi \in S_4$, so $\phi \in H_{X,Y}(t-1) \subseteq H_{X,Y,Z}(t-1)$ and $\phi \perp H_X(t-1)$. By hypothesis, $\phi \perp H_Z(t-1)$, so $\phi \perp H_{X,Z}(t-1)$, hence $\phi \in S_2$. Conversely, let $\eta \in H_Z(t-1)$; since $(\eta|S_2) = \eta - \eta = (\eta|S_4)$, $\eta \perp S_4$. Hence $H_Z(t-1) \perp S_4$. (iii) Since $S_4 \subseteq S_2$, by (ii) $H_Z(t-1) \perp S_4$. Since $S_4 \oplus H_X(t-1) = H_{X,Y}(t-1)$, the result follows. The primed results follow from the unprimed counterparts by interchanging Y and Z. Finally, assume $S_4 \subseteq S_2$, so that $S_4 \oplus S_4^\perp = S_2$, where S_4^\perp is the orthogonal complement of S_4 relative to the space S_2 . Hence $S_3 \oplus S_4 = S_4 \oplus S_4^\perp \oplus S_1$, so that $S_3 = S_1 \oplus S_4^\perp$ and $S_1 \subseteq S_3$. By

symmetry the last equivalence (ii)' is proved. Q.E.D.

It would be interesting to characterize the equality of S_2 and S_4 since this would give a sufficient condition for $(X(t)|S_4) = (X(t)|S_2)$. (i) produces $S_4 = S_2$, but at the too stringent assumption that the Z process is completely orthogonal to the X and Y processes. It turns out that (ii) is a just right condition which brings $Y \overset{t}{\perp} X$ and $Y \overset{b}{\perp} X$ into concordance; the symmetry of the lemma suggests that this same condition also matches up $Z \overset{t}{\perp} X$ and $Z \overset{b}{\perp} X$, and a repetition of the proof shows this to be so.

Motivated by the lemma, we are now able to prove, from a different perspective, all that has been shown, and more, in

Proposition 8 Let $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ be a nondegenerate discrete stochastic

process with finite variances at each time.

Without additional assumptions, there is no logical relation between $Y \overset{t}{\perp} X$ and $Y \overset{b}{\perp} X$. However, maintaining either (or both) of the assumptions (a) $Z \overset{t}{\perp} X$; (b) $H_Z(t-1) \perp \{H_{X,Y}(t-1) \oplus H_X(t-1)\}$; makes the exogeneity events $Y \overset{b}{\perp} X$ and $Y \overset{t}{\perp} X$ logically equivalent.

Proof The first assertion is evident from (6.7) and the

freedoms of choice in specifying properties for projections to obey. We have seen that $Y \overset{t}{\perp} X$ is equivalent to (*) $(X(t)|H_{X,Y,Z}(t-1)) = (X(t)|H_{X,Z}(t-1))$ and $Y \overset{b}{\perp} X$ is equivalent to (**) $(X(t)|H_{X,Y}(t-1)) = (X(t)|H_X(t-1))$. We may write the LHS of (*) uniquely as

$\alpha + \beta + \gamma$, where $\alpha \in H_X(t-1)$, $\beta \in H_Y(t-1)$, and $\gamma \in H_Z(t-1)$. This follows from the (not necessarily orthogonal) decomposition $H_{X,Y,Z}(t-1) = H_X(t-1) + H_Y(t-1) + H_Z(t-1)$, where the directness comes from the same kind of nondegeneracy argument used previously; that is, nonuniqueness would imply that a vector could be perfectly predicted from current and lagged values of the other variables. (*) is equivalent to $\alpha + \beta + \gamma = \alpha + \gamma + (\beta | H_{X,Z}(t-1))$, or, $\beta = (\beta | H_{X,Z}(t-1))$, or $\beta \in H_{X,Z}(t-1)$. This, the non-degeneracy argument has shown, is possible if and only if $\beta = 0$. Now forming the LHS of (**), $(X(t) | H_{X,Y}(t-1)) = \alpha + \beta + (\gamma | H_X(t-1) + H_Y(t-1)) = \alpha + \beta + \gamma_1 + \gamma_2$, where $\gamma_1 \in H_X(t-1)$ and $\gamma_2 \in H_Y(t-1)$ are unique, although again not necessarily orthogonal. In this language, (**) holds if and only if $\beta + \gamma_2 = 0$. That γ_2 is not in general = 0 is the first assertion, and the conditions (a) and (b) provide sufficient conditions to ensure $\gamma_2 = 0$, in which case $\beta = 0$ and $\beta + \gamma_2 = 0$ are in concordance, being true or false together. Condition (a) implies $\gamma = 0$ (by the same argument as established the equivalence of $\beta = 0$ and $Y \stackrel{t}{\vdash} X$), à fortiori $\gamma_2 = 0$. To show that condition (b) implies $\gamma_2 = 0$, we write $\gamma_1 + \gamma_2 = (\gamma | H_X(t-1) \oplus \{H_{X,Y}(t-1) \ominus H_X(t-1)\}) = \gamma_1 + (\gamma | S_4)$ using a previous definition and uniqueness. Subtracting yields $\gamma_2 = (\gamma | S_4)$. But $\gamma \in H_Z(t-1)$ which is orthogonal to S_4 , so $\gamma_2 = 0$. Q.E.D.

The question of the interpretation of these results is a difficult one for several reasons. Since the meaning of

causal orderings within a given system is itself an unsettled subject ([25] may be profitably consulted in this regard), there appears little hope for a universal explanation here. To make matters worse, issues of scientific method and philosophy inevitably enter any such discussion. With great trepidation, then, we contend that an investigator reporting the acceptance of bivariate exogeneity as a null hypothesis has found an "unusual" occurrence which requires an explanation. To be satisfactory, the explanation must show how a "believable" model, usually of dimension higher than two, could generate the observed ordering. Our proposition gives two such mechanisms, both of which are considerably weaker than the assumption that the omitted variables Z are orthogonal to the $\begin{pmatrix} X \\ Y \end{pmatrix}$ system. If neither condition is found acceptable, the mechanism must be more sophisticated, and the investigator must return to the drawing board in search of some other explanation. Bivariate causal orderings imply severe limitations on the higher order system that could have generated it, notwithstanding the many-one nature of the mapping between these systems. Our proposition offers some, per force incomplete, understanding of these restrictions.

Needless to say, applications will specialize the setting to stationary processes where truncated distributed lags represent the projections. Condition (a) is then self-explanatory; condition (b) requires that the information in the omitted variables' histories be different from the

information in Y^S but not X^S history.

Examples in which $Y \overset{t}{\vdash} X$, yet $Y \overset{b}{\vdash} X$ are easily constructed in the following manner. Adopt the normalization III,L,D discussed earlier in this section, and take $b_1(\cdot) \equiv 0$, so that $Y \overset{t}{\vdash} X$. Now take $c_1(\cdot) \neq 0$ and let the trivariate system be first order autoregressive. Use the result $(Y(t)|H_X(-\infty, \infty)) = R_{XX}^{-1} * R_{YX} * X(t) \equiv \mu * X(t)$ and compute a few coefficients of $\mu(\cdot)$, say $\mu(-1)$, $\mu(0)$, $\mu(1)$ for trivial choices of the other coefficients in the system. It would take incredible ingenuity to produce a one-sided $\mu(\cdot)$; any two-sided $\mu(\cdot)$ provides the desired example.

On the other hand, examples of models in which $Y \overset{b}{\vdash} X$, yet $Y \overset{t}{\not\vdash} X$, are hard to generate by adopting an autoregressive representation and parameterization for the trivariate system.

VII. Further Remarks on Applications in Economics

The interpretations of findings of exogeneity in economic data is a delicate and unsettled matter, even at the theoretical level, as recent contributions by Sargent [18], [19], and Sims [26] show. At the very least, owing to the sheer unlikelihood that two economic time series stand in a unidirectional causal relationship, such phenomena represent facts for theory to explain.

More fundamentally, however, the notion of a "structural relation invariant to manipulation of controlled processes which enter it" or "an intervention into the system," which

represent causality in the everyday usage of this term, must be distinguished from causality in the Wiener-Granger-Sims sense. That the two concepts are logically distinct is an important message of [18], in which money creation causes hyperinflation in the "intervention" sense, yet hyperinflation causes money in the sense of this paper; examples in which equations other than the structure show the causal ordering may be found in [26].

Nevertheless, in an important class of cases there may be not only consistency, but a mutual reenforcement, as the following interpretation of the money-income example shows. Suppose that money causes income, but not conversely (as found in [24]). Let $y = a * y + b * m + u$ and $m = c * m + v$ represent the model and exhibit the causal ordering ($m \rightarrow y$, $y \dashv m$). The "intervention" sense of causality refers to a stable relation involving y and m which allows the computation of y whenever an m process is inserted in it. Provided the coefficients a and b are invariant to changes in the m process, the first equation will be such a structural relation, which will yield $y = (1-a)^{-1} * b * m + (1-a)^{-1} * u$. While both variants of causality are present, there are two caveats. First, the empirical finding of causality during a sample is no guarantee of the invariance of a and b to changes in regime, as the "rational expectations" literature has emphasized.

Second, if the second relation were replaced by $m = \bar{c} * m + \bar{d} * y + v$ but the first relation remained invariant to "interventions" which violate the second equation and determine m , then again the concordance is spoiled, since only the causality in the "intervention" sense would be present.

From another point of view, to the extent that the results presented here involve innovations and optimal prediction (a form of optimizing behavior), they are likely to find use in, and enter structurally into, any theories in economics where stochastic elements enter in an essential way. Since the Hilbert spaces projected onto have the natural interpretation of information set, the possibilities for applications are virtually unlimited.

Finally, from an econometric point of view and as emphasized originally in [24], efficient estimation techniques (which are asymptotically the equivalent of generalized least squares) for a regression of $Y(t)$ on $X(t)$, $X(t-1)$, ... require exogeneity of X precisely in the sense of this paper. Thus, the propositions here may be of interest solely on econometric grounds.

VIII. Conclusions and Comments on Future Research

Since the introduction offers a summary statement as well, we confine ourselves here to a very brief paraphrasing of the results.

First, a minor generalization of Granger's first causality result is given in Proposition 1. The technique of proof is one which naturally allows the treatment of the more general cases of statistical causality in multivariate time series, the subject of Proposition 6.

Two characterizations of exogeneity in bivariate, or block-bivariate, systems are given next. In Proposition 2, it is demonstrated that the exogeneity of X with respect to Y is equivalent to the statement that future X be of no additional help in predicting current Y, given only current and past X. Despite its statement in prediction language, which brings to mind the original causality definition, a special case of this result yields Sims' important Theorem 2. Proposition 3 presents a characterization for X being exogenous in terms of univariate innovations of the X and Y processes; such a statement contrasts markedly with the previous results, which all stress bivariate characteristics. This result states that Y does not cause X precisely when past innovations of Y are all orthogonal to current, and by stationarity, all future X innovations. The relation of this result to the unpublished work of others is commented upon.

Proposition 4 is in the nature of a meta-theorem; it asserts that, when the definitions are altered so as to effect a time reversal (we backcast the present from the future) all existing theorems have natural analogues. A sample proof is

provided for the reversed version of Sims' Theorem 1. What is not symmetric, however, is the property of exogeneity itself; specifically, Proposition 5 shows that, only in the special case where the regression of Y on current, past, and future X has all coefficients, except possibly contemporaneous X, vanishing will X be exogenous according to both definitions.

The next result gives a new characterization of trivariate exogeneity of X with respect to Y given Z, in terms of projections of Y onto the past and future of X and Z. The form of the result provides a ready statistical test. Finally, Proposition 8 offers conditions in the presence of which bivariate and trivariate exogeneity events are logically equivalent. This result may provide explanations of bivariate findings in terms of the multivariate models thought more apt to characterize the process under study. Unfortunately, not all models will meet the assumptions we found necessary to obtain this result.

Despite the fact that Proposition 3 puts the characterization of the exogeneity of X, in terms of its own and Y's own innovations, on the same theoretical (Hilbert space) underpinning as the Sims and Granger results, it is a cursed result. As several writers have noted ([25] is most forceful), the natural estimation procedure which it suggests does not have the asymptotic validity of the other two tests. Whether this problem is amenable to correction by some fancy

footwork with distribution theory (thereby validating the procedure used in [5], [9], and [15], to name just three adherents of the "prewhitening school") or whether the difficulty is more deepseated remains an issue on which present opinion is divided.

At the theoretical (population) level, it may be too early to assert with any confidence that all of the interesting characterizations of exogeneity have been discovered. More important, however, will be a better understanding of economic mechanisms which give rise to causal orderings.

Footnotes

¹The symbol \equiv will frequently be used to indicate that the object on the left-hand side is being defined.

²These abbreviations (here, w.s.s.p.) which follow technical definitions will often be used to retain precision and economize on space in the sequel.

³When the range of Z_1 and Z_2 is complex, it is necessary to take the complex conjugate of Z_2 , as the notation indicates. Even though we deal with real processes, their representation in the frequency domain requires this treatment. Since there is essentially no use of the frequency domain in this paper, we will hereafter suppress the conjugation notation.

⁴Because its usage is not uniform, we emphasize that all subspaces for us will be closed (equivalently complete, because a Hilbert space is a complete metric space; [13], p. 116 proves this equivalence).

⁵The set S may be taken as a mnemonic for the span of the elements in the parentheses, or the linear manifold generated by them. Its closure may also be shown to be the intersection of all subspaces containing the generators.

⁶To define orthogonal projection, several related concepts are needed. The first is finite direct sum:

$$X = M_1 + M_2 + \dots + M_n = \sum_{i=1}^n M_i$$

means that any $x \in X$ may be uniquely written

$$x = \sum_{i=1}^n x_i,$$

where $x_i \in M_i$. When the subspaces $M_i \perp M_j$, all $i \neq j$ (any element of one orthogonal to all elements of the other), the direct sum decomposition is said to be orthogonal, and is indicated

$$X = M_1 \oplus M_2 \oplus \dots \oplus M_n = \sum_{i=1}^n \oplus M_i.$$

Secondly, if H is any Hilbert space and M is any linear manifold,

the set $M^\perp \equiv \{x \in H : x \perp m, \text{ all } m \in M\}$ is a subspace; and if M is itself a subspace, it follows that $H = M \oplus M^\perp$. ([1], p. 172). Applying this last result,

$$H_{X,Y}(-\infty, \infty) = H_{X,Y}(t) \oplus [H_{X,Y}(t)]^\perp;$$

for $z \in H_{X,Y}(-\infty, \infty)$,

$$z = u + v, u \perp v, u \in H_{X,Y}(t), v \in [H_{X,Y}(t)]^\perp.$$

Now the projection operator which maps $H_{X,Y}(-\infty, \infty)$ onto the subspace $H_{X,Y}(t)$, $(\cdot | H_{X,Y}(t))$, is defined by $(z | H_{X,Y}(t)) = u$. The special cases of interest in the text involve $z = Z(t+1)$ and $z = Y(t)$. This operator enjoys many important properties: linearity, idempotency, continuity unit norm, self-adjointness, and positivity--none of which are heavily exploited in this paper. That the projection minimizes the mean-square error follows from 5.8(6) of [30].

⁷ $L: H_{X,Y}(-\infty, \infty) \rightarrow H_{X,Y}(-\infty, \infty)$ is defined by $L[X(t)] = X(t+1)$ and $L[Y(t)] = Y(t+1)$, all t , and extended by continuity ([16], p. 14, 15). L is useful not only in proving plausible implications of stationarity, but also in deriving spectral properties of the process, which flow from the spectral properties of a unitary family of operators (cf. Stone's Theorem in [16], Section 137). $\{L^t\}$, $t \in I$ is such a family, when L^t is defined as the composition of L with itself t times.

⁸ The notation is so natural that it should cause no confusion, although strictly speaking we should write

$$\begin{pmatrix} u \\ w \end{pmatrix}(t) \perp H_{X,Y}(t-1) \times H_{X,Y}(t-1)$$

and proceed to extend all concepts to product spaces like these. Such extensions are helpful where extensive proofs are involved, as the Wiener-Masani [30] article demonstrates. Of course, by saying that a vector $\begin{pmatrix} u \\ w \end{pmatrix}(t)$ is \perp to a subspace, we mean that each component of the vector is orthogonal to all elements of the subspace.

⁹ This definition and the assertion of the multidimensional Wold decomposition theorem appears due to Zasuhiin [33] who announced the result without proof ([30], p. 136). The first

proof of the full rank case is Doob's ([2], p. 597); the general rank case was treated by Wiener and Masani [30]. Another definition of the rank of a process as the a.e. (Lebesgue) rank of its spectral density function ([17], p. 39) is present in the literature but needn't concern us in this paper.

¹⁰The testable restrictions which must be in the data for this treatment not to be rejected are quite--perhaps too--severe. For example, $X(t)$ must be perfectly predictable from (current) $Y(t)$ and the joint past. Of course, the finite length of real world data series compromises any strict test, but in principle the same objection applies to all theory.

¹¹These claims may be made good by use of the uniqueness of the orthogonal complement and 6.10(a) and 5.11(b) of [30].

¹²This is 6.10(b) of [30].

¹³6.13(b) of [30].

¹⁴It might be argued that in most applications there is no perfectly predictable component in the original series, again leaving us with a linearly regular process to analyze.

¹⁵In this case, $\Gamma_{u,v}(s) = \Sigma \cdot \delta(s)$, where $\delta(s) = \begin{cases} 1 & s=0 \\ 0 & s \neq 0 \end{cases}$ is the convolution identity ($A * \delta = \delta * A = A$ for all sequences $A(\cdot)$). Thus, $A * \Gamma_{u,v} * A'(k) = A * \Sigma \cdot \delta * A'(k) = A\Sigma * A'(k)$.

¹⁶Thus, one sees expressions like $X(t) = \sum_{j=0}^{\infty} a(j)e^{(t-j)}$ where $\sum_{j=0}^{\infty} |a(j)| < \infty$ is imposed, and vague, unmotivated references to "invertibility" made. First, no covariance stationary process with a discontinuous spectrum can be so represented, so the first assumption is overly strong. Second, no "invertibility," even with a finite order m.a.r., is necessary for $H_X(t) = H_e(t)$, although if the process were "invertible," the desired result would follow immediately from stationarity considerations.

¹⁷The author is not aware of necessary and sufficient conditions for a process to have an a.r. A very natural condition on the spectral density matrix, that there exist $0 < c_1 < c_2 < \infty$ such that $c_1 I < F'(\lambda) < c_2 I$ where $F'(\lambda)$ is the spectral density matrix of the $\begin{pmatrix} X \\ Y \end{pmatrix}$ process, has been used in [17], [30], [27] to arrive at an a.r. This condition is not, however, necessary for an a.r. Like the outright

assumption of existence of an a.r., it is in the nature of a regularity condition which, depending on one's axiomatic point of view, may be preferable.

The fact is, moreover, that this boundedness condition on the spectral density also guarantees that the process has an a.r., and more: the set $\{X(t), Y(t), t \in I\}$ forms a "basis" for $H_{X,Y}(-\infty, \infty)$, so that all elements in $H_{X,Y}(-\infty, \infty)$, not just projections, may be expressed as convergent infinite linear combinations.

18. The result referred to in the text reads: for any real square matrix A there exists a real, orthogonal P such that $PA = T$, where T is upper (real) triangular, with diagonal elements nonnegative. The desired application follows, upon transposition, for the II-L normalization; an analogous theorem for T lower triangular could be proved (by induction) and transposition would again give the II-U normalization; uniqueness is immediate.

19. This fact is the crux of the statement that a Wold causal-chain simultaneous equations model is exactly identified by its requirement of lower triangularity (which embodies the direction of causality in the chain) and diagonal covariance matrix. The situation must be carefully distinguished from a lower triangular Wold decomposition in a time series, which, if imposed, would be a vastly overidentifying restriction.

20. Since $\sigma_v^2 = \sigma_w^2 - \frac{\sigma_{uw}^2}{\sigma_u^2}$ and recalling the definition of

$\begin{pmatrix} u \\ v \end{pmatrix}$, we have

$$\begin{pmatrix} X \\ Y \end{pmatrix}(t) = A \begin{pmatrix} \sigma_u & 0 \\ \frac{\sigma_{uw}}{\sigma_u} & \sigma_w^2 - \frac{\sigma_{uw}^2}{\sigma_u^2} \\ \frac{\sigma_{uw}}{\sigma_u} & \frac{\sigma_{uw}}{\sigma_u} \end{pmatrix} * \begin{pmatrix} \frac{u}{\sigma_u} \\ w - \frac{\sigma_{uw}}{\sigma_u^2} u \\ \frac{\sigma_{uw}}{\sigma_u^2} \\ \frac{\sigma_w^2 - \frac{\sigma_{uw}^2}{\sigma_u^2}}{\sigma_w^2 - \frac{\sigma_{uw}^2}{\sigma_u^2}} \\ \frac{\sigma_{uw}}{\sigma_u^2} \end{pmatrix} (t) \equiv A * \Sigma^{1/2} Q \begin{pmatrix} \bar{e} \\ f \end{pmatrix} (t),$$

in the terminology of p. 18. We may verify directly that the proposed candidate for $\Sigma^{1/2} Q$ works and that $\begin{pmatrix} \bar{e} \\ f \end{pmatrix}$ has covariance matrix the identity.

²¹This follows by standard manipulation of the inner product. $X(t) - (X(t)|\overline{N_1}) \perp \overline{N_1}$ by the characterization of $(X(t)|\overline{N_1})$. But $(X(t)|\overline{N_1})$ is in $\overline{N_1}$, so that $\langle X(t), (X(t)|\overline{N_1}) \rangle - \langle (X(t)|\overline{N_1}), (X(t)|\overline{N_1}) \rangle = 0$, and the first term is zero since $X(t) \perp \overline{N_1}$ by assumption. This leaves $\| (X(t)|\overline{N_1}) \|^2 = 0$, so that $(X(t)|\overline{N_1}) = 0$.

²²Actually, Sims modestly proves a little more than he states. He proves the "only if" part, that $Y(t) = h*X(t)+W(t) \Rightarrow Y$ does not cause X , without the assumption that $\begin{pmatrix} X \\ Y \end{pmatrix}$ has an autoregressive representation. This, of course, is what our Corollary 2 gave. So, we only have a strengthening in the "if" direction, if this change in Sims' statement of Theorem 2 is made.

²³The author wishes to acknowledge his gratitude to Christopher A. Sims not only for suggesting the pursuit of this projection, but more generally for stressing the importance of fundamentalness.

References

- [1] Bachman, G. and Narici, L., Functional Analysis, Academic Press, New York, 1966.
- [2] Doob, J. L., Stochastic Processes, New York, 1953.
- [3] Geweke, J. F., "Wage and Price Dynamics in U.S. Manufacturing," mimeo. University of Wisconsin-Madison, 1975.
- [4] Granger, C. W. J., "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," Econometrica, July 1969, 37, p. 424-38.
- [5] Granger, C. W. J. and Newbold, D., "The Time Series Approach to Econometric Model Building," mimeo, 1975, presented at conference at the Minneapolis Federal Reserve Bank, November 13-14, 1975.
- [6] Graybill, F. A., Introduction to Matrices With Applications in Statistics, Wadsworth Publishing Co., Belmont, California, 1969.
- [7] Hannan, E. J., Multiple Time Series Analysis, Wiley, New York, 1970.
- [8] Haugh, L. D., "The Identification of Time Series Interrelationships with Special Reference to Dynamic Regression," Ph.D. Dissertation, Department of Statistics, University of Wisconsin (unpublished), 1972.
- [9] Haugh, L. D. and Box, G. E. P., "Identification of Dynamic Regression (Distributed Lag) Models Connecting Two Time Series," Technical Report #74, Department of Statistics, University of Florida (1974).
- [10] Hewitt, E. and Stromberg, K., Real and Abstract Analysis, Springer-Verlag, New York, 1969.
- [11] Kolmogorov, A., "Stationary Sequences in Hilbert Space" (Russian) Bull. Math. University Moscow 2, No. 6 (1941) 40 pp. (English translation by Natasha Artin).

- [12] Malinvaud, E., Statistical Methods of Econometrics, Second Revised Edition, North Holland, New York, 1970.
- [13] Naylor, A. and Sell, E., Linear Operator Theory in Engineering and Science, Holt, Rinehart, and Winston, Inc., New York, 1971.
- [14] Nerlove, M., "Distributed Lags and Unobserved Components in Economic Time Series," in Ten Economic Studies in the Tradition of Irving Fisher, John Wiley and Sons, Inc., New York, 1967.
- [15] Pierce, D. A. and Haugh, L. D., "The Assessment and Detection of Causality in Temporal Systems," 1975, mimeo.
- [16] Riesz, F. and Sz.-Nagy, B., Functional Analysis, Ungar, New York, 1955.
- [17] Rozanov, Yu. A., Stationary Random Processes, Holden-Day, San Francisco 1967 (translated by A. Feinstein).
- [18] Sargent, T. J., "The Demand for Money During Hperinflations Under Rational Expectations," manuscript, 1975.
- [19] Sargent, T. J., "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," forthcoming in Journal of Political Economy, (1976).
- [20] Sargent, T. J., "A Classical Macroeconometric Model for the United States," forthcoming in Journal of Political Economy (1976).
- [21] Sargent, T. J., "Rational Expectations, the Real Rate of Interest and the Natural Rate of Unemployment," Brookings Papers on Economic Activity, 1973.
- [22] Sargent, T. J. and Sims, C. A., "Business Cycle Modeling Without Much A Priori Economic Theory," mimeo, 1975, presented at conference at the Minneapolis Federal Reserve Bank, November 13-14, 1975.
- [23] Sargent, T. J. and Wallace, N., "'Rational Expectations, the Optimal Monetary Instrument, and The Optimal Money Supply Rule," Journal of Political Economy, 83, 2; April 1975, pp. 241-254.
- [24] Sims, C. A., "Money, Income, and Causality," The American Economic Review, Vol. 62, No. 4, September, 1972, pp. 540-552.

- [25] Sims, C. A., "Exogeneity Tests and Multivariate Time Series: Part I," mimeo, 1975.
- [26] Sims, C. A., "Exogeneity and Causal Ordering in Macroeconomic Models," mimeo, 1975, presented at conference at the Minneapolis Federal Reserve Bank, November 13-14, 1975.
- [27] Skoog, G., "Systematically Missing Data in Econometric Models: Some Identification Considerations," Staff Report 13, Federal Reserve Bank of Minneapolis, November 1976.
- [28] Whittle, P., Prediction and Regulation, Van-Nostrand-Reinhold, Princeton, New Jersey, 1963.
- [29] Wiener, N., "The Theory of Prediction," in Modern Mathematics for Engineers, Series 1 (edited by E. F. Beckenback) Ch. 8, 1956.
- [30] Wiener, N. and Masani, "The Prediction Theory of Multivariate Stochastic Processes I and II, Acta. Math., 98 and 99; 111-150 and 93-137; 1958 and 1959.
- [31] Wilson, G. T., "The Estimation of Parameters in Multivariate Time Series Models," Journal of the Royal Statistical Society, Series B, pp. 76-85, 1973.
- [32] Wold, H. O., A Study in the Analysis of Stationary Time Series, Almqvist and Wicksell, Uppsala, 1938 (2nd ed. 1954).
- [33] Zasuhin, V., "On the Theory of Multidimensional Stationary Random Processes" (Russian) C. R. (Doklady) Acad. Sci. U.R.S.S. 33 (1941) p. 435.