Historical Copy

Causality Characterizations: Bivariate, Trivariate, and Multivariate Propositions

Gary R. Skoog

May 1975
Revised January 1976

Working Paper #: 57

Rsch. File #: 266

Causality Characterizations: Bivariate, Trivariate, and Multivariate Propositions

by

Gary R. Skoog

Department of Economics*
University of Minnesota

May 1975
Revised January 1976**

*Research costs for this paper were partially borne by the Federal Reserve Bank of Minneapolis, which, of course, retains the right to form its own opinion on the matters discussed here. Discussions about Section IV with Thomas Sargent were especially helpful.

**Because a final revision is expected, please seek the author's permission before citation.

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 3 strengthen and slightly 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 Proposition 2 was proved when the author was unaware of three unpublished works (Haugh [8], Haugh and Box [9], and Pierce and Haugh [15]).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 scrutinization. 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 exogenous. Is it now true that, again trying to predict the current X but now given <u>future</u> X, that <u>future</u> 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. Proposition 7 closely examines the relationship between the events "X is exogenous with respect to Y in the trivariate system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$" and "X is exogenous with respect to Y in the implied bivariate $\binom{X}{Y}$ system." Using previous propositions, the answer is indicated on a case-by-case basis, 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 Theil-type omitted variables theorem. The role of the assumptions concerning instantaneous causality in this result is also investigated. 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 tact as an expository device here. Since the major complications introduced by the general q-variate mathematical theory are already present when q=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 $\mathcal{F}$ and probability measure P is defined a vector family of random variables (measurable functions), indexed by the discrete parameter t, $t \varepsilon I \equiv$ integers, $\binom{X}{Y}(t) \equiv \binom{X(t)}{Y(t)}$,[1/] which is the subject of our study. Following tradition, we have already suppressed the dependence of $\binom{X}{Y}$ on $\omega \varepsilon \Omega$: $\binom{X}{Y}(t,\omega)$ might have appeared more appropriate. Our notation reflects the fact that we will never investigate the behavior of sample paths (a sequence $\{\binom{X}{Y}(t,\bar{\omega})$, t = ... -1, 0, 1, ... for fixed $\bar{\omega}\}$) in the sequel, so there is no need to keep track of a second argument.

We do require that $\binom{X}{Y}(t)$ be a weakly stationary stochastic process (w.s.s.p.),[2] which means: (i) $E\binom{X}{Y}(t) = \binom{0}{0}$, all $t\varepsilon I$; and (ii) the Gramian or autocovariance matrix

$$\left(\binom{X}{Y}(t),\ \binom{X}{Y}(t-k)\right) \equiv \Gamma_{X,Y}(t,k) \equiv E\begin{pmatrix} X(t)X(t-k) & X(t)Y(t-k) \\ Y(t)X(t-k) & Y(t)Y(t-k) \end{pmatrix} \equiv \begin{pmatrix} R_X(k) & R_Y(k) \\ R_{YY}(k) & R_{YY}(k) \end{pmatrix}$$

does not depend on t, and so may be written $\Gamma_{X,Y}(k)$ ($= \Gamma^T_{X,Y}(-k)$, where ^{T} denotes transpose). Here E denotes mathematical expectation, so that one effect of (ii) is that X(t) and Y(t) must be in $L_2(\Omega, F, P)$, the space of all random variables Z(.) such that $\int |Z(\omega)|^2 dP(\omega) < \infty$. This latter space is a Hilbert space, H, with the inner product given by $\langle Z_1(\omega), Z_2(\omega)\rangle \equiv \int Z_1(\omega)\ \overline{Z_2(\omega)}\ dP(\omega)$[3] and norm $||Z_1|| \equiv \left(\int |Z_1(\omega)|^2\ dP(\omega)\right)^{1/2}$; 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 subspace[4] (closed linear manifold) of H, the space of values of the process $\binom{X}{Y}$, to be denoted $H_{X,Y}$ or $H_{X,Y}(-\infty, \infty)$. For any sets of integers $s_1, \ldots s_m$; $t_1, \ldots t_n$ and any sets of real or complex constants $a_1, \ldots a_m$; $b_1, \ldots b_n$ 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 $(\cup_{i\varepsilon I} X(i)) \cup (\cup_{i\varepsilon I} Y(i))$ or $S(X(i), Y(j), i, j\varepsilon I)$;[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, $D(t) \equiv \{\binom{X}{Y}(s), s=t, t-1, \ldots\}$ generated by the w.s.s.p. $\binom{X}{Y}$, a series sufficiently representative to yield perfect knowledge of the covariance sequence $\{\Gamma_s, s = \ldots -1, 0, 1, \ldots\}$. 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 will 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 projecting[6] 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 $\binom{u}{w}(t)$ form a new s.p., the innovations process (i.p.), corresponding to the original $\binom{X}{Y}(t)$ 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 lag operator.[7] More evident is the uncorrelatedness of $\binom{u}{w}(t)$ over time. Since $\binom{u}{w}(t) \perp H_{X,Y}(t-1)$[8] and $\binom{u}{w}(t-k) \varepsilon H_{X,Y}(t-k) \subseteq H_{X,Y}(t-1)$, $\binom{u}{w}(t) \perp \binom{u}{w}(t-k)$ (by which is meant that the autocovariance matrix formed from the two vectors,

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

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, $\Gamma_{u,w}(k) = \sum \cdot \delta_{0,k}$ where $\delta_{0,k} \equiv \begin{cases} 1 & k=0 \\ 0 & k\neq 0 \end{cases}$, the process $\binom{u}{w}(t)$ 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 $\sum$, $\rho(\sum)$, is known as the rank[9] of the $\binom{X}{Y}$ process and indicates an important structural characteristic of the system. Some taxonomy regarding system rank follows: (a) $\binom{X}{Y}$ may be perfectly predicted from its past only if $\binom{u}{w}(t) = \binom{0}{0}$, all t; in this case $\sum$ is the null matrix, $\rho(\sum) = 0$, and the process is said to be deterministic. (b) $\rho(\sum) \geq 1$, the process is nondeterministic (n.d.): it possesses at least one "component" which cannot be perfectly predicted from the past. The subcases are: (i) $\rho(\sum) = 1 < 2$, a degenerate case in which the

bivariate shock $\binom{u}{w}(t)$ 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 physical 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 $H_{u,w}(t) \equiv S(u(s), w(s), s \leq t) = \sum_{s=-\infty}^{s=t} \oplus S(u(s), w(s)) = \sum_{s=-\infty} \oplus D_{u,w}(s)$.[11/] Hence, $D_{u,w}(t) = D_{X,Y}(t)$, $t \in I$. 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 $\binom{X}{Y}$ 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) $\binom{X}{Y}(t) = \binom{X}{Y}_{l.r.}(t) + \binom{X}{Y}_{d.}(t)$, where

$$\binom{X}{Y}_{l.r.}(t) \equiv \left(\binom{X}{Y}(t) \middle| H_{u,w}(t)\right) \perp \binom{X}{Y}_{d.}(t) \equiv \left(\binom{X}{Y}(t) \middle| H_{X,Y}(-\infty)\right)$$

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

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

where $\left|\left|X_{l.r.}(t)\right|\right|^2 + \left|\left|Y_{l.r.}(t)\right|\right|^2 = \text{trace} \sum_{k=0}^{\infty} A(k)\, \Sigma A'(-k) \equiv \text{tr}\, A\Sigma * A'(0) < \infty$

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

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

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. $\binom{X}{Y}(t)$ is said to be linearly regular (purely nondeterministic is also used) if $\left(\binom{X}{Y}(t) \middle| H_{X,Y}(s)\right) \to 0$ as $s \to -\infty$; 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 regularity[13] 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 $\binom{X}{Y}(t)$ and assume that the process we are analyzing is linearly regular.[14/] This assumption will be maintained throughout the remainder of the paper: $\binom{X}{Y}(t)$ 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 succintly 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 $\Gamma_{X,Y}(.)$ of $\binom{X}{Y} = A * \binom{u}{v}$ may easily be expressed in terms of the autocovariance sequence $\Gamma_{u,v}(.)$ 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(.) and B(.) 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 $A\sum * A'(k)$[15/] or $A * \sum A'(k)$; these last formulae suggest the desirability of a representation in which $\sum = I$ so that $R_{X,Y} = A * A'(k)$. This may be done by tucking $\sum^{1/2} . \sum^{-1/2}$ into the convolution, to arrive at $\binom{X}{Y}(t) = A\sum^{1/2} * \sum^{-1/2}\binom{u}{w}(t) \equiv B * \binom{e}{n}(t)$, say (see p. 15 for an elaboration of this procedure). In the new representation $R_{X,Y}(k) = B * B'(k)$, since

Cov $\sum^{-1/2}\binom{u}{w}(t) \equiv E\binom{e}{n}(t)\binom{e}{n}^T(t) = I$; finite variances of X(t) and Y(t) becomes 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

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

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

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; analagous 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: $R_{X,Y}(z) \equiv \sum_{k=-\infty}^{\infty} (\binom{X}{Y}(t), \binom{X}{Y}(t-k))z^k = B(z)\, B^T(z^{-1})$, 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 the analytic $B(z)$ matrices has only 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 $B(.)$ is found which performs the factorization, there is the further requirement that, in $\binom{X}{Y}(t) = B * \binom{e}{n}(t)$, the $\binom{e}{n}$ 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 $\binom{X}{Y}$ and $\binom{e}{n}$ 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 $\binom{e}{n}(t)$) 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 $B(.)$ 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, $R_{X,Y}(s)$, is 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/]

Concepts closely allied to m.a.r. are those of autoregressive representation (a.r.)[17/] and extended autoregressive representation (e.a.r.); only the latter is new. When the l.r. w.s.s.p. $\binom{X}{Y}(t)$ with associated i.p. $\binom{e}{n}(t)$ permits the representation $B * \binom{X}{Y}(t) \equiv \sum_{s=0}^{\infty} B(s)\binom{X}{Y}(t-s) = \binom{e}{n}(t)$, with $B(0) = I$, where the sum converges 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 $(\binom{X}{Y}(t) | H_{X,Y}(t-1))$ has the convenient representation $-\sum_{s=1}^{\infty} B(s)\binom{X}{Y}(t-s)$ rather than the more generally necessary representation as the limit of a sequence of finite sums, with possibly changing weights. We say the process has an e.a.r. if, in addition, all of the projections into the subspaces $H_X(t-1)$ and $H_Y(t-1)$, for example, $(Y(t-i) | H_X(t-1))$ also have representations of the form $\sum_{j=1}^{\infty} h_i(j)\, X(t-j)$. Uniqueness of the various representations of this paragraph is apparently a necessary requirement to maintain consistency with the assumption that the process has full rank.

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: $\binom{X}{Y}(t) = \sum_{k=0}^{\infty} A_k \binom{u}{w}(t-k) \equiv A * \binom{u}{w}(t)$, $\left(\binom{u}{w}(t), \binom{u}{w}(t-k)\right)\Big\rangle^{-1} = A_k$ and $\Big\langle = \left(\binom{u}{w}(t), \binom{u}{w}(t)\right)$ (so that $A_0 = I$). The parameters A_k and $\Big\langle$ 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 $\binom{u}{w}(t)$ implies its subordination to $\binom{X}{Y}(t)$; thus $\binom{u}{w}(t)$ are mutually subordinate (m.s.): $H_{X,Y}(t) \equiv H_{u,v}(t)$.[17/] 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 $\binom{X}{Y}$. Other convolution representations, say $\binom{X}{Y} = \underline{A} * \binom{\underline{u}}{\underline{w}}$ which, like $\binom{u}{w}$, are vector white noise, also exist; as will be clearer from the ensuing paragraphs, there will be no difficulty in normalizing these $\binom{\underline{u}}{\underline{w}}$ so that their contemporaneous covariance matrix is the identity. Yet only for $\binom{\underline{u}}{\underline{w}} = T\binom{u}{w}$, $|T| \neq 0$ will $H_{\underline{u},\underline{w}}(t) \subseteq H_{X,Y}(t)$; those $\binom{\underline{u}}{\underline{w}}$ which are not subordinate to $\binom{X}{Y}$ yield <u>lower</u> prediction variances for $\binom{X}{Y}$. They are not suitable for prediction purposes because they carry information about the future of the $\binom{X}{Y}$ process.

For some purposes we may wish a representation $\binom{X}{Y}(t) = D * \binom{e}{n}(t)$, where the v.w.n. $\binom{e}{n}(t)$ is mutually subordinate to $\binom{X}{Y}(t)$ <u>and</u> $\left(\binom{e}{n}(t), \binom{e}{n}(t)\right) = I$. Rozanov terms $\binom{e}{n}$ 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. $\binom{X}{Y}(t) = A * \binom{u}{w}(t)$, $\mathrm{Cov}\binom{u}{w} = \sum$, we arrive at a f.p. by: diagonalizing $\sum$, $P\sum P' = \Lambda$; writing $\sum = P'\Lambda P = (P'\Lambda^{1/2}P)(P'\Lambda^{1/2}P) \equiv \sum^{1/2}\sum^{1/2}$, $\sum^{-1/2} \equiv P'\Lambda^{-1/2}P$, $\Lambda \equiv \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$ and $\Lambda^{1/2} \equiv \begin{pmatrix}\lambda_1^{1/2} & 0\\ 0 & \lambda_2^{1/2}\end{pmatrix}$; and finally taking $\binom{e}{n}(t) \equiv \sum^{-1/2}\binom{u}{w}(t)$. $\binom{e}{n}(t)$ has covariance matrix the identity, and $\binom{X}{Y}(t) = A\sum^{1/2} * \sum^{-1/2}\binom{u}{w}(t) \equiv A\sum^{1/2} * \binom{e}{n}(t)$ is a m.a.r. with $\binom{e}{n}(t)$ as f.p. Any Q, $Q'Q = I$ results in $\binom{X}{Y}(t) = A\sum^{1/2}Q' * Q\binom{e}{n}(t) \equiv A\sum^{1/2}Q' * \binom{\bar{e}}{\bar{n}}(t)$, so that $\binom{\bar{e}}{\bar{n}}(t) \equiv Q\binom{e}{n}(t)$ 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 convolution is $\sum^{1/2}Q'$ where $\sum^{1/2}$ is unique, but Q' may be any orthonormal matrix. It is a well-known result in matrix theory[18] ([6], p. 191-192) that any symmetric matrix ($\sum^{-1/2}$) may be lower (respectively, upper) tringularized 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.: $\binom{X}{Y}(t) = A * \binom{u}{w}(t)$, $\mathrm{Cov}\binom{u}{w}(t) = \begin{pmatrix}\sigma_u^2 & \sigma_{uw}\\ \sigma_{wu} & \sigma_w^2\end{pmatrix}$ and $A(0) = I$. We retain u(t), but replace w(t) in our basis

by $v(t)$, where $v(t) \equiv w(t) - (w(t)|u(t)) = w(t) - \frac{\sigma_{uw}}{\sigma_u^2} u(t)$. In other words, $\binom{u}{v}(t) = \begin{pmatrix} 1 & 0 \\ -\frac{\sigma_{uw}}{\sigma_u^2} & 1 \end{pmatrix} \binom{u}{w}(t)$, or $\binom{u}{w}(t) = \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u^2} & 1 \end{pmatrix} \binom{u}{v}(t)$.

Consequently, $\binom{X}{Y}(t) = A \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u^2} & 1 \end{pmatrix} * \binom{u}{v}(t)$ in which the m.s. process $\binom{u}{v}(t)$ 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 $A(k) \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u^2} & 1 \end{pmatrix}$, $k \varepsilon I$, 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: $A(0) \begin{pmatrix} 1 & 0 \\ \frac{\sigma_{uw}}{\sigma_u^2} & 1 \end{pmatrix} T^{-1}$, $T \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} T'$. 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 $\binom{u}{v}(t)$ so that they have unit variance. In other words, $\binom{X}{Y}(t) =$

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

or again $A\begin{pmatrix} \sigma_u & 0 \\ \frac{\sigma_{uw}}{\sigma_u} & \sigma_v \end{pmatrix} * \begin{pmatrix} \frac{u}{\sigma_u} \\ \frac{v}{\sigma_v} \end{pmatrix}$, 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 $II^{\underline{s}}$ 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, arbitrary 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. $\binom{X}{Y}(t)$. 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: $Y \overset{i}{\rightarrow} X$). In symbols, $Y \overset{i}{\rightarrow} X$ 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 $Y \overset{i}{\rightarrow} X$ if and only if $X \overset{i}{\rightarrow} Y$, the notion of Wiener-Granger causality does not permit any distinction as regards instantaneous causality; consequently, the definition is most meaningful only 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 $\binom{u}{w}(t)$ is fundamental--v.w.n. 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. $\binom{X}{Y}(t)$, Y does not cause X if and only if the Wold (m.a.) representation subject to II-L is $\binom{X}{Y}(t) = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} * \binom{u}{w}(t)$.

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 cannot resist supplying 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 $\underset{\sim}{a}(z) = g(z)\, a(z)$, with $g(z)\, g(z^{-1}) = 1$ on $|z| = 1$, such that $|\underset{\sim}{a}(0)|^2 > |a(0)|^2$ (here $a(z)$ is a scalar). Then if we consider a competing $\underset{\sim}{B}(z) \equiv \begin{pmatrix} \underset{\sim}{a}(z) & 0 \\ c(z)\, g(z) & d(z) \end{pmatrix}$, $\underset{\sim}{B}(z)$ still factors the spectral density and $\det|\underset{\sim}{B}(0)|^2 = |\underset{\sim}{a}(0)|^2 |d(0)|^2 > |a(0)|^2 |d(0)|^2$, contradicting the maximality of $B(z)$ and hence the joint fundamentalness of $\begin{pmatrix} u \\ w \end{pmatrix}$. It is indeed unfortunate that what is essentially a time domain argument does depend on frequency domain considerations at this stage, but we can find no alternative that is worth the work.

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

III. Bivariate Characterizations

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

Proposition 1. Assume that the 1.r.w.s.s.p. $\binom{X}{Y}$ has the extended autoregressive representation (e.a.r.) $\binom{X}{Y}(t) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} * \binom{X}{Y}(t) + \binom{u}{w}(t) \equiv \begin{pmatrix} \sum_{i=1}^{\infty} a(i)X(t-i) + \sum_{i=1}^{\infty} b(i)Y(t-i) \\ \sum_{i=1}^{\infty} c(i)X(t-i) + \sum_{i=1}^{\infty} d(i)Y(t-i) \end{pmatrix} + \binom{u}{w}(t)$ with $E\binom{u(t)}{w(t)}(u(t)w(t)) \equiv \begin{pmatrix} \sigma_u^2 & \sigma_{uw} \\ \sigma_{wu} & \sigma_w^2 \end{pmatrix} \equiv \Sigma$. Then: (i) Y does not cause X if and only if $b(.) \equiv 0$; (ii) whether or not $b(.) \equiv 0$, instantaneous feedback (or instantaneous causality) is present if and only if $\sigma_{uw} \neq 0$, this last result holding even if no e.a.r. exists, where then $\binom{u}{w}(t)$ is interpreted as the innovations process in the m.a.

Proof: (i) Assume first that $b(.) \equiv 0$. Then by the definition of the a.r. $\binom{u}{w}(t) \perp H_{(X,Y)}(t-1)$, 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 $b(.) \not\equiv 0$, we have a contradiction: $(Y(t-i)|H_X(t-1)) = \sum_{j=1}^{\infty} h_i(j)X(t-j)$,

say, by the e.a.r. assumption, yielding $\sum_{i=1}^{\infty} b(i)[Y(t-i) - \sum_{j=1}^{\infty} h_i(j)X(t-j)] = 0$ which contradicts the uniqueness of the a.r. since this term may be added to the R.H.S. of the a.r. with impunity. (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 $\langle u(t) - \frac{\sigma_{uw}}{\sigma_w^2} w(t), w(t)\rangle = 0$, 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 $(X(t)|w(t)) = (u(t)|w(t)) = \frac{\sigma_{uw}}{\sigma_w^2} w(t) + (v(t)|w(t)) = \frac{\sigma_{uw}}{\sigma_w^2} w(t)$; 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 $\frac{\sigma_{uw}}{\sigma_u^2} u(t)$ 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 clever 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. Since as footnote 17 has observed, the natural sufficient condition (and only one the author has seen enunciated) for the existence of the a.r. also implies the existence of the e.a.r., any weakness inflicted by the e.a.r. assumption is minimal. Further, 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 $b(.) \equiv 0$ 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, which was referred to earlier on p. 14 as well, may be found in the second of the Wiener-Masani references [30].

With the characterization of instantaneous causality in hand, let us momentarily return to the Sims theorem and observe that i.c. 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 $c(.)$, 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 $b(s) = 0$, $s=1, 2, \ldots$.

In a comparison of these theorems, the Sims result has the advantage of mathematical generality, in that its hypothesis are met for any l.r.w.s.s.p.; the Granger result, while requiring in addition an a.r. (or e.a.r.), yields an immediate statistical test. We refer interested parties 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.

Proposition 2. In the l.r.w.s.s.p. $\binom{X}{Y}(t)$, 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 equality and stationarity, we substitute, shift, and define to arrive at $n_j \equiv Y(t-j) - (Y(t-j)|H_X(t-j)) \perp X(t+k)$, $j=1, 2, \ldots$ 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

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

Since clearly $S(X(t-j), Y(t-j), j=1, 2, \ldots) = S(\bigcup_{j=1}^{\infty} X(t-j), N_1)$, the closures must be the same set, which by definition is $H_{X,Y}(t-1)$. Thus,

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

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, if Y does not cause X, we may take $b(.) \equiv 0$ in the Wold representation by Sims' Theorem 1. Consequently, $X(t) = a*e(t)$ and $Y(t) = c*e(t) + d*n(t)$; we have argued that $e(t)$ is univariate fundamental for $X(t)$, so that $H_X(t) = H_e(t)$. Now forming the two projections whose equality we seek to establish,

$$(Y(t)|H_X(-\infty, \infty)) = (c*e(t)+d*n(t)|H_e(-\infty, \infty)) = c*e(t)$$

and

$$(Y(t)|H_X(t)) = (c*e(t)+d*n(t)|H_e(t)) = c*e(t).$$

The LHS of the expressions above are thus equal. Q.E.D.

Corollary 1. If in the l.r.w.s.s.p. $\binom{X}{Y}$, 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 $(Y(t)|H_X(t)) \varepsilon H_X(t) = H_e(t)$ and $\{e(s)\}_{s=-\infty}^{s=t}$ is a complete orthonormal set, we have the Fourier representation

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

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 double[22/] strengthening (and alternate proof) of Sims' Theorem 2. Thus, the Sims test for exogeneity--testing whether "future" coefficients of h(.) vanish--is an implication of "Y does not cause X" under the milder assumption that only X (and not $\binom{X}{Y}$ jointly) possesses an a.r. On the other hand, the presence of a one-sided f(.), 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 available (and 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. $\binom{X}{Y}$, 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), ... are uncorrelated with e(t+1), e(t+2), ...; 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 proposition 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

(*) $Y(t) \perp e(t+j)$ all $j > 0 \iff Y(t) \perp D_e(t+j)$, all $j > 0 \iff$

$(Y(t)|D_e(t+j)) = 0$ all $j > 0 \iff ||(Y(t)|D_e(t+j))||^2 = 0$ all $j > 0 \iff$

$\sum_{s=t+1}^{\infty} ||Y(t)|D_e(s)||^2 = 0 \iff (Y(t)|H_X(t)) = (Y(t)|H_X) \iff$ Y does not

cause X, using Proposition 4 at the last step. By joint stationarity, (*) is equivalent to $Y(t-k) \perp e(t+j)$, all $j > 0$ and all $k \leq 0$. Since $H_v(t) = H_Y(t) = \sum_{s=-\infty}^{s=t} \oplus D_v(s)$ and $\sum_{s=t+1}^{\infty} \oplus D_e(s) = S(e(t+j),\ j > 0)$, 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, $(1-L)^{-1} = 1 + L + L^2 + \ldots$ results in the unpleasant $\lim_{n\to\infty} \sum_{i=0}^{n} X(t-i) = \lim_{n\to\infty} \{u(t) - u(t-n-1)\}$, 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 out of the blue, 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 $u(t) \varepsilon H_x(t)$. 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 $u(t) \varepsilon H_X(t)$. We take for $\hat{X}_n$ the projection of $X(t)$ onto $\langle X(t-1), X(t-2), \ldots X(t-n) \rangle$.[23/] Writing out the normal equations yields $\hat{X}_n = \sum_{i=1}^{n} c(i) X(t-i)$ where the $c(i)$ satisfy

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

Since A is symmetric, so is A^{-1}; the first row or column may be verified to be $\left(\frac{n}{n+1} \quad \frac{n-1}{n+1} \quad \ldots \quad \frac{1}{n+1} \right)$ which allows the determination of c. Hence, $\hat{X}_n = -\sum_{i=1}^{n} \left(\frac{n+1-i}{n+1} \right) X(t-i)$. Now $(n+1)[X(t) - \hat{X}_n] = (n+1)X(t) + nX(t-1) +$

$\ldots + X(t-n) = (n+1)u(t) - u(t-1) - u(t-2) - \ldots - u(t-n) - u(t-n-1)$.

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

$u(t)$ 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

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

with $d(0) = 1$ which converges. We use $(\dagger)$ in concluding. Because $\hat{X}_n + u(t) \to X(t)$ as $n \to \infty$, if an a.r. exists, say $X(t) = \sum_{i=1}^{\infty} a(i)X(t-i) + u(t)$, then $\lim_{n\to\infty} \sum_{i=1}^{n} a(i)X(t-i) = \lim_{n\to\infty} \hat{X}_n$, or $\lim_{n\to\infty} \sum_{i=1}^{n} (a(i) + \frac{n-i}{n+1})X(t-i) = 0(*)$.

If $a(i) \neq -1$ for some i, let i' be the first such i. We then have, after renormalizing via $d(i-i') \equiv (a(i) + \frac{n-i}{n+1})/(a(i') + \frac{n-i'}{n+1})$,

$\lim_{n\to\infty} \sum_{i=0}^{n} d(i)X(t-i'-i) \equiv d*X(t-i') = 0$, $d(0) = 1$. Stationarity implies $d*X(t) = 0$, contradicting $(*)$. 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: $u(t-1) \in \overline{\bigcup_{j=1}^{\infty} X(t-j)}$ 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

$$(Y(t)\,|\,H_X(t)) = (u(t) + v(t) - v(t-1)\,|\,H_X(t)) =$$

$$u(t) = (Y(t)\,|\,H_X(-\infty, \infty));$$

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

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

(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

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

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 $\lambda\varepsilon[0,2\pi]$ 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 $Y \dashv\!\!\rightarrow X$ 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, $S(X(t+i), Y(t+j), i,j=1,2,...)$, 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: $Y \overset{t.r.}{\dashv\!\!\rightarrow} X$) 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 smaller subspace; as before, time reversal exogeneity of X with respect to Y (synonymous with $Y \overset{t.r.}{\dashv\!\!\rightarrow} X$) 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 $\underline{H}_{X,Y}(t) = \sum_{s=t}^{\infty} \oplus \underline{D}_{X,Y}(s) \oplus H_{X,Y}(\infty)$, the latter term representing the infinite future, $\bigcap_{s=-\infty}^{\infty} \underline{H}_{X,Y}(s)$, 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, $H_{X,Y}(s, \infty)$, 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 $H_{X,Y}(-\infty)$ and $H_{X,Y}(\infty)$ 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. $\binom{X}{Y}(t)$ 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 $\underline{H}_{X,Y}(t, \infty) = \sum_{s=t}^{\infty} \oplus \underline{D}_{u,w}(s) \oplus H_{X,Y}(\infty)$ is available, a reversed version of a moving average representation falls out, just as before, by projecting $\binom{X}{Y}(t)$ 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. $\binom{X}{Y}$ has the moving average representation forward (m.a.r.f.)

$$\binom{X}{Y}(t) = \sum_{k=0}^{\infty} \underline{A}(k)\binom{u}{w}(t+k) \equiv \underline{A} * \binom{u}{w}(t), \text{ where } \underline{A}(0) = I, \text{ Cov } \binom{u}{w} = \underline{\Sigma},$$

$$\underline{A}(k) = \left(\binom{X}{Y}(0), \binom{u}{w}(k)\right)\underline{\Sigma}^{-1}, \text{ and trace } \sum_{k=0}^{\infty} \underline{A}(k)\underline{\Sigma}\,\underline{A}'(-k) < \infty.$$

$$\underline{H}_{X,Y}(t) = \underline{H}_{u,v}(t), \text{ so that } \binom{u}{v} \text{ is the innovations process.}$$

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

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

where again $\binom{u}{w}$ and $\binom{X}{Y}$ are mutually subordinate into the future. And again, projections are guaranteed to have convenient representations in terms of convergent infinite sums by the definition of an extended a.r.f. (e.a.r.f.).

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 $X(t) = b\underline{*}\,\underline{u}(t)$ (the same b(.) sequence) again represents X(t), because $R_X(.)$ is a symmetric function. Consequently, there is no nonzero element in $H_X(+\infty)$ either. And the maximality condition which b(.) is known to satisfy is precisely that which guarantees that $\underline{u}(t)$ is future fundamental. Thus, $X(t) - (X(t)|\underline{H}_X(t+1)) = b(0)\underline{u}(t)$, Var $b(0)\underline{u}(t) = |b(0)|^2 =$ 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 equal. Q.E.D.

We remark that the reason this result does not carry over to vector processes is because the matrix analogue of b(.), B(.), does not continue to factor $R_{X,Y}(.) = B*B'(.)$, 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 $|\sum|$, that is, $|\sum| = |\underline{\sum}|$: 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 $\binom{X}{Y}(t)$ be a l.r.f. w.s.s.p. Then X is time-reversed exogenous with respect to Y ($Y \overset{t.r.}{\dashv\!\rightarrow} X$) if and only if:

(i) The m.a.r.f. is given by $\binom{X}{Y}(t) = \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} * \binom{u}{w}(t)$, i.e., $\underline{b}(.) \equiv 0$; where the normalization is $\underline{a}(0) = \underline{d}(0) = 1$, $\underline{b}(0) = 0$; Cov $\binom{u}{w} = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix}$; (the analogue of III-L).

(ii) When an a.r.f. exists, it is given by $\begin{pmatrix} \alpha & 0 \\ \gamma & \delta \end{pmatrix} * \binom{X}{Y}(t) = \binom{u}{w}(t)$, where the normalization is as in (i), III-L.

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

(iv) $\underline{e}(t), \underline{e}(t-1), \ldots$ are uncorrelated with $\underline{v}(t+1), \underline{v}(t+2), \ldots$, where $Y(t) = \underline{\mu} * \underline{v}(t)$ and $X(t) = \underline{a} * \underline{e}(t)$, $\underline{\mu}(0) = \underline{a}(0) = 1$, are univariate m.a.r.f.'s.

Proof:

(i)<=: With the given m.a.r.f., X(t) lies in $\underline{H}_u(t)$. By definition of m.a.r.f., $\underline{H}_{X,Y}(t) = \underline{H}_{u,v}(t)$, and the earlier remarks in Section II show that $\underline{H}_X(t) = \underline{H}_u(t)$, or else, by an analagous maximality argument, the mutual subordination of the $\binom{X}{Y}$ and $\binom{u}{v}$ processes would be contradicted. Now forming the projection $(X(t)|\underline{H}_{X,Y}(t+1)) = \sum_{i=1}^{\infty} \underline{a}(t)\underline{u}(t+i))$, 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. =>: Assuming now the equality of these projections, and our definition of $Y \overset{t.r.}{\dashv\!\rightarrow} X$, we may now write

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

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

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, $(\underline{u}(s), \underline{v}(s), s=t,t+1,\ldots)$ form a complete orthonormal system, $\underline{H}_{X,Y}(t) = \underline{H}_{u,v}(t)$, 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 $Y \overset{t.r.}{\dashv\!\rightarrow} X$, then the analagous results hold, and conversely. However, it quickly becomes apparent that $Y \dashv\!\rightarrow X$ does not in general hold up under time reversal, except in the special case where $(Y(t) \,|\, H_X(-\infty, \infty))$ is in $\overline{S(X(t))}$, as we prove for totally regular processes in

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

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

where $a(0) = d(0) = 1$, $\mathrm{Cov}\binom{u}{w}(t) = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_w^2 \end{pmatrix} = \mathrm{Cov}\ \binom{u}{\underline{v}}(t)$, and $k=0$ if and only if $X \overset{i.c.}{\longrightarrow} Y$.

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 $Y \nrightarrow X$, 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 $Y \overset{t.r.}{\nrightarrow} X$ to hold, there must also be such a lower triangular representation on the future, $\begin{pmatrix} X \\ Y \end{pmatrix}(t) = \underline{\begin{pmatrix} a & 0 \\ e & f \end{pmatrix}} * \underline{\begin{pmatrix} u \\ w \end{pmatrix}}(t)$, where we have used a fact encountered in the lemma: the same a(.) must be present in both. But writing out the cross-autocovariance $R_{YX}(.)$ 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 $a^{-1}(.)$ 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 $c(.) = ka(.)$. 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 $(a*a')^{-1}*ka*a' = k.\delta(.)$. $\therefore e(.) = c(.)$, 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 $Y \nrightarrow X$ 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 $\underline{H}_{X,Y}(t+1)$ 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 <u>not</u> 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. $\underset{nx1}{X(t)} = \underset{nxn}{A} * \underset{nx1}{e(t)}$, $H_e(t) = H_X(t)$ for the m.a. and $B * X(t) = e(t)$, $H_e(t) = H_X(t)$ 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 bloc-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. 19. Since bloc-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 1x1 and X_2 is (n-1)x1, 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: $X_2 \not\mapsto x_1$). 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 $x_1 \not\mapsto X_2$.

The general notion, of course, is the relation "does not cause," indicated by $\not\mapsto$, which is defined as in the last sentence but with $\underset{n_1x1}{X_1}$ and $\underset{n_2x1}{X_2}$ representing different subvectors of $\underset{nx1}{X}$, where $n_1 + n_2 \leq n$. We understand this interpretation in the sequel. Since $X_1 \not\mapsto X_2$ means

that no component of $X_1^{n_1 x 1}$ helps predict any component of $X_2^{n_2 x 1}$, the symbol $\nrightarrow$ refers to $n_1.n_2$ elementary causality events of the form $x_i^{1x1} \nrightarrow x_j^{1x1}$. To characterize these latter events, the a.r. is the more convenient, as Proposition 6 (i) shows. However, in describing results involving one component, say x_1, and $(x_2, \ldots x_n)^T \equiv X_2$, the rest of the system, the a.r. and m.a. again have the same qualitative appearance, if a relation $\nrightarrow$ 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. $X(t) = A^{nxn} * e(t)^{nx1}$, e.a.r. $B^{nxn} * X(t) = e(t)^{nx1}$, and $X^T = (x_1^{1x1 T} \; X_2^{1x(n-1) T})$,

(i) $x_i \nrightarrow x_j$ if and only if $b_{ji}(.) \equiv 0$ in the e.a.r.

(ii) $X_2 \nrightarrow x_1$, or x_1 is exogenous, if and only if either of the equivalent conditions hold:

(a) $(b_{12}(.) \cdots b_{1n}(.)) \equiv 0$ in the e.a.r.

(b) $(a_{12}(.) \cdots a_{1n}(.)) \equiv 0$ in the m.a.

(iii) $x_1 \nrightarrow X_2$, or x_1 does not cause any other variable in the system, if and only if either of the equivalent conditions hold:

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

(iv) In the results (ii) and (iii), x_1^{1x1} may be replaced by $X_1^{n_1 x 1}$, $X_2^{(n-1)x1}$ by $X_2^{n_2 x 1}$, $(n_1+n_2 = n)$ and the conditions (a) and (b)

by the upper right and lower left matrices in the conformably partitioned e.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 e.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, q-variate prediction theory are present for q=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. It is hoped that the inelegance of the brute force method applied here will be at least partially offset by the usefulness of the results.

The first question addressed 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 system $\begin{pmatrix} X \\ Y \end{pmatrix}$, when will this implied bivariate system inherit the exogeneity of X? By giving a set of conditions which are necessary and sufficient, and by answering the same question when, in the trivariate system, Y causes X, an exhaustive classification of all possible interactions is achieved: this result may help the researcher who has some fragmentary knowledge of the direction of flows of causation between potentially omitted time series and the series he is modeling.

We will cast our main result in terms of a variant of the autoregressive representation, a normalization analagous to III-L in the

m.a.r. The reason is that, as Proposition 6 has shown, "zero restrictions" on these parameters are easiest to interpret directly as causality statements; an added advantage is that the a.r. is more familiar to econometricians than the m.a.r. Logically prior to the 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: $Z \overset{i_1}{\rightarrow} X$) 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

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

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

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

As in the bivariate case, both relations are symmetric (that is, $X \overset{i_1 \text{ (or } i_2)}{\longrightarrow} Z$ if and only if $Z \overset{i_1 \text{ (or } i_2)}{\longrightarrow} X$). Consequently, the notation $X \overset{i_1 \text{ (or } i_2)}{\leftrightarrow} Z$ is now adopted. And, as in the bivariate case, the covariances between X and Z provide a handy criterion:

$\sigma_{XZ} \neq 0 \Leftrightarrow X \overset{i_2}{\leftrightarrow} Z$ and $\sigma_{XZ.Y} \neq 0 \Leftrightarrow X \overset{i_1}{\leftrightarrow} Z$.

The technique of proof is the same as was used in Proposition 1, (ii), so the treatment here is terse. We take the e.a.r. to be

$B^*\begin{pmatrix}X\\Y\\Z\end{pmatrix}(t) = e(t)$, $B(0) = I$, Cov $e(t) = \Sigma$. Since the forecast error of $X(t)$ given the 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 $e_1(t) - (e_1(t)|\overline{e_2(t) \cup e_3(t)})$. Solving for the two indicated projections, $e_1(t) - \frac{\sigma_{12}}{\sigma_{22}} e_2(t)$ and $e_1(t) - (e_2(t)\ e_3(t)) \begin{pmatrix}\sigma_{22} & \sigma_{23}\\ \sigma_{32} & \sigma_{33}\end{pmatrix}^{-1} \begin{pmatrix}\sigma_{12}\\ \sigma_{13}\end{pmatrix} = e_1(t) - (e_2(t)\ e_3(t)) \frac{1}{\sigma_{22}\sigma_{33}-\sigma_{23}^2} \begin{pmatrix}\sigma_{33}\sigma_{12}-\sigma_{32}\sigma_{13}\\ -\sigma_{23}\sigma_{12}+\sigma_{22}\sigma_{13}\end{pmatrix}$. Thus, it is clear that, $Z \overset{i_1}{\leftrightarrow} X$ iff Cof $\sigma_{31} \equiv \begin{vmatrix}\sigma_{12} & \sigma_{13}\\ \sigma_{22} & \sigma_{23}\end{vmatrix} \neq 0$. Analagously, $X \overset{i_1}{\leftrightarrow} Z$ iff, since the projection is $e_1(t) - (e_1(t)\ e_2(t)) \frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}^2} \begin{pmatrix}\sigma_{22} & -\sigma_{12}\\ -\sigma_{21} & \sigma_{11}\end{pmatrix}\begin{pmatrix}\sigma_{31}\\ \sigma_{32}\end{pmatrix}$, Cof $\sigma_{13} \neq 0$.

The symmetry of Σ thus entails symmetry of $\overset{i_1}{\leftrightarrow}$. As in the bivariate case, $Z \overset{i_2}{\leftrightarrow} X$ if and only if $\sigma_{12} \neq 0$, as a computation above shows; symmetry of Σ thus extends to $\overset{i_2}{\leftrightarrow}$. Only in the case where $\sigma_{12}\ \sigma_{23} = 0$ will there necessarily be agreement between $\overset{i_1}{\leftrightarrow}$ and $\overset{i_2}{\leftrightarrow}$ for X and Z, although a.e. (lebesgue) in the space of positive definite Σ matrices it will be the case that $\overset{i_1}{\leftrightarrow} \iff \overset{i_2}{\leftrightarrow}$ .24/

We now proceed to derive a normalization of the a.r. (or e.a.r.), which is analagous to III-L,D. This particular normalization will be found convenient in the proof below; further, we sketch its derivation 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 e.a.r. 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 $\text{Cov}\begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}(t) = \sum$, 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 $\underline{e}_2(t) \perp \overline{H_{X,Y,Z}(t-1) \cup X(t) \cup Y(t)}$; 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:

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

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 - \frac{\sigma_{21}}{\sigma_{11}} a_1 & b_2 - \frac{\sigma_{21}}{\sigma_{11}} b_1 & c_2 - \frac{\sigma_{21}}{\sigma_{11}} c_1 \\ a_3 - r_1 a_1 - r_2 a_2 & b_3 - r_1 b_1 - r_2 b_2 & c_3 - r_1 c_1 - r_2 c_2 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) + \begin{pmatrix} \underline{e}_1(t) \\ \underline{e}_2(t) \\ \underline{e}_3(t) \end{pmatrix}$$

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

This is the parameterization we adopt; it seems reasonable to name it 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, ... 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,..., etc. No confusion will arise if we drop the _ and reuse the previous notation: $a_1, a_2, \ldots$ Hence, what is important to notice in our final representation,

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

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

is that $a_2(0) = \frac{\sigma_{21}}{\sigma_{11}}$, $a_3(0) = \frac{\mathrm{Cof}\ \sigma_{13}}{\mathrm{Cof}\ \sigma_{33}}$, $b_3(0) = \frac{-\mathrm{Cof}\ \sigma_{23}}{\mathrm{Cof}\ \sigma_{33}}$. Consequently, the assumption $a_2(.) \equiv 0$ rules out $X \overset{i_2}{\leftrightarrow} Y$, while the assumption $b_1(.) \equiv 0$ does not have any implication for instantaneous causality. Similar statements hold for $a_3(.) \equiv 0$ and $c_1(.) \equiv 0$ regarding $X \overset{i_1}{\leftrightarrow} Z$; and for $b_3(.) \equiv 0$ and $c_2(.) \equiv 0$ regarding $Y \overset{i_1}{\leftrightarrow} Z$.

We now begin a development which might be properly termed the beginning of the proof of the next proposition. The first observation is that the m.a.r. with normalization III-L is obtained when (6.4) is rewritten, with 1 denoting the identity convolution, as

$$-\begin{pmatrix} \alpha & b_1 & c_1 \\ a_2 & \beta & c_2 \\ a_3 & b_3 & \gamma \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) \equiv \begin{pmatrix} 1-a_1 & -b_1 & -c_1 \\ -a_2 & 1-b_2 & -c_2 \\ -a_2 & -b_3 & 1-c_3 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}(t)$$

and inverted (under convolution) to

$$(6.5)\qquad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) = -\Delta^{-1} * \begin{pmatrix} \beta\gamma - c_2 b_3 & c_1 b_3 - b_1\gamma & b_1 c_2 - c_1\beta \\ c_2 a_3 - a_2\gamma & \alpha\gamma - c_1 a_3 & c_1 a_2 - \alpha c_2 \\ a_2 b_3 - a_3\beta & b_1 a_3 - \alpha b_3 & \alpha\beta - b_1 a_2 \end{pmatrix} * \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}(t),$$

where Δ^{-1} is the inverse under convolution of Δ,

$$\Delta = \alpha\beta\gamma + b_1 c_2 a_3 + c_1 a_2 b_3 - c_1 \beta a_3 - \alpha c_2 b_3 - b_1 a_2 \gamma,$$

and in the last two expressions, convolutions have been juxtaposed without benefit of *, e.g., $\beta\gamma - c_2 b_3$ might have been tediously written $\beta * \gamma - c_2 * b_3$, etc. We emphasize that (6.5) in a variant of the (Wold) m.a.r. which always exists. Now the general formulae of Section II may be applied to find the own and cross autocovariance functions between X and Y, $R_X(.)$ and $R_{YX}(.)$. Finally, exogeneity of X in the bivariate system may be investigated by checking the lag distribution of $(Y(t) \,|\, H_X(-\infty, \infty)) = \sum_{-\infty}^{\infty} \mu(i) X(t-i)$, say, to see if it is one-sided on the past. Without any "zero restrictions" on the individual lag distributions, i.e., specifications that $a_i(.)$, $b_j(.)$, or $c_k(.)$ vanish identically, we will have

$$\begin{aligned}(6.6)\qquad \mu(.) = R_X^{-1} * R_{YX}(.) = {} & [(\beta\gamma - c_2 b_3) * (\beta'\gamma' - c_2' b_3')\sigma_1^2 + \\ & (c_1 b_3 - b_1\gamma) * (c_1' b_3' - b_1'\gamma')\sigma_2^2 + (b_1 c_2 - c_1\beta) * (b_1' c_2' - c_1'\beta')\sigma_3^2]^{-1} * \\ & [(c_2 a_3 - a_2\gamma) * (\beta'\gamma' - c_2' b_3')\sigma_1^2 + (\alpha\gamma - c_1 a_3) * (c_1' b_3' - b_1'\gamma')\sigma_2^2 + \\ & (c_1 a_2 - \alpha c_2) * (b_1' c_2' - c_1'\beta')\sigma_3^2].\end{aligned}$$

While it is conceivable for $\mu(.)$ to be one-sided on the past without any of the individual lag distributions vanishing, the possibility of this happening brings to mind the notion of identification by, say, zero restrictions in classical econometrics: parameters are so identified

(often loosely stated, a.e.--almost everywhere--in the parameter space) only on the condition that other parameters, thought not to vanish, do not, in fact, vanish. So it is here: we will say $\mu(.)$ is two-sided in general (or, two-sided) whenever $\mu(.)$ will be two-sided except for a meager subset[25/] of the possible choices of parameter values. All such statements in the sequel will be subject to this qualification. Also, we will say $\mu(.)$ is one-sided under a class of restrictions only when it is one-sided for all elements within the class.

We now offer an argument that, in the general case, where $b_1(.) \not\equiv 0$ ($Y \to X$ in the trivariate system), $\mu(.)$ will be two-sided; and, moreover, $\mu(.)$ will remain two-sided under any and all additional assumptions of the zero restriction variety (each of which correspond to causality events, as did $b_1(.)$). Indeed, the reader who is not convinced from the sight of (6.6) that $\mu(.)$ will be two-sided in general might consider expressing the convolutions as z-transforms, replacing transposes by z^{-1}, and considering the case where all lag distributions are generated by ratios of polynomials. Let, say, $R_X(z) = A(z)A(z^{-1})$, where $A(z) = \frac{P_1(z)}{Q_1(z)}$, and let the sum of terms which represents $R_{YX}(z)$ be factored as $\frac{P_2(z)P_3(z^{-1})}{Q_2(z)Q_3(z^{-1})}$ where these factors are uniquely determined up to a constant. Then $\mu(z) = \frac{Q_1(z)}{P_1(z)} \frac{Q_1(z^{-1})}{P_1(z^{-1})} \frac{P_2(z)P_3(z^{-1})}{Q_2(z)Q_3(z^{-1})}$, so that $\mu(.)$ is one-sided on the past only when an incredible matchup of zeros of the polynomials occurs so that $Q_1(.)P_3(.) = k.P_1(.)Q_3(.)$, k some constant. This is precisely the kind of event excluded by the previous discussion.

It is also the case that the situation does not change when lag distributions which may be set identically to zero are so set. Indeed, assume $c_1(.) \equiv c_2(.) \equiv a_2(.) \equiv 0$. In this case, (6.6) simplifies to

$$(6.7) \qquad \mu(.) = -[b_1\gamma b_1'\gamma'.\sigma_2^2 + \sigma_1^2.\beta\gamma\beta'\gamma']^{-1}(\alpha\gamma b_1'\gamma').$$

Now, note that, since $b_1(0) = 0$ by the normalization, after bringing $\gamma\gamma'$ out of the inverse, we are left with $(b_1b_1'.\sigma_2^2 + \sigma_1^2.\beta\beta')^{-1}$, which will <u>not</u> be of the form const $(b_1b_1')^{-1}$. So even in this special case, which simplifies $\mu(.)$ as much as is possible, when $b_1(.) \not\equiv 0$ the result is a two-sided bivariate regression.

There is one instance related to the general argument that (6.7) must be two-sided, which shows two-sidedness to be the case without allowing the possibility of even a chance cancellation. This follows from the

<u>Proportionality Factorization Lemma</u>. Let $a(0) = d(0) = 1$ be a normalization of the one-sided, ℓ_2, rational sequences a(.), b(.), c(.), d(.) (that is, $a(s) = 0$, $s < 0$ and $\sum_{s=0}^{\infty} |a(s)|^2 < \infty$ and similarly for b, c, and d). Assume further that a^{-1}, $(a*a' + b*b')^{-1}$ exist in ℓ_2 and (Z): the zeros of $\det\begin{pmatrix} a(z) & b(z) \\ c(z) & d(z) \end{pmatrix}$ lie outside $|z| = 1$. Then $(a*a' + b*b')^{-1} *(a'*c + b'*d)$ is one-sided if and only if (P) $k.a(s) = b(s)$, all $s \geq 0$. (k may be zero.)

Proof: Sufficiency is straightforward: Under (P),

$$(a*a' + b*b')^{-1} *(a'*c + b'*d) = [(1+k^2)\ a*a']^{-1} *a'(c+k.d) = (\frac{1}{1+k^2}).a*(c+kd)$$

which is one-sided on the past. Necessity is less evident, since it seems possible that $k.a(.) \ddagger b(.)$ and yet fortuitous cancellation at the negative lags might still yield one-sidedness. To see that this is not possible, define the process $\binom{X}{Y}(t) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} * \binom{e}{n}(t)$. By the assumptions on a,b,c,d, $\binom{e}{n}$, if taken as a white noise process, not only defines an $\binom{X}{Y}$ process, but is fundamental for it provided the condition (Z) is met. (Since $\binom{X}{Y}$ has a rational spectral density, this follows from Remark 2, p. 88, and Remarks 1-3, p. 62 of [17]; this is a "deep" result, although universally not appreciated as such, involving the relationship between maximal functions and fundamental representations.) Then, since all fundamental representations must be of the form $\binom{X}{Y}(t) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} P * P' \binom{e(t)}{n(t)}$, P orthonormal, we must have $a(s)p_{11} + b(s)p_{21}$ and $a(s)p_{12} + b(s)p_{22}$ in the top row of such a representation. But fundamental or not, the general formula for the two-sided projection $(Y(t) \mid H_X)$ is $R_{XX}^{-1} * R_{YX}$ and, since this is one-sided, by Sims' Theorem 2, (or Proposition 2) Y does not cause X. Application of Sims' Theorem 1 guarantees that $a(s)p_{12} + b(s)p_{22} \equiv 0$ for some p_{12}, p_{22}, which contradicts the hypothetical lack of proportionality. Necessity, and therefore the result, follow. Q.E.D.

Of course, there will be cases of rational lag distributions a(.), b(.) c(.), d(.) for which the determinental condition does not apply, but even in these cases, only for certain small sets of exact linear relations among the coefficients will the resultant $\mu(.)$ be one-sided. We have now proved half of, and essentially given the crux of the argument used in the converse of, our last result.

Proposition 7. Let the l.r.w.s.s.p. $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$ have the extended autoregressive representation (e.a.r.) $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}(t) + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}(t)$, where the error covariance is diagonal and we have normalized by setting $b_1(0) = c_1(0) = c_2(0) = 0$ and, of course, $a_1(0) = b_2(0) = c_3(0) = 0$. Then Y causes X ($b_1(.) \neq 0$) implies that Y causes X in the implied bivariate system $\binom{X}{Y}$. Conversely, if Y does not cause X in the (correctly specified) trivariate system ($b_1(.) \equiv 0$) then Y does not cause X in the corresponding bivariate system $\binom{X}{Y}$ in the following cases: (i) $c_1(.) \equiv 0$, or, (ii) $c_2(.) \equiv b_3(.) \equiv 0$. In case neither of these conditions hold, omitting Z will induce spurious causality from Y to X in the implied bivariate model.

Remark 1. The intended interpretation of the notation is, for example, in (i), when $c_1(.) \equiv 0$, the remaining lag distributions may or may not be zero. Causality requiring the very particular form of autocovariance sequence that it does, the first part of the proposition should come as no surprise; lack of causality in the trivariate system cannot, even with the aid of an imaginatively chosen set of further zero restrictions, produce spurious causality in the bivariate system.

Remark 2. Using Proposition 6 and the above discussion on instantaneous causality and normalization to translate the zero restrictions into the equivalent causality statements, we may paraphrase the converse to say that, when $Y \nrightarrow X$ relative to $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$, $Y \nrightarrow X$ relative to $\binom{X}{Y}$ if and only if

at least one of the additional conditions prevail: (i) $Z \nrightarrow X$; or (ii) $Z \nrightarrow Y$, $Y \nrightarrow Z$, $Z \overset{i_1}{\not\leftrightarrow} Y$. As a check of both our intuition and the plausibility of the result, we observe that a very special case occurs when (i), (ii), $X \nrightarrow Z$, and $X \overset{i_2}{\not\leftrightarrow} Z$ all obtain. Here $\sigma_{YZ} = 0$ as well, and Proposition 6 ensures that the trivariate Wold representation is bloc-diagonal: the Z and $\binom{X}{Y}$ processes are orthogonal. The conclusion that trivariate exogeneity carries over to the bivariate system under all these assumptions comes as no surprise. The point of the converse of this proposition is that it specifies the extent to which these restrictions may be relaxed, in terms conditions which are individually sufficient and jointly necessary. The conditions (i) and (ii) generally have a natural interpretation as a Theil-like omitted variables theorem in Hilbert space. Indeed, as in the proof of Proposition 1, if we project X(t) first onto $H_{X,Y,Z}(t-1)$, then onto $H_{X,Y}(t-1)$, we get

$$(X(t)\,|\,H_{X,Y}(t-1)) = \sum_{i=1}^{\infty} a_1(i)X(t-i) + \sum_{i=1}^{\infty} b_1(i)Y(t-i) + \sum_{i=1}^{\infty} c_1(i)(Z(t-i)\,|\,H_{X,Y}(t-1)).$$

Since we hypothesize $b_1(.) \equiv 0$, and the Granger criterion (Proposition 1) tells us that all coefficients on Y(t-i), i=1,2,... must vanish, the last term tells the story: $c_1(.) = 0$ is (i), or $Z \nrightarrow X$--the omitted variable didn't enter the true relation; if $c_1(.) \neq 0$, but

$$(\dagger) \qquad (Z(t-i)\,|\,H_{X,Y}(t-1)) = (Z(t-i)\,|\,H_X(t-1)), \text{ all } i=1,2,...$$

X will still be exogenous with respect to Y in the bivariate system. This latter condition might intuitively be read "the coefficients of the auxiliary regression, of omitted (Z) on included (Y) variable, vanish" and this latter projection equality thought logically equivalent to condition (ii) $Z \nrightarrow Y$, $Y \nrightarrow Z$, $Z \overset{i_2}{\not\leftrightarrow} Y$. Indeed, the proof of Proposition 7

provides more than casual support for this view. Actually, we have already seen a special case of the equivalence of the conditions (ii) and (†) in the case (iii) discussion; we postpone a formal statement of this equivalence, which appears important and interesting in its own right, until such time as a more straightforward and elegant proof can be given.

Remark 3. The proof given so far, and to be continued, is by brute force, examining all possibilities in the trivariate e.a.r., computing the implied two-sided projection $(Y(t)\,|H_X(-\infty, \infty))$ and checking to see if it is one-sided. The e.a.r. assumption obviates the need to use Proposition 2, as Sims' Theorem 2 is in force. An alternative strategy might try to find the implied, bivariate Wold decomposition from the trivariate e.a.r. or its inverse, the trivariate m.a.r. Factoring the implied $\binom{X}{Y}$ autocovariance function is then necessary to make any progress, and this appears a theoretically intractable problem. Were it possible to make a statement such as, "the factorization cannot be lower triangular" then Sims' Theorem 1 could be used. The technique employed here gets around this difficulty. As will be clear from its study, the sufficiency of the conditions is much more straightforward than their necessity.

Remark 4. It is interesting to note that, from the point of view of this result, there is some reason to favor the first definition of instantaneous trivariate causality because $\overset{i_1}{\leftrightarrow}$ does enter into the statement of condition (ii).

Continuation of Proof. For the converse, we maintain $b_1(.) \equiv 0$, and consider the three cases in order:

(i) When $c_1(.) \equiv 0$ as well,

(6.8) $$\mu(.) = [(\beta\gamma - c_2 b_3) * (\beta'\gamma' - c_2 b_3')\sigma_1^2]^{-1} \cdot \sigma_1^2 \cdot$$

$$(c_2 a_3 - a_2 \gamma) * (\beta'\gamma' - c_2' b_3')$$

i.e., $\mu(.) = (\beta\gamma - c_2 b_3)^{-1} * (c_2 a_3 - a_2 \gamma)$, which is one-sided, regardless of whether $\alpha(.)$, $\beta(.)$, or $\gamma(.)$ are concentrated at zero (equivalently, $a_1(.)$, $b_2(.)$, or $c_3(.)$ vanish).

(ii) When $c_2(.) \equiv b_3(.) \equiv 0$, then

(6.9) $$\mu(.) = [\beta\gamma\beta'\gamma' . \sigma_1^2 + c_1\beta c_1'\beta' . \sigma_3^2]^{-1}(-a_2\gamma\beta'\gamma'\sigma_1^2 - c_1 a_2 c_1'\beta'\sigma_3^2),$$

or, again,

$$\mu(.) = -\beta^{-1}\beta'^{-1}(\gamma\gamma'\sigma_1^2 + c_1 c_1'\sigma_3^2)^{-1} \; a_2\beta'(\gamma\gamma'\sigma_1^2 + c_1 c_1'\sigma_3^2),$$

and finally,

$\mu(.) = -\beta^{-1} a_2$, which is one-sided. Again, whether $\beta(.) = b_2(.) - \delta$, $\alpha(.)$ or $\gamma(.)$ are multiples of the Kronecker delta, δ, is irrelevant.

To show the necessity of these conditions, we must argue that, in all other cases, except for the occurrence of a cancellation of zeros, $\mu(.)$

will be two-sided. By making statements conditional on $a_1(.)$, $b_2(.)$, $c_3(.)$ not vanishing and since $b_1(.) \equiv 0$ has been maintained, 2^9 cases have been held to a manageable 2^5, of which 20 have been considered. The observation that the conditional statements were actually unconditional has cut down our effort and yet cost no generality. We group the remaining 12 (conditional) cases as:

(a) $b_1(.) \equiv c_2(.) \equiv 0$; $b_3(.)$, $c_1(.) \not\equiv 0$, $a_2(.)$, $a_3(.)$ anything.

Here, $\mu(.) = (\beta\beta'\gamma\gamma'.\sigma_1^2 + c_1 b_3 c_1' b_3'.\sigma_2^2 + c_1\beta c_1'\beta'\sigma_3^2)^{-1}$

$$(-\gamma a_2\gamma'\beta'.\sigma_1^2 - c_1 c_1' b_3' a_3'\sigma_2^2 + \alpha\gamma c_1' b_3'\sigma_2^2 - c_1 c_1'\beta' a_2.\sigma_3^2$$

Conditionally or unconditionally, the term which represents $R_{XX}^{-1}(.)$ will, under these assumptions, not be able to cancel the primed terms in $R_{YX}(.)$.

(b) $b_1(.) \equiv b_3(.) \equiv 0$; $c_1(.)$, $c_2(.) \not\equiv 0$; a_2, a_3 anything.

Here, $\mu(.) = \beta^{-1}(\gamma\gamma'.\sigma_1^2 + c_1 c_1'.\sigma_3^2)^{-1}[c_2 a_3\gamma'\sigma_1^2 - \gamma\gamma' a_2.\sigma_1^2 -$

$$c_1 c_1' a_2\sigma_3^2 + \alpha c_2 c_1'.\sigma_3^2].$$

Since $c_1(.) \not\equiv 0$, the first term must factor into something proportional to neither $\gamma(.)$ nor $c_1(.)$. Hence $\mu(.)$ is two-sided here, even under all assignments to $\alpha(.)$, $\beta(.)$, and $\gamma(.)$.

(c) $b_1(.) \equiv 0$, $c_2(.)$, $b_3(.) \not\equiv 0$; a_2, a_3 anything.

Here, $\mu(.) = [(\beta\gamma - c_2 b_3)*(\beta\gamma - c_2 b_3)'\sigma_1^2 + c_1 c_1' b_3 b_3' \sigma_2^2 + c_1 \beta c_1' \beta' . \sigma_3^2]^{-1} *$

$[(c_2 a_3 - a_2 \gamma)*(\beta\gamma - c_2 b_3)' . \sigma_1^2 + (\alpha\gamma - c_1 a_3) c_1' b_3' . \sigma_2^2 +$

$-(c_1 a_2 - \alpha c_2)*(c_1 \beta)' . \sigma_3^2]$

which is as two-sided as can be! Q.E.D.

VII. 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.

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 projections. The "intervention" sense of causality means finding 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 $m = \underline{c} * m + \underline{d} * y + v$ 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 last result, Proposition 7, confronts the criticism that bivariate findings are statistical artifacts which, when found, are likely to represent specification bias. By adding a third variable and analyzing the trivariate system $\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$, two lessons are learned: (i) The exogeneity of X in the bivariate system $\binom{X}{Y}$ will, if found, almost surely reflect trivariate causality; and (ii) The exogeneity of X in the trivariate system will be preserved in the bivariate system if and only if: (a) Z doesn't cause X either, or (b) neither Z nor Y cause the other, instantaneously or over time. (These conditions are individually sufficient and jointly necessary.) When these additional interactions cannot be ruled out, true exogeneity of X, as referred to the trivariate system, is blurred by the omitted series Z when perceived in the bivariate system.

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 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. Should the latter view win out, it would shed some needed light on the cause of the innovations approach's sampling shortcomings.

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. Even for bivariate systems, certain natural variational conjectures come to mind; and the issue of the relation, if any, of $H_{X,Y}(-\infty)$ to $H_{X,Y}(\infty)$ remains open. An elegant proof of the result in Remark 2 in Section VI will doubtless be forthcoming. Also, the specific question posed relating $Y \nrightarrow X$ in the bivariate and trivariate systems was only one of many that could be considered. Others, involving the passage of the notions of instantaneous causality from the trivariate to the bivariate system, come immediately to mind.

Finally, additional contributions which clarify, or, otherwise pronounce upon, any reconciliation of statistical causality with "intervention-based" notions of causality would be gratefully received.

Footnotes

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

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

3/ When 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.

4/ Because 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).

5/ The 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.

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

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

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

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

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 $z = X(t+1)$ and $z = Y(t)$. This operator enjoys many important properties: linearity, idempotency, continuity unit norm, self-adjointness, and positivity--none of which are exploited in this paper. That the projection minimizes the mean-square error follows from 5.8(6) of [30].

7/ $L: H_{X,Y}(-\infty, \infty) \to H_{X,Y}(-\infty, \infty)$ is defined by $L[X(t)] = L[X(t+1)]$ 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

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

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 deomonstrates. Of course, by saying that a vector $\binom{u}{w}(t)$ 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 Zasuhin [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.

10/ The 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.

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

12/ This is 6.10(b) of [30].

13/ 6.13(b) of [30].

14/ It 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.

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

16/ Thus, 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.

17/ The 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 $\binom{X}{Y}$ 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 e.a.r., and more: the set $\{X(t), Y(t), t\varepsilon I\}$ 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 simulataneous 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 $\binom{u}{v}$, we have

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

In the terminology of p. 15. We may verify directly that the proposed candidate for $\sum^{1/2} Q$ works and that $\binom{e}{f}$ has covariance matrix the identity.

21/ This 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 $(X(t)|\overline{N_1}) = 0$.

22/ Actually, 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 $\binom{X}{Y}$ 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.

23/ The 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 fundamentalness of fundamentalness. His oral tradition is reflected in the references to the work of Rozanov found in these pages.

24/ More explicitly, $X \overset{i_1}{\leftrightarrow} Z \iff \sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22} \neq 0 \iff \sigma_{13} - \sigma_{12}\sigma_{22}^{-1}\sigma_{23} \neq 0 \iff \sigma_{XZ.Y} \neq 0$, and $X \overset{i_2}{\leftrightarrow} Z \iff \sigma_{13} \neq 0$. Consequently, only when $\sigma_{12}\sigma_{23} = 0$ will $\sigma_{XZ.Y}$ and σ_{13} necessarily agree; in other words, $X \overset{i_1}{\leftrightarrow} Z$ whenever $X \overset{i_2}{\leftrightarrow} Z$ iff either $X \overset{i_2}{\leftrightarrow} Y$ or $Y \overset{i_2}{\leftrightarrow} Z$. Viewed as

subsets of the space Σ positive definite matrices, the events $X \overset{i_1}{\leftrightarrow} Z$ and $X \overset{i_2}{\leftrightarrow} Z$ both have lebesgue measure, m(.), zero, so that trivially their complements agree almost everywhere, as asserted. We do not pursue these and related issues here, because our chief concern is with causality flows over time.

25/ The word is chosen to connote technical senses in which this concept might be made more precise.

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, Novebmer 13-14, 1975.

[27] Skoog, G., "Systematically Missing Data in Econometric Models: Some Identification Considerations," mimeo June 1975.

[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, Almquist and Wicksell, Uppsalen, 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.