Comput Stat (2013) 28:1151–1167
DOI 10.1007/s00180-012-0351-z

ORIGINAL PAPER

Extracting informative variables in the validation of two-group causal relationship

Ying-Chao Hung · Neng-Fang Tseng

Received: 14 December 2011 / Accepted: 13 July 2012 / Published online: 31 July 2012

Abstract The validation of causal relationship between two groups of multivariate time series data often requires the precedence knowledge of all variables. However, in practice one finds that some variables may be negligible in describing the underlying causal structure. In this article we provide an explicit definition of "non-informative variables" in a two-group causal relationship and introduce various automatic computer-search algorithms that can be utilized to extract informative variables based on a hypothesis testing procedure. The result allows us to represent a simplified causal relationship by using minimum possible information on two groups of variables.

Keywords Causal relationship · Vector autoregression model · Informative variables · Modified Wald test · Automatic computer-search algorithm

1 Introduction

Over the years the causality system described by the multivariate time series process has been one of the most flexible and popular statistical techniques to measure the dynamic relationships between groups of variables in the areas of economics, finance, medicine, science, and engineering. The primary study of causal relationships can

Y.-C. Hung
Department of Statistics, National Chengchi University, NO. 64, Sec. 2, ZhiNan Rd.,
Wenshan District, Taipei 11605, Taiwan
e-mail: hungy@nccu.edu.tw

N.-F. Tseng (✉)
Department of Mathematical Statistics and Actuarial Science, Aletheia University,
32 Chen-Li Street, Tamsui, Taipei 25103, Taiwan
e-mail: au4225@mail.au.edu.tw

date back to the work by Granger (1969), wherein the vector autoregression (VAR) model, which is a generalization of the univariate AR models, was used to identify the "causality" between two groups of time series data based on precedence and predictability. Afterward, there exists a fairly rich literature on its extensive studies. Some remarkable works are: Granger (1980) proposed a statistical hypothesis testing procedure to validate the bivariate causal relationship; Osborn (1984) discussed about "Unidirectional Granger Causality" based on the ARMA model and probed it into a statistical hypothesis testing procedures; Geweke (1982, 1984) considered measures of linear dependence and feedback between multiple time series data and provided a comprehensive literature survey of Granger-causality; Boudjellaba et al. (1992) tested causality between two vectors in multivariate autoregressive moving average models; Granger and Lin (1995) talked about the measure of causality by using the spectral decomposition based on the vector error correction model (VECM); Mosconi and Giannine (1992) investigated the Granger causality based on a non-stationary VAR model; Roebroech et al. (2005) used the Granger causality mapping (GCM) to explore directed influences between neuronal populations in fMRI data; Hacker and Hatemi (2006) developed a method that is not sensitive to deviations from the assumption that the error term is normally distributed; Fujita et al. (2007) proposed an improved VAR model (called DVAR) to estimate time-varying gene regulatory networks based on gene expression profiles obtained from microarray experiments; Haufe et al. (2010) estimated causal interactions in multivariate time series using the VAR model; just to name a few.

The study of causal relationships usually include all variable information in the analysis. However, in many practical situations one finds that some variables are not particularly informative and can mislead the interpretation of the underlying causal structure. The work by Hsiao (1982) was closely related to such a concept. He introduced three different types of causal relationships (called direct, indirect, and spurious causality) by reducing the information set in a three-variate time series model. However, when the number of variables becomes large, it is a much harder task to characterize all the causal patterns due to model complexity. To overcome this problem, some graphical techniques have been successfully developed to identify and visualize the causal relationships between the components of multivariate time series data. The readers can refer to the works by Koster (1996, 1999), Lauritzen (1996, 2000), Pearl (1995, 2000), Whittaker (1990), and Arnold et al. (2007) for this type of approaches.

The goal of this study is to extract informative variables in the validation of causal relationship between two groups of multivariate time series data. These extracted variables are important and useful in the sense that it allows us to forecast the future quantity of explicit variables by utilizing the minimum data information. The remainder of this paper is organized as follows. In Sect. 2, we introduce some background knowledge required for defining and identifying informative variables in the validation of two-group causal relationship. In Sect. 3, we introduce how to extract all informative variables by utilizing a hypothesis testing procedure (called the modified Wald test) and various automatic computer-search algorithms. In Sect. 4, the computer-search algorithms are illustrated on a real example. Some concluding remarks are given in Sect. 5.

2 Background knowledge

The notion of causality in multivariate time series data is often discussed by the stationary pth-order VAR model (denoted by VAR(p)):

$$W_t = b + \sum_{j=1}^{p} A_j W_{t-j} + a_t, \quad t = 1, \ldots, T, \tag{1}$$

where b is a $(K \times 1)$ constant vector, $W_t = (W_{1,t}, W_{2,t}, \ldots, W_{K,t})'$ is a $(K \times 1)$ random vector, A_j is a $(K \times K)$ coefficient matrix for all $j = 1, \ldots, p$, and a_t is a $(K \times 1)$ error (or noise) vector satisfying that (i) $E(a_t) = \mathbf{0}$; (ii) the covariance matrix $E(a_t a_t')$ is positive definite (thus non-singular); and (iii) $E(a_t a_{t-k}') = \mathbf{0}$ for any non-zero k. Dividing all the variables of interest into two groups X_t and Y_t, we see that W_t can be further represented as

$$W_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \sum_{j=1}^{p} \begin{pmatrix} A_{XX,j} & A_{XY,j} \\ A_{YX,j} & A_{YY,j} \end{pmatrix} \begin{pmatrix} X_{t-j} \\ Y_{t-j} \end{pmatrix} + \begin{pmatrix} a_{X,t} \\ a_{Y,t} \end{pmatrix}, \quad t = 1, \ldots, T, \tag{2}$$

where $X_t = (X_{1,t}, \ldots, X_{n,t})$ and $Y_t = (Y_{1,t}, \ldots, Y_{m,t})$ are $(n \times 1)$ and $(m \times 1)$ random vectors, b_1 and b_2 are $(n \times 1)$ and $(m \times 1)$ constant vectors, $A_{XX,j}$, $A_{XY,j}$, $A_{YX,j}$, and $A_{YY,j}$ are sub-matrices of A_j with orders $(n \times n)$, $(n \times m)$, $(m \times n)$, and $(m \times m)$, respectively, $a_{X,t}$ and $a_{Y,t}$ are $(n \times 1)$ and $(m \times 1)$ error vectors. The primary goal of the so-called "Granger causality" is to examine whether or not the time series Y_t is useful in forecasting the time series X_t.

Given any point in time t, let us consider the two *information sets*

$$\Omega_{XY} = \{X_{1,t}, \ldots, X_{n,t}, \ldots, X_{1,1}, \ldots, X_{n,1}, Y_{1,t}, \ldots, Y_{m,t}, \ldots, Y_{1,1}, \ldots, Y_{m,1}\}$$

and

$$\Omega_X = \{X_{1,t}, \ldots, X_{n,t}, \ldots, X_{1,1}, \ldots, X_{n,1}\}.$$

For any given future time $(t + h)$, we denote the *best linear predictor* of X_{t+h} based on the information sets Ω_{XY} and Ω_X by

$$\hat{X}_t(h|\Omega_{XY}) = (\hat{X}_{1,t}(h|\Omega_{XY}), \ldots, \hat{X}_{n,t}(h|\Omega_{XY}))$$

and

$$\hat{X}_t(h|\Omega_X) = (\hat{X}_{1,t}(h|\Omega_X), \ldots, \hat{X}_{n,t}(h|\Omega_X),$$

respectively. The two-group causality (also known as generalization of Granger causality) is defined as follows.

Definition 1 (*Two-group Causality up to Horizon c*) Given any positive integer c, if $\hat{X}_t(h|\Omega_X) \neq \hat{X}_t(h|\Omega_{XY})$ for some $h \leq c$, then we say that Y_t causes X_t up to horizon c and denote it by $Y \underset{(c)}{\rightarrow} X$. On the other hand, if $\hat{X}_t(h|\Omega_X) = \hat{X}_t(h|\Omega_{XY})$ for all $h \leq c$, then we say that Y_t does not cause X_t up to horizon c and denote it by $Y \underset{(c)}{\nrightarrow} X$.

The following proposition is useful for identifying the causality/non-causality between X_t and Y_t.

Proposition 1 *Based on the model in Eqs. (1)–(2), for any positive integer c we have that* $Y \underset{(c)}{\nrightarrow} X$ *if and only if* $A_{XY,j} = \mathbf{0}_{n\times m}$ *for all* $j = 1, \ldots, p$.

Proof Since we know that $Y \underset{(c)}{\nrightarrow} X$ is equivalent to $Y \underset{(\infty)}{\nrightarrow} X$ (see Dufour and Renault (1998), Proposition 2.3) and $Y \underset{(\infty)}{\nrightarrow} X$ if and only if $A_{XY,j} = \mathbf{0}_{n\times m}$ for all $j = 1, \ldots, p$ (see Lütkepohl 2005, Corollary 2.2.1), the result is simply obtained. □

Proposition 1 indicates that the two-group causality based on the VAR model can be determined by examining the coefficient matrix $A_{XY,j}$. We next review some properties that are necessary for establishing the procedure of extracting informative variables in the later section.

As a result of Definition 1, if $Y \underset{(c)}{\rightarrow} X$ then there exists at least one pair $(i, h) \in \{1, \ldots, n\} \times \{1, \ldots, c\}$ such that

$$E\left(\hat{X}_{i,t}(h|\Omega_{XY}) - X_{i,t+h}\right)^2 < E\left(\hat{X}_{i,t}(h|\Omega_X) - X_{i,t+h}\right)^2,$$

where $\hat{X}_{i,t}(h|\Omega_{XY})$ and $\hat{X}_{i,t}(h|\Omega_X)$ are the ith element of $\hat{X}_t(h|\Omega_{XY})$ and $\hat{X}_t(h|\Omega_X)$, respectively. Now we introduce how to calculate $\hat{X}_t(h|\Omega_{XY})$. Based on Eq. (1), for any given time lag $h > 0$ we have that

$$W_{t+h} = \sum_{k=0}^{h-1} A_1^{(k)}(b + a_{t+h-k}) + \sum_{j=1}^{p} A_j^{(h)} W_{t+1-j}, \tag{3}$$

where $A_j^{(k)}$ is a matrix obtained from the recursive formula

$$A_j^{(k)} = \begin{cases} A_j & k = 1 \\ A_{j+1}^{(k-1)} + A_1^{(k-1)} A_j & k = 2, 3, \ldots, h, \end{cases} \tag{4}$$

and $j = 1, \ldots, p$. Consider the following partition of matrix $A_j^{(h)}$:

$$A_j^{(h)} = \begin{pmatrix} A_{XX,j}^{(h)} & A_{XY,j}^{(h)} \\ A_{YX,j}^{(h)} & A_{YY,j}^{(h)} \end{pmatrix},$$

where $A^{(h)}_{XX,j}$ and $A^{(h)}_{XY,j}$ are two sub-matrices with orders $(n \times n)$ and $(n \times m)$, respectively. Denote the *identity matrix* of order n by $\mathbf{I}_n$, it is clear that $X_{t+h} = (\mathbf{I}_n, \mathbf{0}_{n\times m})W_{t+h}$. Based on the notations introduced above, the best linear predictor (in matrix form) is given by

$$\hat{X}_t(h|\Omega_{XY}) = b_{1,h} + \sum_{j=1}^{p} \left(A^{(h)}_{XX,j} X_{t+1-j} + A^{(h)}_{XY,j} Y_{t+1-j} \right), \tag{5}$$

where $b_{1,h} = (\mathbf{I}_n, \mathbf{0}_{n\times m}) \sum_{k=0}^{h-1} A_1^{(k)} b$. Equation (5) shows that the best linear predictor $\hat{X}_t(h|\Omega_{XY})$ relates to Y_t merely through the coefficient matrix $A^{(h)}_{XY,j}$. This will serve as an important benchmark for the remaining of this study.

Note that Definition 1 focuses on the causal relationship between the two random vectors X_t and Y_t. In particular, it explicitly defines whether or not Y_t can improve the forecasting of X_{t+h}. However, by the preceding arguments we learn that if $Y \underset{(c)}{\to} X$, then it is guaranteed that adding all variables in Y_t into the information set will improve the forecasting of "some" variables in X_t—but not definitely all. On the other hand, the forecasting of X_{t+h} may be improved by utilizing merely the information of "some" variables in Y_t—but not necessarily all. Therefore, our goal here is to provide a statistical procedure to extract those "informative variables" in both X_t and Y_t. To do this, we first introduce the definition of "non-informative variables" in both X_t and Y_t.

Definition 2 (*Non-informative Variables in X_t and Y_t*) Consider the VAR (p) model described in Eqs. (1)–(2) and assume that $Y \underset{(c)}{\to} X$ for some given integer $c > 0$.

(a) The variable $Y_{i,t}$ in $Y_t = (Y_{1,t}, \ldots, Y_{m,t})'$ is *non-informative* if

$$\hat{X}_t(h|\Omega_{XY}) = \hat{X}_t(h|\Omega_{XY_{-i}}) \text{ for all } h \le c, \tag{6}$$

where $\Omega_{XY_{-i}} = \Omega_{XY} \setminus \{Y_{i,t}, \ldots, Y_{i,1}\}$ refers to the reduced information set with the ith variable in Y_t being excluded.

(b) The variable $X_{i,t}$ in $X_t = (X_{1,t}, \ldots, X_{n,t})'$ is *non-informative* if

$$\hat{X}_{i,t}(h|\Omega_{XY}) = \hat{X}_{i,t}(h|\Omega_X) \text{ for all } h \le c. \tag{7}$$

The result of Definition 2 directly implies that, if the prediction of X_{t+h} based on Ω_{XY} is the same as that based on the reduced information set $\Omega_{XY_{-i}}$, then the variable $Y_{i,t}$ can be excluded from analysis (since it is non-informative in predicting X_t). Analogously, if the prediction of $X_{i,t+h}$ based on Ω_{XY} is the same as that based on Ω_X, then the variable $X_{i,t}$ can be excluded from analysis. The following two theorems provide useful guidelines for finding the non-informative (or informative) variables in both X_t and Y_t.

Theorem 1 (Identification of Non-informative Variables in Y_t) *Consider the matrix $A^{(h)}_{XY,j}$ given in Eq. (5) and its column partition*

$$A_{XY,j}^{(h)} = \left(A_{XY,j}^{(h)}(:,1), A_{XY,j}^{(h)}(:,2), \ldots, A_{XY,j}^{(h)}(:,m) \right), \tag{8}$$

where $A_{XY,j}^{(h)}(:,i)$ *refers to the* i*th column of* $A_{XY,j}^{(h)}$*. Then for any given* $i \in \{1,\ldots,m\}$*,* $Y_{i,t}$ *is non-informative if and only if* $A_{XY,j}^{(h)}(:,i) = \mathbf{0}$ *for all* $(h,j) \in \{1,\ldots,c\} \times \{1,\ldots,p\}$.

Proof Although the proof is quite similar to the one shown by Dufour and Renault (1998), we sketch it here for the sake of completeness. Based on Eq. (5), $\hat{X}_t(h|\Omega_{XY})$ can be further represented as

$$\begin{aligned}\hat{X}_t(h|\Omega_{XY}) &= b_{1,h} + \sum_{j=1}^{p} A_{XX,j}^{(h)} X_{t+1-j} + \sum_{l=1}^{m}\sum_{j=1}^{p} A_{XY,j}^{(h)}(:,l) Y_{l,t+1-j} \\ &= b_{1,h} + \sum_{j=1}^{p} A_{XX,j}^{(h)} X_{t+1-j} + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,i) Y_{i,t+1-j} \\ &\quad + \sum_{l\neq i}^{m}\sum_{j=1}^{p} A_{XY,j}^{(h)}(:,l) Y_{l,t+1-j},\end{aligned}$$

where the last equality is obtained by dividing the information set Ω_Y into $\{Y_{i,t}\}$ and $\Omega_{Y_{-i}}$. Thus, by treating $\Omega_{Y_{-i}}$ as the set of "auxiliary variables", we can conclude that $A_{XY,j}^{(h)}(:,i) = \mathbf{0}$ for all $(h,j) \in \{1,\ldots,c\} \times \{1,\ldots,p\}$ is the necessary and sufficient condition for $\hat{X}_t(h|\Omega_{XY}) = \hat{X}_t(h|\Omega_{XY_{-i}})$ (the result of Theorem 3.1 by Dufour and Renault 1998). The result then follows.

Theorem 2 (Identification of Non-informative Variables in X_t) *Consider the matrix* $A_{XY,j}^{(h)}$ *given in Eq. (5) and its row partition*

$$A_{XY,j}^{(h)} = \begin{pmatrix} A_{XY,j}^{(h)}(1,:) \\ A_{XY,j}^{(h)}(2,:) \\ \vdots \\ A_{XY,j}^{(h)}(n,:) \end{pmatrix}, \tag{9}$$

where $A_{XY,j}^{(h)}(i,:)$ *refers to the* i*th row of* $A_{XY,j}^{(h)}$*. Then for any given* $i \in \{1,\ldots,n\}$*,* $X_{i,t}$ *is non-informative if and only if* $A_{XY,j}^{(h)}(i,:) = \mathbf{0}$ *for all* $(h,j) \in \{1,\ldots,c\} \times \{1,\ldots,p\}$.

Proof The proof is quite similar to that of Theorem 1. By extending the formula given in Eq. (5), we have that

$$\begin{aligned}\hat{X}_{i,t}(h|\Omega_{XY}) &= b_{1,h}(i) + \sum_{j=1}^{p} A^{(h)}_{XX,j}(i,:)X_{t+1-j} + \sum_{j=1}^{p} A^{(h)}_{XY,j}(i,:)Y_{t+1-j} \\ &= b_{1,h}(i) + \sum_{j=1}^{p} A^{(h)}_{XX,j}(i,i)X_{i,t+1-j} + \sum_{l\neq i}^{n}\sum_{j=1}^{p} A^{(h)}_{XX,j}(i,l)X_{l,t+1-j} \\ &\quad + \sum_{j=1}^{p} A^{(h)}_{XY,j}(i,:)Y_{t+1-j}\end{aligned}$$

where $b_{1,h}(i)$ refers to the ith element of vector $b_{1,h}$, and the last equality is obtained by dividing the information set Ω_X into $\{X_{i,t}\}$ and $\Omega_{X_{-i}}$. Thus, by treating $\Omega_{X_{-i}}$ as the set of "auxiliary variables", we can conclude that $A^{(h)}_{XY,j}(i,:) = \mathbf{0}$ for all $(h, j) \in \{1, \ldots, c\} \times \{1, \ldots, p\}$ is the necessary and sufficient condition for $\hat{X}_{i,t}(h|\Omega_{XY}) = \hat{X}_{i,t}(h|\Omega_X)$ (the result of Theorem 3.1 by Dufour and Renault 1998). The result then follows.

3 Extracting informative variables

Theorems 1 and 2 state that the informative variables for two-group causality can be explicitly identified by examining the row and column vectors of the coefficient matrix $A^{(h)}_{XY,j}$. However, in practice the parameters in $A^{(h)}_{XY,j}$ are usually unknown and need to be estimated. Therefore, to extract all informative variables one can resort to a study analogous to "model selection". When the number of variables is large, some commonly used algorithms are: stepwise, forward selection, and backward elimination. These algorithms involve a multi-stage procedure of variable selection and/or elimination that are executed based on the so-called *modified Wald test* proposed by Lütkepohl and Burda (1997). Before we proceed, let us first look at the following simple example that illustrates how a desired modified Wald test is performed.

3.1 The modified Wald test

Let us consider the following three-variate VAR(1) process with

$$\begin{pmatrix} X_{1,t} \\ Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{pmatrix} A_{X_1X_1} & A_{X_1Y_1} & A_{X_1Y_2} \\ A_{Y_1X_1} & A_{Y_1Y_1} & A_{Y_1Y_2} \\ A_{Y_2X_1} & A_{Y_2Y_1} & A_{Y_2Y_2} \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + a_t.$$

Given $c = 2$, suppose we would like to test whether or not $Y_{1,t}$ is an informative variable in the causal relation $Y \underset{(2)}{\rightarrow} X$, by Definition 2 we can test the null hypothesis

$$H_0 : \hat{X}_t(h|\Omega_{XY}) = \hat{X}_t(h|\Omega_{XY_{-1}}) \quad \text{for } h = 1, 2. \qquad (10)$$

Based on the result of Theorem 1, if $A^{(1)}_{XY,1}(:,1)$ and $A^{(2)}_{XY,1}(:,1)$ are both close to zero, then the null hypothesis is not rejected and $Y_{1,t}$ is characterized as a non-informative variable; otherwise it is characterized as an informative variable. To perform a general test for any given c and VAR(p) model, we consider the matrix $\mathbf{A}^{(h)} = (A^{(h)}_1, \ldots, A^{(h)}_p)$ for $h = 1, \ldots, c$ (recall that $A^{(h)}_j$ are matrices defined in Eq. (4)) and the column vector

$$\boldsymbol{\alpha} = \begin{pmatrix} \mathrm{vec}(\mathbf{A}^{(1)}) \\ \vdots \\ \mathrm{vec}(\mathbf{A}^{(c)}) \end{pmatrix},$$

where "vec" is the *column stacking operator* that creates a column vector from the matrix by stacking its column vectors below one another. Thus, at each stage the hypotheses for testing whether or not a particular variable $Y_{i,t}$ (or $X_{i,t}$) is informative can be written as the form

$$\begin{cases} H_0 : (\mathbf{I}_c \otimes R)\boldsymbol{\alpha} = \mathbf{0} \\ H_a : (\mathbf{I}_c \otimes R)\boldsymbol{\alpha} \neq \mathbf{0} \end{cases} \tag{11}$$

where $\otimes$ is the *Kronecker product* so that

$$\mathbf{I}_c \otimes R = \begin{pmatrix} 1 \cdot R & 0 \cdot R & \cdots & 0 \cdot R \\ 0 \cdot R & 1 \cdot R & \cdots & 0 \cdot R \\ \vdots & \vdots & \ddots & \vdots \\ 0 \cdot R & 0 \cdot R & \cdots & 1 \cdot R \end{pmatrix}_{cr \times cpK^2},$$

and R is a "designated" $(r \times pK^2)$ matrix that corresponds to the null hypothesis (here $r = \#$ of variables in X_t or Y_t considered in the null hypothesis and $K = m + n$). To illustrate, the preceding example for the null hypothesis in Eq. (9) gives that

$$\mathbf{A}^{(h)} = \begin{pmatrix} A_{X_1X_1} & A_{X_1Y_1} & A_{X_1Y_2} \\ A_{Y_1X_1} & A_{Y_1Y_1} & A_{Y_1Y_2} \\ A_{Y_2X_1} & A_{Y_2Y_1} & A_{Y_2Y_2} \end{pmatrix}^h \quad \text{for } h = 1, 2, \text{ and } \boldsymbol{\alpha} = \begin{pmatrix} \mathrm{vec}(\mathbf{A}^{(1)}) \\ \mathrm{vec}(\mathbf{A}^{(2)}) \end{pmatrix}.$$

To test if $Y_{1,t}$ is informative, we can simply choose

$$R = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

so as to satisfy that

$$(\mathbf{I}_2 \otimes R)\boldsymbol{\alpha} = \begin{pmatrix} R & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} \mathrm{vec}(\mathbf{A}^{(1)}) \\ \mathrm{vec}(\mathbf{A}^{(2)}) \end{pmatrix} = \begin{pmatrix} A^{(1)}_{XY,1}(:,1) \\ A^{(2)}_{XY,1}(:,1) \end{pmatrix}.$$

As a result, testing the null hypothesis that

$$(\mathbf{I}_2 \otimes R)\boldsymbol{\alpha} = \mathbf{0} \quad \text{is then equivalent to testing} \begin{pmatrix} A^{(1)}_{XY,1}(:,1) \\ A^{(2)}_{XY,1}(:,1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Analogously, to test if $Y_{2,t}$ is informative, we can simply choose

$$R = \begin{pmatrix} 0\;0\;0\;0\;0\;0\;1\;0\;0 \end{pmatrix};$$

while to test if $X_{1,t}$ is informative, we can simply choose

$$R = \begin{pmatrix} 0\;0\;0\;1\;0\;0\;0\;0\;0 \\ 0\;0\;0\;0\;0\;0\;1\;0\;0 \end{pmatrix}.$$

Note that once the hypothesis in Eq. (11) is specified, the *modified Wald test statistic* is given by

$$\lambda^{mod}_{Wald} = T\left((\mathbf{I}_c \otimes R)\hat{\boldsymbol{\alpha}} + \frac{\hat{w}}{\sqrt{T}}\right)' \left((\mathbf{I}_c \otimes R)\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}}(\mathbf{I}_c \otimes R)' + \lambda\hat{\Sigma}_{\hat{w}}\right)^{-1} \cdot \left((\mathbf{I}_c \otimes R)\hat{\boldsymbol{\alpha}} + \frac{\hat{w}}{\sqrt{T}}\right), \tag{12}$$

where T is the number of data observations, $\hat{\boldsymbol{\alpha}}$ is the least square estimator of $\boldsymbol{\alpha}$, $\hat{w}$ is a random vector (independent of $\hat{\boldsymbol{\alpha}}$) drawn from a multivariate normal distribution $MN(\mathbf{0}, \lambda\hat{\Sigma}_{\hat{w}})$, λ is usually a small positive value (relative to T) chosen by the user,

$$\hat{\Sigma}_{\hat{w}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{c-1} \otimes diag(R\hat{\Sigma}_{\hat{\beta}}R') \end{pmatrix},$$

where $\hat{\beta}$ is the least square estimator of $\beta = \text{vec}(\mathbf{A}^{(1)})$, $\hat{\Sigma}_{\hat{\beta}}/T$ is the estimated covariance matrix of $\hat{\beta}$,

$$\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}} = \begin{pmatrix} \mathbf{I} \\ \sum_{i=0}^{1} \mathbf{B}^{1-i} \otimes J\mathbf{B}^i J' \\ \vdots \\ \sum_{i=0}^{c-1} \mathbf{B}^{c-1-i} \otimes J\mathbf{B}^i J' \end{pmatrix} \hat{\Sigma}_{\hat{\beta}} \begin{pmatrix} \mathbf{I} \\ \sum_{i=0}^{1} \mathbf{B}^{1-i} \otimes J\mathbf{B}^i J' \\ \vdots \\ \sum_{i=0}^{c-1} \mathbf{B}^{c-1-i} \otimes J\mathbf{B}^i J' \end{pmatrix}',$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{A}^{(1)} \\ \mathbf{I}_{K(p-1)} \quad \mathbf{0}_{K(p-1)\times K} \end{pmatrix} \quad \text{and} \quad J = \begin{pmatrix} \mathbf{I}_K, \mathbf{0}_{K\times K(p-1)} \end{pmatrix}.$$

It has been shown (Lütkepohl and Burda 1997) that

$$\lambda^{mod}_{Wald} \xrightarrow{d} \chi^2(rc) \quad \text{under the null hypothesis in Eq. (11).}$$

Therefore, given a significance level α , the null hypothesis in Eq. (11) is rejected when $\lambda_{Wald}^{mod} > \chi^2_{1-\alpha}(rc)$.

Remark 1 If $c = 1$, the random vector $\hat{w}$ has a degenerate distribution localized at zero (i.e., $\hat{w} = \mathbf{0}$ and $\hat{\Sigma}_{\hat{w}} = \mathbf{0}$ almost surely). In this case the modified Wald test reduces to the standard Wald test.

3.2 Automatic computer-search algorithms

Note that if the identification of every informative variable relies merely on one modified Wald test (i.e., with all remaining variables included in analysis), the search procedure can lead to the dropping of "true" informative variables (the issue is similar to that in regression analysis the full model is considered while variables are selected merely based on the t statistics, see Kutner et al. 2008). On the other hand, if the modified Wald test is conducted by considering all possible subsets of variables, the search procedure can be computationally expensive (especially when the number of variables is large). To overcome these problems, we introduce some automatic computer-search algorithms that have been widely used in regression analysis, such as the *forward selection*, *backward elimination*, and $stepwise$. We introduce the ideas of these algorithms in the following.

Forward Selection: At each stage the algorithm includes one "most informative" variable based on a predetermined level of the corresponding p value, but omitting the test whether an included variable should be removed. The algorithm terminates if no further variables can be included.

Backward Elimination: This is the opposite of forward selection. The algorithm starts with the model containing all variables and at each stage remove one "most non-informative" variable based on a predetermined level of the corresponding p value. The algorithm terminates if no further variables can be removed.

Forward Stepwise: At each stage the algorithm first includes one "most informative" variable based on a predetermined level of the corresponding p value (called α-to-enter) and, if there are any of the other variables in the model, remove one "most non-informative" variable based on another predetermined level of the corresponding p value (called α-to-remove). The algorithm terminates if no further variables can either be included or removed.

It should be mentioned here that, the algorithms introduced above all result in approximations to the "best set" of informative variables. In addition, there is no guarantee that the search results of different algorithms will be the same. We summarize the steps of implementing these algorithms as follows, in which we start with extracting the informative variables in Y_t and the informative variables in X_t afterwards.

Step 1: Select one algorithm and denote the corresponding initial set of informative variables in Y_t and X_t by $Y(0)$ and $X(0)$, respectively. Set the initial stage $k = 0$.

Step 2: Set the stage $k = k + 1$. Update the set of informative variables in Y_t based on $X(0)$ and the associated modified Wald tests, denote the updated set by $Y(k)$.

Step 3: Repeat Step 2 until the stopping criterion is satisfied. Denote the resulting estimated set of informative variables in Y_t by $\tilde{Y}_t$. Set the stage $k = 0$.

Table 1 The ADF test for the transformed eight time series data based on the AR(1) model

Variables	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$
Test statistic	−3.62	−3.16	−4.26	−2.78	−3.70	−3.32	−3.54	−2.75
p value	0.006	0.024	0.000	0.064	0.005	0.015	0.008	0.067

Step 4: Set the stage $k = k + 1$. Update the set of informative variables in X_t based on $\tilde{Y}_t$ and the associated modified Wald tests, denote the updated set by $X(k)$.
Step 5: Repeat Step 4 until the stopping criterion is satisfied. Denote the resulting estimated set of informative variables in X_t by $\tilde{X}_t$.
Step 6: Extract the obtained two sets $\tilde{Y}_t$ and $\tilde{X}_t$.

4 A real example

Now we illustrate the algorithms introduced in Sect. 3.2 on a real example. Let us consider the following two groups of econometric variables in the United States, which were retrieved from the database of Taiwan Economic Journal (http://www.finasia.biz):

$X_{1,t}$: Exports of Goods (seasonally adjusted, in millions USD)
$X_{2,t}$: Imports of Goods (seasonally adjusted, in millions USD)
$X_{3,t}$: Dow Jones Industrial Average Index
$X_{4,t}$: S&P 500 Index
$Y_{1,t}$: Consumer Price Index (seasonally adjusted)
$Y_{2,t}$: New Orders for Manufactured Goods (in millions USD)
$Y_{3,t}$: ISM Manufacturing Index
$Y_{4,t}$: Leading Indicators

Note that in order to make the primary time series become stationary, each variable is transformed into the "growth rate" (i.e., the natural logarithm of (the value in the present period)/(the value in the previous period)). The resulting time series plots are then given in Fig. 1, wherein the growth rates are presented monthly over the period from January 2001 to September 2011.

To validate the property of stationarity, the Augmented Dickey–Fuller (ADF) test is performed for each of the transformed time series data. For example, if $Y_{i,t}$ is the transformed time series, then a Unit Root test based on the simple assumption of AR(1) model is performed using the regression equation

$$\triangle Y_{i,t} = \mu + \beta Y_{i,t-1} - \alpha_1 \triangle Y_{i,t-1} + \varepsilon_t,$$

where μ is a constant and $\triangle$ is the first difference operator. The results for all the transformed time series are given in Table 1.

As can be seen from Table 1, the *p* values for the ADF tests of these eight variables are rather small. This strongly supports that all the time series data are stationary after transformation.

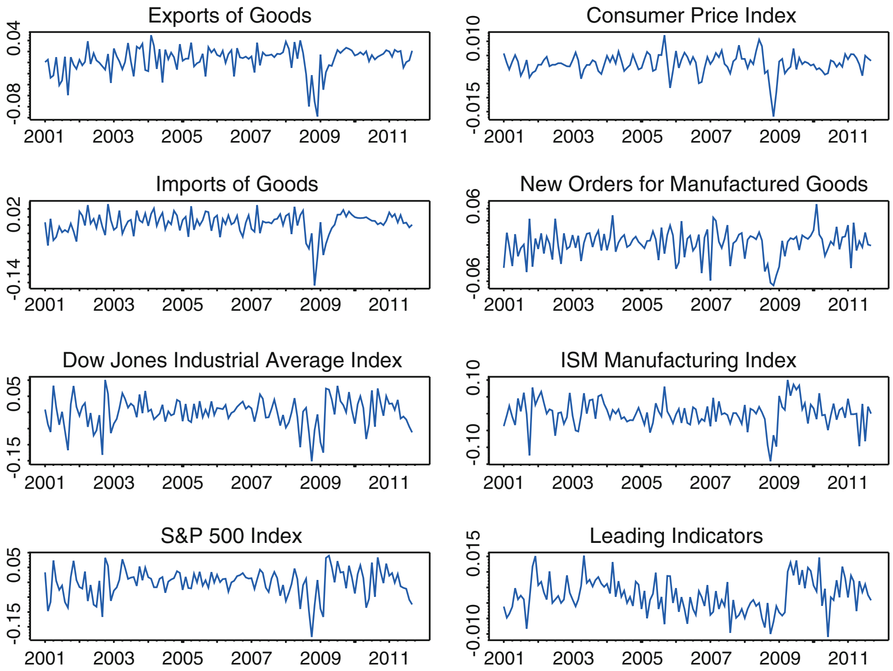

Fig. 1 The time series plots for the growth rates of the eight variables recorded monthly from January 2001 to September 2011

We next consider fitting a VAR(p) model for the eight variables based on the data observed from January 2001 to December 2010. The data observed from January 2011 to September 2011 are left for testing the predictability of the informative variables extracted by different algorithms (see later in Sect. 4). After a complete model selection procedure based on the Akaike information criterion (AIC), a VAR(2) model is chosen so as to validate the desired causal relationship. The stage-by-stage results of the three algorithms (forward stepwise, forward selection, and backward elimination) for extracting the informative variables along with the corresponding computing times are given in Tables 2, 3, 4, 5, 6 and 7. Note that in this study we simply choose $c = 2$. In addition, for each modified Wald test the value of α-to-enter is chosen to be 0.05, while the value of α-to-remove is chosen to be 0.10 (Kutner et al. 2008) suggested that the value of α-to-enter should be less than the value of α-to-remove for model selection). All numerical results in this section were performed by using the software package R (version 2.13.0) and executed on 3.0 GHz AMD Athlon II X2 250 processors with 4GB of cache under the operating system of Microsoft Windows 7 32-bit Service Pack 1 (SP1).

Based on Tables 2, 3, 4, 5, 6 and 7, the two sets of informative variables extracted by the forward selection algorithm are $\tilde{Y}_t = \{Y_{1,t}, Y_{2,t}, Y_{3,t}\}$ and $\tilde{X}_t = \{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$; the two sets of informative variables extracted by the backward elimination algorithm are $\tilde{Y}_t = \{Y_{1,t}\}$ and $\tilde{X}_t = \{X_{1,t}, X_{2,t}, X_{3,t}\}$; and the two sets of informative variables extracted by the forward stepwise algorithm are

Table 2 The stage-by-stage result of the "forward selection" algorithm for extracting informative variables in Y_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Enter
	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$	
$k = 1$	0.000	0.022	0.007	0.021	$Y_{1,t}$
$k = 2$	*	0.009	0.087	0.680	$Y_{2,t}$
$k = 3$	*	*	0.019	0.107	$Y_{3,t}$
$k = 4$	*	*	*	0.154	*
Computing time	7.66 (second)				

Note that for each modified Wald test the values of α-to-enter is chosen to be 0.05
* Represents the invalid cases

Table 3 The stage-by-stage result of the "forward selection" algorithm for extracting informative variables in X_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Enter
	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	
$k = 1$	6.7×10^{-9}	2.1×10^{-132}	0.022	0.014	$X_{2,t}$
$k = 2$	9.5×10^{-58}	*	0.011	1.1×10^{-5}	$X_{1,t}$
$k = 3$	*	*	1.9×10^{-7}	2.9×10^{-6}	$X_{3,t}$
$k = 4$	*	*	*	2.2×10^{-24}	$X_{4,t}$
Computing time	7.75 (second)				

Note that for each modified Wald test the values of α-to-enter is chosen to be 0.05
* Represents the invalid cases

Table 4 The stage-by-stage result of the "backward elimination" algorithm for extracting informative variables in Y_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Remove
	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$	
$k = 1$	5.1×10^{-10}	1.8×10^{-5}	0.040	0.358	$Y_{4,t}$
$k = 2$	4.7×10^{-6}	0.131	0.240	*	$Y_{3,t}$
$k = 3$	4.9×10^{-6}	0.117	*	*	$Y_{2,t}$
$k = 4$	0.005	*	*	*	*
Computing time	10.34 (second)				

Note that for each modified Wald test the values of α-to-remove is chosen to be 0.10
* Represents the invalid cases

$\tilde{Y}_t = \{Y_{1,t}, Y_{2,t}\}$ and $\tilde{X}_t = \{X_{1,t}, X_{2,t}\}$. For comparison purposes, the informative variables extracted by the three algorithms are summarized in Table 8.

Remark 2 It is noted that for small samples, the size of the modified Wald test can be sensitive to the choice of λ in Eq. (12). To avoid this problem, for this particular data set we suggest the following rule of thumb for choosing λ at each stage of the modified Wald test:

Table 5 The stage-by-stage result of the "backward elimination" algorithm for extracting informative variables in X_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Remove
	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	
$k = 1$	0.001	1.4×10^{-4}	0.013	0.143	$X_{4,t}$
$k = 2$	3.6×10^{-4}	0.001	4.6×10^{-4}	*	*
Computing time	0.43 (second)				

Note that for each modified Wald test the values of α-to-remove is chosen to be 0.10

* Represents the invalid cases

Table 6 The stage-by-stage result of the "forward stepwise" algorithm for extracting informative variables in Y_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Enter/Remove
	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$	
$k = 1$	0.013	0.022	0.031	0.384	$Y_{1,t}$ (Enter)
	*	*	*	*	*
$k = 2$	*	0.092	0.013	0.260	$Y_{3,t}$ (Enter)
	0.000	*	*	*	*
$k = 3$	*	0.003	*	0.616	$Y_{2,t}$ (Enter)
	0.000	*	0.180	*	$Y_{3,t}$ (Remove)
$k = 4$	*	*	0.094	0.113	*
	*	*	*	*	*
Computing time	12.16 (second)				

Note that for each modified Wald test the values of α-to-enter and α-to-remove are chosen to be 0.05 and 0.10, respectively

* Represents the invalid cases

Table 7 The stage-by-stage result of the "forward stepwise" algorithm for extracting informative variables in X_t and the associated computing time

Stage	The p value of λ_{Wald}^{mod}				Enter/Remove
	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	
$k = 1$	7.2×10^{-14}	2.5×10^{-45}	0.196	0.061	$X_{2,t}$ (Enter)
	*	*	*	*	*
$k = 2$	2.5×10^{-24}	*	0.213	0.015	$X_{1,t}$ (Enter)
	*	1.1×10^{-42}	*	*	*
$k = 3$	*	*	0.389	0.104	*
	*	*	*	*	*
Computing time	3.86 (second)				

Note that for each modified Wald test the values of α-to-enter and α-to-remove are chosen to be 0.05 and 0.10, respectively

* Represents the invalid cases

Table 8 The two sets of informative variables extracted by the three algorithms

Algorithm	The estimated set $\tilde{Y}_t$	The estimated set $\tilde{X}_t$
Forward selection	$\{Y_{1,t}, Y_{2,t}, Y_{3,t}\}$	$\{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}\}$
Backward elimination	$\{Y_{1,t}\}$	$\{X_{1,t}, X_{2,t}, X_{3,t}\}$
Forward stepwise	$\{Y_{1,t}, Y_{2,t}\}$	$\{X_{1,t}, X_{2,t}\}$

Table 9 The forecasting of $X_{1,t+h}$ and $X_{2,t+h}$ based on (i) the estimated sets $\tilde{Y}_t$ and $\tilde{X}_t$ of the three algorithms; and (ii) all the variables in Y_t and X_t

Algorithm	Forecasting of $X_{1,t+h}$		Forecasting of $X_{2,t+h}$	
	MSEP	MAEP	MSEP	MAEP
Forward stepwise	1.19×10^{-4}	0.0094	1.05×10^{-4}	0.0073
Forward selection	1.01×10^{-4}	0.0087	6.77×10^{-5}	0.0064
Backward elimination	1.08×10^{-4}	0.0092	7.98×10^{-5}	0.0067
All variable information	1.03×10^{-4}	0.0082	9.49×10^{-5}	0.0087

Let N be the number of variables included in the VAR model. Choose $\lambda = 9 - N$ and $\lambda = 35 - 5N$ when extracting the informative variables in Y_t and X_t, respectively.

We next evaluate the algorithms in terms of their predictability based on the extracted informative variables by using the data from January 2011 to September 2011. Specifically, we compare their accuracy in forecasting two common variables in $\tilde{X}_t$, viz., $X_{1,t}$ and $X_{2,t}$ (see Table 8). To carry out this comparison, the following two performance measures, called the mean squared error of prediction (MSEP) and the mean absolute error of prediction (MAEP), are considered:

$$\text{MSEP} = \frac{1}{M}\sum_{h=1}^{M}\left[X_{i,t+h} - \hat{X}_{i,t}(h|\Omega_{\tilde{X}_t\tilde{Y}_t})\right]^2$$

$$\text{MAEP} = \frac{1}{M}\sum_{h=1}^{M}\left|X_{i,t+h} - \hat{X}_{i,t}(h|\Omega_{\tilde{X}_t\tilde{Y}_t})\right|$$

Note that M is the number of observations we wish to forecast in the future (here $M = 9$), whereas $\Omega_{\tilde{X}_t\tilde{Y}_t}$ is the set of all informative variables extracted by a particular algorithm based on current observations. The numerical results are given in Table 9.

As can be seen from Table 9, for this particular data set the forward selection algorithm clearly outperforms the other two algorithms in forecasting $X_{1,t+h}$ and $X_{2,t+h}$. This supports that if one wishes to forecast the growth rates of "Exports of Goods" and "Imports of Goods", it suffices to utilize merely the information of "Consumer Price Index", "New Orders for Manufactured Goods", and " ISM Manufacturing Index". In summary, compared to the method that includes all variable information, the three algorithms perform well in terms of both the MSEP and MAEP.

5 Conclusions

In this article we introduced various computer-search algorithms so that a hypothesis testing procedure can be utilized to extract informative variables in the validation of causal relationship between two groups of time series. The result is useful in the sense that, it allows us to forecast the future quantity of explicit variables by utilizing the minimum data information. The numerical results show that the algorithms considered in this study have fairly high accuracy in forecasting the future quantity of selected variables. There are some issues we would like to address here. First, it should be mentioned that the algorithms used in this study may fail (i.e., $\tilde{Y}_t = \tilde{X}_t = \emptyset$, the readers can refer to Makridakis et al. (1983) for an example that the forward stepwise algorithm fails). To overcome this problem, we can consider the following two strategies: (i) increase the values of α-to-enter and/or α-to-remove; (ii) utilize other search algorithms or perform an exhaustive search of all possible subsets of variables. If none of the strategies work, it is possible that there is no causal relationship between the two groups of variables. Second, one may suspect that which information sets $\tilde{Y}_t$ and $\tilde{X}_t$ obtained from the algorithms are the "best". This answer is, in fact, subject to cases. One possible solution is to consider the accuracy of forecasting in some selected variables. To illustrate, let us recall the numerical results in Table 9. Suppose now we are more interested in forecasting the growth rates of "Exports of Goods" ($X_{1,t}$) and "Imports of Goods" ($X_{2,t}$), then the informative variables extracted by the forward selection algorithm should work better (since they result in the smallest values of MSEP and MAEP). Finally, we are currently investigating the computational cost when the number of variables becomes large, and how to extract informative variables based on a suitable hypothesis testing procedure for non-stationary processes.

References

Arnold A, Liu Y, Abe N (2007) Temporal causal modeling with graphical Granger methods. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, pp 66–75

Boudjellaba H, Dufour JN, Roy R (1992) Testing causality between two vectors in multivariate autoregressive moving average models. J Am Stat Assoc 87:1082–1090

Dufour JM, Renault E (1998) Short-run and long-run causality in time series theory. Econometrica 66: 1099–1125

Fujita A, Sato JR, Garay-Malpartida HM, Morettin PA, Sogayar MC, Ferreira CE (2007) Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. Bioinformatics 23:1623–1630

Geweke J (1982) Measurement of linear dependence and feedback between multiple time series. J Am Stat Assoc 77:304–313

Geweke J (1984) Inference and causality in economic time series. In: Griliches Z, Intriligator MM (eds) Handbook of econometrics, vol 2. North-Holland, Amsterdam, pp 1101–1144

Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438

Granger CWJ (1980) Testing for causality: a personal viewpoint. J Econ Dyn Control 2:329–352

Granger CWJ, Lin JL (1995) Causality in the long run. Econ Theory 11:530–536

Hacker RS, Hatemi JA (2006) Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application. Appl Econ 38:1489–1500

Haufe S, Müller K-R, Nolte G, Krämer N (2010) Sparse causal discovery in multivariate time series. In: NIPS 2008 workshop on causality. JMLR workshop and conference proceedings, vol 6, pp 97–106

Hsiao C (1982) Autoregressive modeling and causal ordering of econometric variables. J Econ Dyn Control 4:243–259

Koster JTA (1996) Markov properties of nonrecursive causal models. Ann Stat 24:2148–2177

Koster JTA (1999) On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. Scand J Stat 26:413–431

Kutner MH, Nachtsheim CJ, Neter J (2008) Applied linear regression models, 4th edn. McGraw Hill, New York

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lauritzen SL (2000) Causal inference from graphical models. In: Barndorff-Nielsen E, Cox DR, Klüppelberg C (eds) Complex stochastic systems. CRC Press, London

Lütkepohl H (2005) New introduction to multiple time series analysis, 1st edn. 2nd printing, Springer, Berlin

Lütkepohl H, Burda MM (1997) Modified Wald tests under nonregular conditions. J Econ 78:315–332

Makridakis SG, Wheelwright SC, McGee VE (1983) Forecasting: methods and applications. Wiley, New York

Mosconi R, Giannine C (1992) Non-causality in cointegrated system: representation. Estimation and testing. Oxf Bull Econ Stat 54:399–417

Osborn DR (1984) Causality testing and its implication for dynamic econometric models. Econ J 94:82–96

Pearl J (1995) Causal diagrams for empirical research (with discussion). Biometrika 82:669–710

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

Roebroech A, Formisano E, Goebel R (2005) Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage 25:230–242

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester