Electronic Journal of Statistics
Vol. 8 (2014) 1940–1972
ISSN: 1935-7524
DOI: 10.1214/14-EJS940

Trimmed Granger causality between two groups of time series

Ying-Chao Hung[*]

Department of Statistics
National Chengchi University
Taipei 11605
Taiwan
e-mail: hungy@nccu.edu.tw

Neng-Fang Tseng

Department of Mathematical Statistics and Actuarial Science
Aletheia University
Taipei 25103
Taiwan
e-mail: au4225@mail.au.edu.tw

and

Narayanaswamy Balakrishnan

Department of Mathematics and Statistics
McMaster University
Hamilton
Ontario
Canada L8S 4K1
e-mail: bala@mcmaster.ca

Abstract: The identification of causal effects between two groups of time series has been an important topic in a wide range of applications such as economics, engineering, medicine, neuroscience, and biology. In this paper, a simplified causal relationship (called trimmed Granger causality) based on the context of Granger causality and vector autoregressive (VAR) model is introduced. The idea is to characterize a subset of "important variables" for both groups of time series so that the underlying causal structure can be presented based on minimum variable information. When the VAR model is specified, explicit solutions are provided for the identification of important variables. When the parameters of the VAR model are unknown, an efficient statistical hypothesis testing procedure is introduced to estimate the solution. An example representing the stock indices of different countries is used to illustrate the proposed methods. In addition, a simulation study shows that the proposed methods significantly outperform the Lasso-type methods in terms of the accuracy of characterizing the simplified causal relationship.

AMS 2000 subject classifications: 62M10, 62F03.

[*] Corresponding author.

Keywords and phrases: Granger causality, multivariate time series, vector autoregressive process, Wald test with constrained parameter spaces, multiple hypothesis testing, Lasso-penalized VAR approach.

Received December 2013.

Contents

1. Introduction

Over the years, the causality system described by a multivariate time series process has been one of the most flexible and popular statistical techniques used for measuring the dynamic relationships between groups of variables in such diverse fields as economics, engineering, medicine, neuroscience, and biology. A pioneering study of causal relationships dates back to the work of Granger [12], wherein a 2-variate autoregressive model was used to identify the "causality" between two time series based on precedence and predictability. Afterwards, a fairly rich literature has emerged on this topic by utilizing the vector autoregressive (VAR) model - a general macroeconometric framework introduced by Sims [34] and shown promising in describing the temporal dependence in multiple time series. For example, Hsiao [17] introduced different types of causal relationships for a 3-variate VAR model. Osborn [27] discussed the "Unidirectional Granger Causality" based on the VAR model with MA errors and adopted it in statistical hypothesis testing procedures. Geweke [10, 11] considered measures of linear dependence and feedback between multiple time series data and provided a comprehensive survey of the literature on Granger causality. Boudjellaba et al. [5] tested Granger causality between two vectors in multivariate ARMA models. Mosconi and Giannine [26] investigated the Granger causality based on a non-stationary VAR model. Roebroech et al. [29] used the Granger causality mapping (GCM) to explore directed influences between neuronal populations in a fMRI data. While Hacker and Hatemi-J [13] developed a method that is not sensitive to deviations from the assumption of normal error, Fujita et al. [9] proposed an improved VAR model (called DVAR) for estimating time-varying gene regulatory networks based on gene expression profiles obtained from microarray experiments. Shojaie and Michailidis [30] proposed truncating lasso penalty to estimate causal relationships from time-course gene expression data.

In this work, we focus on the notion of causality in multivariate stochastic processes introduced by Granger [12] and Dufour and Renault [8]. Then by definition, the so-called Granger causality between two groups of time series indicates that the past values of variables in one group can help predict the future values of variables in the other group. It is shown that the existence of Granger causality can be validated by examining a designated subset of the VAR coefficients. For large VAR models where some of the coefficients are zero/insignificant (the so-called spare VAR model), the result of the Granger causality test can be misinterpreted since some variables in the model may not contribute to the prediction of variables of interest (i.e., some variables are not Granger-causal for the variables of interest). Obviously, these variables are "unimportant" in describing the underlying causal structure and may mislead the causal inference. Therefore, our goal here is to identify the unimportant variables in the primarily introduced VAR model so that a simplified causal relationship, called *trimmed Granger causality*, can be presented by a small subset of "important variables".

The estimation of coefficients for large VAR models appears to be a challenging task due to a limited number of stationary observations. A straightforward way to tackle this problem is to utilize the idea of l_1-penalized regression (Lasso) so that some of the VAR coefficients are shrunk to zero (Tibshirani [36]). The Lasso-penalized VAR modeling approach and its variants (Arnold et al. [1], Song and Bickel [35], Davis et al. [7], Basu and Michailidis [3]) provide a convenient tool for solving the desired variable-selection problem. However, this type of approaches have the disadvantage that the accuracy of the estimated important variables (or the estimated trimmed Granger causality) can not be satisfactorily controlled based on finite samples (see Section 5.2 for numerical illustrations). To overcome this issue, we propose an alternative framework based on which the important variables are identified by validating a class of designated constraints on the VAR coefficients throughout a sequential hypothesis testing procedure. By setting appropriate bounds on the type I error rates of the associated tests, the proposed framework allows us to better control the accuracy of the estimated important variables without requiring much computational cost.

The rest of this paper is organized as follows. Section 2 introduces the VAR model, assumptions, some basic concepts of projection theory, and preliminaries needed for establishing the desired causal relationship. Section 3 depicts how the important variables are defined so as to form the trimmed Granger causality when the VAR model is specified. Theoretical basis that facilitates the hypothesis testing procedure introduced in Section 4 is also established. Section 4 presents a hypothesis testing procedure for identifying the important variables when the VAR coefficients are unknown and so need to be estimated based on observed data. The presented procedure has many stages involving efficient search algorithms (i.e., the backward and forward search), the Wald tests with constrained parameter spaces, bootstrap sampling, and control of type I error rates for multiple testing. Section 5 illustrates the proposed procedure and discusses the power of the associated Wald tests using a real example. A simulation study is also carried out to compare the accuracy of the proposed search methods with that of the Lasso-penalized VAR approach. Finally, some concluding remarks are made in Section 6.

2. Preliminaries

Many different definitions of causal relationships have been proposed in the literature. In this work, we mainly focus on the notion of causality in multivariate stochastic processes introduced by Granger [12] and Dufour and Renault [8]. We shall introduce it briefly now. Consider a measurable space on which the set of square integrable L^2-functions form a vector space, called the L^2-space. The "causality" discussed here is defined in terms of projection of an L^2-space onto the Hilbert space with respect to a probability measure μ, where for any integrable L^2-functions f and g, the "inner product" in the Hilbert space is defined as

$$< f, g >= \int_X fg d\mu = E(fg).$$

Consider a random vector $W = (W_1, W_2, \ldots, W_K)'$ with each element $W_i \in L^2$, and let I_W be the *linear manifold* spanned by all $W_1, \ldots, W_K$. For any given random variable X in L^2, a reasonable quantity to estimate X is the projection onto I_W such that the mean squared error (M.S.E.) can be minimized; that is, to consider the following estimator

$$P_{I_W}(X) = \arg\min_{c \in I_W} E(X - c)^2.$$

By definition, $P_{I_W}(X)$ is also the *best linear predictor* of X in I_W. The same notation can also be used if X represents a random vector.

Suppose there are two multivariate time series $X_t = (X_{1,t}, \ldots, X_{n,t})'$ and $Y_t = (Y_{1,t}, \ldots, Y_{m,t})'$, which were possibly identified by economic theory, expert knowledge, and experience. 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. As shown in the literature, this type of study largely relies on the pth-order Vector Autoregression (VAR) model of the form

$$W_t = c + \sum_{j=1}^{p} A_j W_{t-j} + a_t, \tag{2.1}$$

where c is a $K \times 1$ constant vector, $K = n + m$, $W_t = \text{vec}(X_t, Y_t) = (W_{1,t}, \ldots, W_{K,t})'$ is a $K \times 1$ random vector ("vec" is the column stacking operator), 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 $E(a_t) = \mathbf{0}$, (ii) the covariance matrix $E(a_t a_t') = \Sigma_a$ is positive definite (thus non-singular), and (iii) $E(a_t a'_{t-k}) = \mathbf{0}$ for any non-zero k. Since W_t is composed of X_t and Y_t, (2.1) can be further represented as

$$W_t = \begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} c_X \\ c_Y \end{pmatrix} + \sum_{j=1}^{p} \begin{pmatrix} A_{XX,j} & A_{XY,j} \\ A_{YX,j} & A_{YY,j} \end{pmatrix} \begin{pmatrix} X_{t-j} \\ Y_{t-j} \end{pmatrix} + \begin{pmatrix} a_{X,t} \\ a_{Y,t} \end{pmatrix}, \tag{2.2}$$

where c_X and c_Y 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 of orders $n \times n$, $n \times m$, $m \times n$, and $m \times m$, respectively, and $a_{X,t}$ and $a_{Y,t}$ are $n \times 1$ and $m \times 1$ error vectors.

Define two *information sets* based on all the observations up to time t as

$$\Omega_{XY} = \{(X_s, Y_s) : s \leq t\}$$

and

$$\Omega_X = \{X_s : s \leq t\}.$$

Let I_{XY} and I_X be the *linear manifolds* spanned by all the random variables in Ω_{XY} and Ω_X, respectively. For any given future time $(t+h)$, we denote the best linear predictors 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. Then, according to the definitions introduced earlier, it is clear that

$$\hat{X}_t(h|\Omega_{XY}) = P_{I_{XY}}(X_{t+h}) \quad \text{and} \quad \hat{X}_t(h|\Omega_X) = P_{I_X}(X_{t+h}).$$

The two-group causality, also known as a generalization of Granger causality, is defined as follows.

Definition 2.1 (Two-group Causality/Non-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 $Y \underset{(c)}{\to} X$. 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 $Y \underset{(c)}{\nrightarrow} X$.

Remark 1.
(a) In practice, the value of c in Definition 2.1 is chosen by the user/designer. If c is chosen to be ∞, the two-group causality is known as the Granger causality;
(b) It is clear that if $Y \underset{(c)}{\to} X$, 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 i-th elements of $\hat{X}_t(h|\Omega_{XY})$ and $\hat{X}_t(h|\Omega_X)$, respectively.

Fact 2.1. Based on the model in (2.1) and (2.2), for any positive integer c, we have $Y \underset{(c)}{\nrightarrow} X$ if and only if $A_{XY,j} = \mathbf{0}_{n\times m}$ for all $j = 1, \dots, p$.

Proof. Since $Y \underset{(c)}{\nrightarrow} X$ is equivalent to $Y \underset{(\infty)}{\nrightarrow} X$ (see Dufour and Renault [8], Proposition 2.3) and $Y \underset{(\infty)}{\nrightarrow} X$ if and only if $A_{XY,j} = \mathbf{0}_{n\times m}$ for all $j = 1, \dots, p$ (see Lütkepohl [23], Corollary 2.2.1), the result follows immediately. □

Fact 2.1 indicates that for any given prediction horizon c, the existence of Granger causality between Y_t and X_t can be validated by examining merely the coefficient matrix $A_{XY,j}$. However, as pointed by Lütkepohl [21], the result does not hold if the information set is changed (e.g., adding or removing variables from the information set). Since changing the information set is the main theme of our proposed methods shown later in Section 4, it is then necessary to examine the coefficient matrix $A^{(h)}_{XY,j}$ associated with the h-step predictor for all $h \le c$ (see (2.7) for details). We now review some important properties that are essential for establishing the theoretical results in Section 3.

Lemma 2.1. *Let I_1 and I_2 be two subspaces of a linear manifold I. Then, we have:*
(a) for any random variable $X \in I_1$, $P_{I_1}(X) = X$;
(b) for any random variable $X \in I$, $P_{I_1}(X) = P_{I_1}[P_{I_2}(X)]$ if and only if $I_1 \subset I_2$.

Proof. For a proof, one may refer to Berberian (1961) [4]. □

Lemma 2.2. *Consider the VAR(p) model described in (2.1), and let I_W be the linear manifold spanned by all the variables in $\Omega_W = \{(W_{1,s}, \dots, W_{K,s}) : s \le t\}$. For any random variable $V \in I_W$, there exist a sequence of constant c_0^n and row vectors $c_1^n, c_2^n, \dots$ such that*

$$c_0^n + \sum_{i=1}^{n} c_i^n W_{t+1-i} \xrightarrow{L^2} V \ \ \textit{as } n \to \infty, \tag{2.3}$$

where "$\xrightarrow{L^2}$" stands for "convergence in quadratic mean". Let $\lim_{n\to\infty} c_i^n = c_i$ for $i = 0, 1, \dots,$ then $c_0, c_1, \dots$ are unique.

Proof. The existence of $c_0^n, c_1^n, \dots$ satisfying (2.3) comes directly from the result of projection theorem (Luenberger [20], p. 51). To prove that $c_0, c_1, \dots$ are uniquely determined, we assume there exists another sequence of constant b_0^n and row vectors $b_1^n, b_2^n, \dots$ such that

$$b_0^n + \sum_{i=1}^{n} b_i^n W_{t+1-i} \xrightarrow{L^2} V \text{ as } n \to \infty, \tag{2.4}$$

where $\lim_{n\to\infty} b_i^n = b_i$ for $i = 0, 1, \dots$. Upon subtracting (2.4) from (2.3), we have

$$(c_0^n - b_0^n) + \sum_{i=1}^{n} (c_i^n - b_i^n) W_{t+1-i} \xrightarrow{L^2} 0 \text{ as } n \to \infty, \tag{2.5}$$

and thus

$$E\left[(c_0^n - b_0^n) + \sum_{i=1}^{n}(c_i^n - b_i^n)W_{t+1-i}\right] \to 0 \text{ as } n \to \infty. \tag{2.6}$$

Recall that a_t represents the white noise in the VAR(p) model. Multiplying a_t on the left hand side of (2.5) and applying Cauchy-Schwarz inequality for expectations, we have

$$E\left(\left[(c_0^n - b_0^n) + \sum_{i=1}^{n}(c_i^n - b_i^n)W_{t+1-i}\right]a_t'\right) \to \mathbf{0} \text{ as } n \to \infty.$$

Since $E(a_t') = \mathbf{0}$ and $E(W_{t+1-i}a_t') = \mathbf{0}$ for all $i \geq 2$, we then have

$$\lim_{n\to\infty}(c_1^n - b_1^n)E(W_t a_t') = \mathbf{0},$$

and thus

$$\lim_{n\to\infty}(c_1^n - b_1^n)E(a_t a_t') = \mathbf{0}.$$

Since $E(a_t a_t')$ is a non-singular matrix, we have $\lim_{n\to\infty}(c_1^n - b_1^n) = \mathbf{0}$, and thus $c_1 = b_1$. Taking the result back to (2.6), we have a new expression

$$E\left[(c_0^n - b_0^n) + \sum_{i=2}^{n}(c_i^n - b_i^n)W_{t+1-i}\right] \to 0 \text{ as } n \to \infty.$$

Continuing in this fashion, we can show $b_i = c_i$ for all $i = 0, 1, 2, \ldots$, that is, all c_i are uniquely determined. □

Note that for any given time lag $h > 0$, (2.1) can be further represented as

$$W_{t+h} = \sum_{k=0}^{h-1} A_1^{(k)}(c + a_{t+h-k}) + \sum_{j=1}^{p} A_j^{(h)} W_{t+1-j}, \tag{2.7}$$

where $A_1^{(0)} = \mathbf{I}_{m+n}$ is the identity matrix of order $m+n$, and $A_j^{(k)}$ is a matrix obtained from the recursive formula

$$A_j^{(k)} = \begin{cases} A_j, & k = 1, \\ A_{j+1}^{(k-1)} + A_1^{(k-1)} A_j, & k = 2, 3, \ldots, h \end{cases} \tag{2.8}$$

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

$$A_j^{(h)} = \begin{pmatrix} A_{XX,j}^{(h)} & A_{XY,j}^{(h)} \\ A_{YX,j}^{(h)} & A_{YY,j}^{(h)} \end{pmatrix}, \tag{2.9}$$

where $A_{XX,j}^{(h)}$ and $A_{XY,j}^{(h)}$ are two sub-matrices of orders $n \times n$ and $n \times m$, respectively. Denoting the identity matrix of order n by $\mathbf{I}_n$, then $X_{t+h} = (\mathbf{I}_n, \mathbf{0}_{n\times m})W_{t+h}$. The following lemma yields the best linear predictor of X_{t+h} (in matrix form) based on the information set Ω_{XY}.

Lemma 2.3. *Based on (2.1) and (2.2), we have*

$$\hat{X}_t(h|\Omega_{XY}) = c_{X,h} + \sum_{j=1}^{p}(A_{XX,j}^{(h)} X_{t+1-j} + A_{XY,j}^{(h)} Y_{t+1-j}), \tag{2.10}$$

where $c_{X,h} = (\mathbf{I}_n, \mathbf{0}_{n\times m}) \sum_{k=0}^{h-1} A_1^{(k)} c$.

Proof. Since the result is similar to that of Dufour and Renault [8], the proof is omitted for the sake of brevity. □

Lemma 2.3 shows that the best linear predictor of X_{t+h} relates to Y_t only through the coefficient matrix $A_{XY,j}^{(h)}$. This will serve as a benchmark for the rest of this work.

3. Identification of important variables when model is specified

Based on the discussions in Section 2, the existence of Granger causality describes that Y_t can improve the prediction of X_{t+h} for some finite lag h. However, from Remark 1(b), we learn that if $Y \underset{(c)}{\rightarrow} X$, then it is guaranteed that adding all variables in Y_t into the information set will improve the prediction of "some" variables in X_t – but not necessarily all. On the other hand, the prediction 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 formal procedure to extract those "important variables" in both X_t and Y_t so that a trimmed causal relationship can be presented. The following example illustrates this idea.

Consider two groups of variables $Y_t = (Y_{1,t}, Y_{2,t}, Y_{3,t})'$ and $X_t = (X_{1,t}, X_{2,t}, X_{3,t})'$, and assume that $Y \underset{(c)}{\rightarrow} X$ for some integer $c > 0$. Fig. 1 shows three possible structures that are characterized as having the same causal relationship $Y \underset{(c)}{\rightarrow} X$. Structure (a) in Fig. 1 indicates that variable $Y_{3,t}$ has no influence on the prediction of all the variables in X_t. Therefore, it is reasonable to present a simplified causal relationship between $(Y_{1,t}, Y_{2,t})$ and X_t by excluding $Y_{3,t}$ from the analysis. Structure (b) in Fig. 1 indicates that the prediction of $X_{1,t}$ is not influenced by any of the variables in Y_t. Therefore, $X_{1,t}$ can be excluded from the analysis so that a simplified causal relationship between Y_t and $(X_{2,t}, X_{3,t})$ can be presented. Similarly, Structure (c) in Fig. 1 indicates that a simplified causal relationship between $(Y_{1,t}, Y_{2,t})$ and $(X_{2,t}, X_{3,t})$ can be presented by excluding $X_{1,t}$ and $Y_{3,t}$ from the analysis. We shall now introduce a formal definition of "important variables" in both X_t and Y_t.

Definition 3.1 (Important Variables in X_t and Y_t)**.** Consider the model expressed in (2.1) and (2.2), and assume that $Y \underset{(c)}{\rightarrow} X$ for some given integer $c > 0$.

(a) The set of important variables in $Y_t = (Y_{1,t}, \ldots, Y_{m,t})'$ is defined as

$$S_Y = \{Y_{i,t} \in Y_t : \hat{X}_t(h|\Omega_{XY}) \neq \hat{X}_t(h|\Omega_{XY_{-i}}) \text{ for some } h \leq c\}, \tag{3.1}$$

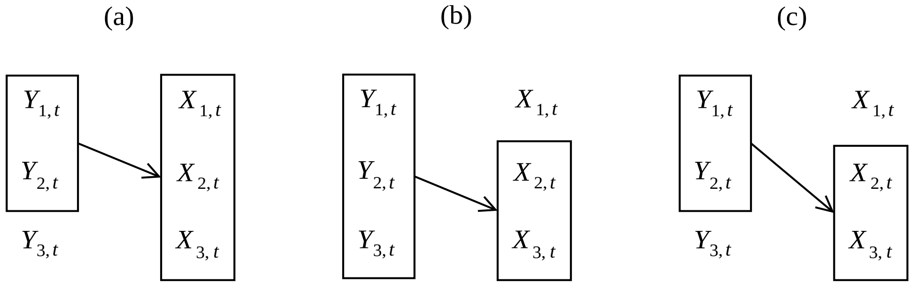

FIG 1. *Three possible causal structures that are characterized as having the same causal relationship* $Y \underset{(c)}{\rightarrow} X$.

where $\Omega_{XY_{-i}} = \Omega_{XY} \setminus \{Y_{i,t}, Y_{i,t-1}, \ldots\}$ refers to a reduced information set with the i-th variable in Y_t being excluded;
(b) The set of important variables in $X_t = (X_{1,t}, \ldots, X_{n,t})'$ is similarly defined as

$$S_X = \{X_{i,t} \in X_t : \hat{X}_{i,t}(h|\Omega_{XY}) \neq \hat{X}_{i,t}(h|\Omega_X) \text{ for some } h \leq c\}. \tag{3.2}$$

Definition 3.1 states 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 $Y_{i,t}$ is characterized as an *unimportant variable* in Y_t and can be excluded from the analysis. On the other hand, if the prediction of $X_{i,t+h}$ based on Ω_{XY} is the same as that based on Ω_X, then $X_{i,t}$ is characterized as an *unimportant variable* in X_t and can be excluded from the analysis. Then, based on the two sets of identified important variables S_Y and S_X, we can present the following *trimmed Granger causality*:

$$S_Y \underset{(c)}{\rightarrow} S_X. \tag{3.3}$$

Remark 2. Note that by Definition 2.1, $S_Y = \emptyset$ implies that $Y \underset{(c)}{\nrightarrow} X$ (i.e., Y_t does not cause X_t).

We shall now introduce some useful guidelines for finding the two important sets S_Y and S_X.

Theorem 3.1 (Identification of S_Y)**.** *Consider the matrix $A^{(h)}_{XY,j}$ given in (2.9), and its column partition as*

$$A^{(h)}_{XY,j} = (A^{(h)}_{XY,j}(:,1), A^{(h)}_{XY,j}(:,2), \ldots, A^{(h)}_{XY,j}(:,m)), \tag{3.4}$$

where $A^{(h)}_{XY,j}(:,i)$ refers to the i-th column of $A^{(h)}_{XY,j}$. Then, for any given $i \in \{1, \ldots, m\}$, $Y_{i,t} \in S_Y$ if and only if there exists at least one pair $(h, j) \in \{1, \ldots, c\} \times \{1, \ldots, p\}$ such that $A^{(h)}_{XY,j}(:,i) \neq \mathbf{0}$.

Proof. First, we show that $Y_{i,t} \in S_Y$ is a sufficient condition for the desired result. Note that by (2.10), $\hat{X}_t(h|\Omega_{XY})$ can be expressed as

$$\hat{X}_t(h|\Omega_{XY}) = c_{X,h} + \sum_{j=1}^{p} A_{XX,j}^{(h)} X_{t+1-j} + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,1) Y_{1,t+1-j} + \cdots + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,m) Y_{m,t+1-j}. \tag{3.5}$$

Suppose $Y_{i,t} \in S_Y$, and given any h we assume that $A_{XY,j}^{(h)}(:,i) = \mathbf{0}$ for all $j \in \{1, \ldots, p\}$. Then, we have

$$\begin{aligned}\hat{X}_t(h|\Omega_{XY}) = c_{X,h} &+ \sum_{j=1}^{p} A_{XX,j}^{(h)} X_{t+1-j} + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,1) Y_{1,t+1-j} + \cdots \\ &+ \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,i-1) Y_{i-1,t+1-j} + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,i+1) Y_{i+1,t+1-j} \\ &+ \cdots + \sum_{j=1}^{p} A_{XY,j}^{(h)}(:,m) Y_{m,t+1-j}.\end{aligned} \tag{3.6}$$

Equation (3.6) implies that each element of $\hat{X}_t(h|\Omega_{XY})$ belongs to the linear manifold $I_{XY_{-i}}$ spanned by all the random variables in $\Omega_{XY_{-i}}$. The result of Lemma 2.1 then yields

$$\begin{aligned}\hat{X}_t(h|\Omega_{XY_{-i}}) &= P_{I_{XY_{-i}}}(X_{t+h}) = P_{I_{XY_{-i}}}\left(P_{I_{XY}}(X_{t+h})\right) \\ &= P_{I_{XY_{-i}}}\left(\hat{X}_t(h|\Omega_{XY})\right) = \hat{X}_t(h|\Omega_{XY}),\end{aligned}$$

which leads to a contradiction with the primary assumption that $Y_{i,t} \in S_Y$. Therefore, we conclude that there must exist at least one pair $(h,j) \in \{1, \ldots, c\} \times \{1, \ldots, p\}$ such that $A_{XY,j}^{(h)}(:,i) \neq \mathbf{0}$. Next, we show that $Y_{i,t} \in S_Y$ is a necessary condition for the desired result. Suppose there exists one pair $(h,j) \in \{1, \ldots, c\} \times \{1, \ldots, p\}$ such that $A_{XY,j}^{(h)}(:,i) \neq \mathbf{0}$, and we assume that $Y_{i,t} \notin S_Y$ (i.e., $\hat{X}_t(h|\Omega_{XY}) = \hat{X}_t(h|\Omega_{XY_{-i}})$). Since $\hat{X}_t(h|\Omega_{XY_{-i}})$ can be analogously expressed as

$$\begin{aligned}\hat{X}_t(h|\Omega_{XY_{-i}}) = b_{X,h} &+ \sum_{j=1}^{p} B_{XX,j}^{(h)} X_{t+1-j} + \sum_{j=1}^{p} B_{XY,j}^{(h)}(:,1) Y_{1,t+1-j} + \cdots \\ &+ \sum_{j=1}^{p} B_{XY,j}^{(h)}(:,i-1) Y_{i-1,t+1-j} + \sum_{j=1}^{p} B_{XY,j}^{(h)}(:,i+1) Y_{i+1,t+1-j} \\ &+ \cdots + \sum_{j=1}^{p} B_{XY,j}^{(h)}(:,m) Y_{m,t+1-j},\end{aligned} \tag{3.7}$$

upon subtracting (3.7) from (3.5), we obtain

$$\begin{aligned}\mathbf{0} = (c_{X,h} - b_{X,h}) &+ \sum_{j=1}^{p} (A^{(h)}_{XX,j} - B^{(h)}_{XX,j}) X_{t+1-j} \\ &+ \sum_{j=1}^{p} (A^{(h)}_{XY,j}(:,1) - B^{(h)}_{XY,j}(:,1)) Y_{1,t+1-j} + \cdots + \sum_{j=1}^{p} (A^{(h)}_{XY,j}(:,i) - \mathbf{0}) Y_{i,t+1-j} \\ &+ \cdots + \sum_{j=1}^{p} (A^{(h)}_{XY,j}(:,m) - B^{(h)}_{XY,j}(:,m)) Y_{m,t+1-j}\end{aligned}$$

almost surely. By Lemma 2.2, we then obtain $A^{(h)}_{XY,j}(:,i) = \mathbf{0}$ for all $j \in \{1,\dots,p\}$, which leads to a contradiction with the assumption that there exists one pair $(h,j) \in \{1,\dots,c\} \times \{1,\dots,p\}$ such that $A^{(h)}_{XY,j}(:,i) \neq \mathbf{0}$. Hence, we must have $Y_{i,t} \in S_Y$, which completes the proof. □

The following theorem describes an important property of S_Y, which we state without showing the detailed proof.

Theorem 3.2. *Let V_Y be any subset of $\{Y_{1,t},\dots,Y_{m,t}\}$, then $S_Y \subseteq V_Y$ if and only if*

$$\hat{X}_t(h|\Omega_{XY}) = \hat{X}_t(h|\Omega_{XV_Y}) \quad \textit{for all } 1 \le h \le c. \tag{3.8}$$

Remark 3.

(a) Any subset V_Y of $\{Y_{1,t},\dots,Y_{m,t}\}$ that satisfies (3.8) is called an *important set* in the validation of the desired causal relationship. Based on Theorem 3.2, we have

$$S_Y = \inf\{V_Y : V_Y \text{ is an important set}\} = \cap(\text{all important sets } V_Y), \tag{3.9}$$

which is clearly the *minimal important set* that provides the same information as Y_t does for predicting X_{t+h}. As shown later in Section 4, this property is useful for finding S_Y when the VAR coefficients are unknown.

(b) Combining the results of Theorems 3.1 and 3.2, we get

$$\begin{aligned}\hat{X}_t(h|\Omega_{XY}) &= \hat{X}_t(h|\Omega_{XS_Y}) \\ &= c_{1,h} + \sum_{j=1}^{p} A^{(h)}_{XX,j} X_{t+1-j} + \sum_{j=1}^{p} A^{(h)}_{XY,j}(:,s_1) Y_{s_1,t+1-j} \\ &\quad + \sum_{j=1}^{p} A^{(h)}_{XY,j}(:,s_2) Y_{s_2,t+1-j} + \cdots + \sum_{j=1}^{p} A^{(h)}_{XY,j}(:,s_q) Y_{s_q,t+1-j},\end{aligned}$$

where $\{s_1,s_2,\dots,s_q\}$ is a sub-sequence of $\{1,2,\dots,m\}$, and $A^{(h)}_{XY,j}(:,s) \neq \mathbf{0}$ for all $s \in \{s_1,s_2,\dots,s_q\}$. Thus, we have $S_Y = \{Y_{s_1,t}, Y_{s_2,t},\dots,Y_{s_q,t}\}$.

Theorem 3.3 (Identification of S_X)**.** *Consider the matrix $A^{(h)}_{XY,j}$ given in (2.9), and its row partition as*

$$A^{(h)}_{XY,j} = \begin{pmatrix} A^{(h)}_{XY,j}(1,:) \\ A^{(h)}_{XY,j}(2,:) \\ \vdots \\ A^{(h)}_{XY,j}(n,:) \end{pmatrix}, \tag{3.10}$$

where $A^{(h)}_{XY,j}(i,:)$ refers to the i-th row of $A^{(h)}_{XY,j}$. Then, for any given $i \in \{1,\ldots,n\}$, $X_{i,t} \in S_X$ if and only if there exists at least one pair $(h,j) \in \{1,\ldots,c\} \times \{1,\ldots,p\}$ such that $A^{(h)}_{XY,j}(i,:) \neq \mathbf{0}$.

Proof. From (2.10), the i-th element of $\hat{X}_t(h|\Omega_{XY})$ can be expressed as

$$\hat{X}_{i,t}(h|\Omega_{XY}) = c_{X,h}(i) + \sum_{j=1}^{p} A^{(h)}_{XX,j}(i,:)X_{t+1-j} + \sum_{j=1}^{p} A^{(h)}_{XY,j}(i,:)Y_{t+1-j},$$

where $c_{X,h}(i)$ is the i-th element of vector $c_{X,h}$. As in the proof of Theorem 3.1, we can simply obtain $\hat{X}_{i,t}(h|\Omega_{XY}) \neq \hat{X}_{i,t}(h|\Omega_X)$ for some $h \le c$ if and only if $A^{(h)}_{XY,j}(i,:) \neq \mathbf{0}$ for at least one pair $(h,j) \in \{1,\ldots,c\} \times \{1,\ldots,p\}$. The result thus follows. □

Theorem 3.4. *The set S_X can be equivalently expressed as*

$$S_X = \{X_{i,t} \in X_t : \hat{X}_{i,t}(h|\Omega_{XS_Y}) \neq \hat{X}_{i,t}(h|\Omega_X) \textit{ for some } h \le c\}. \tag{3.11}$$

Proof. The result follows immediately from Definition 3.1(b) and Remark 3(b). □

Theorem 3.4 indicates that S_X can be obtained by first identifying the set S_Y. The following is a dual result of Theorem 3.2, which we present without a proof.

Theorem 3.5. *Let $\bar{V}_X$ be any subset of $\{X_{1,t},\ldots,X_{n,t}\}$. Then, $S_X \subseteq \{X_{1,t},\ldots,X_{n,t}\} \setminus \bar{V}_X$ if and only if*

$$\hat{X}_{i,t}(h|\Omega_{XY}) = \hat{X}_{i,t}(h|\Omega_X) \quad \textit{for all } X_{i,t} \in \bar{V}_X \textit{ and } 1 \le h \le c. \tag{3.12}$$

Remark 4. If $\bar{V}_X$ is any subset of $\{X_{1,t},\ldots,X_{n,t}\}$ that satisfies (3.12), then $V_X = \{X_{1,t},\ldots,X_{n,t}\} \setminus \bar{V}_X$ is called an *important set* in the validation of the desired causal relationship. Therefore, we have

$$S_X = \inf\{V_X : V_X \text{ is an important set}\} = \cap(\text{all important sets } V_X), \tag{3.13}$$

which is clearly the *minimal important set* so that the prediction of each variable $X_{i,t}$ in the set is influenced by Y_t (or S_Y). As shown later in Section 4, this property is useful for finding S_X when the VAR coefficients are unknown.

4. Identification of important variables: A hypothesis testing procedure

We have shown that, by examining the coefficient matrices $A^{(h)}_{XY,j}$, the important variables in the trimmed causal relationship can be explicitly identified. However, in practice, the parameters in $A^{(h)}_{XY,j}$ are usually unknown and need to be estimated. Thus, to best identify the causal relationship, one can resort to a statistical hypothesis testing procedure. A naive approach is to perform an exhaustive search on finding all possible important sets V_X and V_Y, and then identify the two minimal important sets S_X and S_Y by using (3.9) and (3.13). This means that for each $1 \leq h \leq c$, one needs to conduct $2^m \times 2^n = 2^{m+n}$ possible hypothesis tests for validating the associated constraints on the elements of the coefficient matrix $A^{(h)}_{XY,j}$. Obviously, such a procedure will become computationally involved (even for small values of m and n, say, $m = n = 5$, there are over 10^3 tests needed to be conducted for each h) and the false rate of the resulting causal relationship (i.e., the overall type I error rate) will also be difficult to control.

Here we propose a more efficient procedure so that the two minimal important sets S_X and S_Y can be adequately estimated without requiring a large number of hypothesis tests. The proposed procedure is sequential and utilizes the ideas of the backward and forward search method that have been successfully applied to various variable selection problems (Kutner et al. [19], Miller [24]). The detailed steps are introduced in the following sections.

4.1. Backward and forward search of important variables

We first explain how to utilize the concept of backward search in our efficient hypothesis testing procedure, which basically comes from the ideas in Theorems 3.2 and 3.5. Suppose $Y \underset{(c)}{\rightarrow} X$ for some positive integer c, where $Y_t = (Y_{1,t}, \ldots, Y_{m,t})'$ and $X_t = (X_{1,t}, \ldots, X_{n,t})'$, and now we would like to characterize the minimal important set S_Y. The backward search starts with an initial estimate of S_Y, say, the full set $V_Y^{(0)} = \{Y_{1,t}, \ldots, Y_{m,t}\}$. In the first stage, all the potential unimportant variables are identified (by using the hypothesis tests shown in the next section) and the most unimportant one, say, $Y_{i_1,t}$, is excluded from $V_Y^{(0)}$. Next, the estimated important set is updated by setting $V_Y^{(1)} = V_Y^{(0)} \setminus \{Y_{i_1,t}\}$. Continue in this fashion until none of the unimportant variables are found, say, if the search stops at the sth stage, then the minimal important set S_Y can be estimated as $\tilde{S}_Y = V_Y^{(s-1)}$. Note that another minimal important set $\tilde{S}_X$ can be obtained in a similar way. Thus, the trimmed causal relationship can be estimated by

$$\tilde{S}_Y \underset{(c)}{\rightarrow} \tilde{S}_X.$$

The forward search, conversely, starts with an empty set $V_Y^{(0)} = \emptyset$ and ends up with an estimate $\tilde{S}_Y = V_Y^{(s')}$ by sequentially adding one most important variable (if exists) into the important set. It should be mentioned that the required number of hypothesis tests for either the backward or the forward search is at most $O(m^2+n^2)$, which is much smaller than that based on the exhaustive search, viz., $O(2^{m+n})$. In concluding, as the required number of hypothesis tests becomes smaller, a more reasonable bound on the type I error rate of each individual test may be used without lowering the overall accuracy of the estimated causal relationship.

4.2. The Wald test with constrained parameter spaces

Note that if the causal relationship is defined by the one-step-ahead predictor (i.e., $c = 1$ or $h = 1$), then at each stage, determining whether or not a particular variable belongs to the estimated important set is equivalent to validating a designated linear constraint on the VAR coefficients under a given linear constraint on the VAR coefficients. In this case, the Wald statistic can be used to perform the test. It should be noted that in order to perform the Wald test associated with each stage of our search procedure, we need a stronger assumption that the term a_t in (2.1) is a “standard white noise”. Denoting $a_t = (a_{1,t}, \dots, a_{K,t})'$, this assumption indicates that a_t are continuous random vectors satisfying the following: (i) $E(a_t) = \mathbf{0}$; (ii) $E(a_t a_t')$ is nonsingular; (iii) a_t and a_s are independent for $t \neq s$; and (iv) there exists some constant $\bar{c} < \infty$ such that

$$E|a_{i,t}a_{j,t}a_{l,t}a_{m,t}| \le \bar{c} \quad \text{for } i, j, l, m = 1, \dots, K, \text{ and all } t.$$

These assumptions allow us to establish consistency and asymptotic normality of the least square estimators of the VAR coefficients (see Lütkepohl [23] for details).

Consider the stationary VAR(p) model in (2.1) and let $\theta = \text{vec}(c, A_1, \dots, A_p)$, which represents a $(K^2p + K) \times 1$ vector comprising all VAR coefficients. We next introduce two multiple hypothesis testing procedures associated with the backward and forward search for estimating the minimal important sets.

The backward search. Suppose we would like now to estimate the minimal important set S_Y by using the backward search. Assume that the search does not terminate at the $(s-1)$th stage and denote the updated important set by $V_Y^{(s-1)} = \{Y_{i_1,t}, Y_{i_2,t}, \dots, Y_{i_{m-s+1},t}\}$, where $\{i_1, i_2, \dots, i_{m-s+1}\}$ is a subsequence of $\{1, \dots, m\}$. Then, at the sth stage, every variable in $V_Y^{(s-1)}$ will be tested so as to see if it is a potential unimportant variable (and thus can be possibly excluded from $V_Y^{(s-1)}$). Note that this is equivalent to conducting a multiple testing procedure with $m - s + 1$ null hypotheses

$$H_0^1 : Q_{i_1}\theta = \mathbf{0},\ H_0^2 : Q_{i_2}\theta = \mathbf{0}, \dots,\ H_0^{m-s+1} : Q_{i_{m-s+1}}\theta = \mathbf{0}, \tag{4.1}$$

where each null hypothesis H_0^k is now tested under a designated linear constraint

$$R_s^k\theta = \mathbf{0}, \quad k = 1, \dots, m - s + 1. \tag{4.2}$$

Here, $Q_{i_1}, \ldots, Q_{i_{m-s+1}}$ are full-ranked matrices of order $np \times (K^2p + K)$ so that $Q_{i_k}\theta = \mathbf{0}$ represents equivalently $A^{(1)}_{XY,j}(:, i_k) = \mathbf{0}$ for all $j = 1, \ldots, p$, $k = 1, \ldots, m - s + 1$. By Theorem 3.1, each Q_{i_k} is a designated matrix for testing if $Y_{i_k,t} \in V_Y^{(s-1)}$ is an unimportant variable. In addition, $R_1^k \equiv R_1$ is a zero matrix (i.e., no constraints at the first stage) and $R_s^k \equiv R_s$ $(s \geq 2)$ is a full-ranked matrix of order $(s-1)np \times (K^2p+K)$ so that the constraint $R_s^k\theta = \mathbf{0}$ represents equivalently an identity of the $s-1$ unimportant variable(s) that have been excluded from Y_t. Solving the homogeneous system in (4.2) yields

$$\theta = B_s^k \theta_s, \tag{4.3}$$

where $B_s^k \equiv B_s$ is a full-ranked matrix of order $(K^2p+K) \times [K^2p+K-(s-1)np]$ and θ_s represents a $[K^2p+K-(s-1)np] \times 1$ vector of unconstrained parameters in θ, $k = 1, \ldots, m - s + 1$.

The forward search. When the forward search is used to estimate S_Y, the above formulation needs to be justified. Assume that the search does not terminate at the $(s - 1)$th stage and denote the complement of the updated important set by $\bar{V}_Y^{(s-1)} = \{Y_{1,t}, \ldots, Y_{m,t}\} \setminus V_Y^{(s-1)} = \{Y_{i_1,t}, Y_{i_2,t}, \ldots, Y_{i_{m-s+1},t}\}$, where $\{i_1, i_2, \ldots, i_{m-s+1}\}$ is a subsequence of $\{1, \ldots, m\}$. Then, at the sth stage, every variable in $\bar{V}_Y^{(s-1)}$ will be tested so as to see if it is a potential important variable (and thus can be possibly added into $V_Y^{(s-1)}$). Analogously, this is equivalent to considering $m - s + 1$ null hypotheses shown in (4.1), while each null hypothesis H_0^k is now tested under a designated linear constraint

$$R_s^k\theta = \mathbf{0}, \quad k = 1, \ldots, m - s + 1. \tag{4.4}$$

Analogously, here $Q_{i_1}, \ldots, Q_{i_{m-s+1}}$ are full-ranked matrices of order $np \times (K^2p + K)$ so that $Q_{i_k}\theta = \mathbf{0}$ represents equivalently $A^{(1)}_{XY,j}(:, i_k) = \mathbf{0}$ for all $j = 1, \ldots, p$, $k = 1, \ldots, m-s+1$. Therefore, each Q_{i_k} is a designated matrix for testing if $Y_{i_k,t} \in \bar{V}_Y^{(s-1)}$ is an important variable. In addition, R_s^k is a full-ranked matrix of order $(m-s)np \times (K^2p+K)$ so that $R_s^k\theta = \mathbf{0}$ represents equivalently an identity of all the unimportant variables in $\bar{V}_Y^{(s-1)}$, with the tested variable $Y_{i_k,t}$ being excluded from the set. Solving the homogeneous system in (4.4) based on these justifications, we obtain

$$\theta = B_s^k \theta_s^k,$$

where B_s^k is a full-ranked matrix of order $(K^2p+K) \times [K^2p+K-(m-s)np]$ and θ_s^k represents a $[K^2p + K - (m - s)np] \times 1$ vector of unconstrained parameters in θ, $k = 1, \ldots, m - s + 1$.

The following theorem can be used to test the constrained null hypotheses in (4.1) for both the backward and forward search.

Theorem 4.1. *Consider the stationary VAR(p) model in (2.1) and let* $\theta = \operatorname{vec}(c, A_1, \ldots, A_p)$. *Define the* $K(p+1) \times 1$ *vector* $Z_t = \operatorname{vec}(1_{K\times1}, W_t, W_{t-1}, \ldots, W_{t-p+1})$ *and let* $\Gamma = E(Z_t Z_t')$. *Suppose we would like to estimate the minimal*

important set S_Y based on T periods of observations. At the sth stage of the search procedure, let $\hat{\theta}$ be the generalized least squares estimator of θ under the constraint in (4.2) (for the backward search) or (4.4) (for the forward search). Then for each $k \in \{1, \dots, m-s+1\}$, we have:
(a)

$$\sqrt{T}(Q_{i_k}\hat{\theta} - Q_{i_k}\theta) \xrightarrow{d} N\left(\mathbf{0}, Q_{i_k}B_s^k\left[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k\right]^{-1}(B_s^k)'Q_{i_k}'\right) \quad \text{as } T \to \infty, \tag{4.5}$$

where $Q_{i_k}B_s^k[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k]^{-1}(B_s^k)'Q_{i_k}'$ is a positive definite matrix with "$\otimes$" representing the Kronecker product;
(b) Define the Wald statistic by

$$\lambda_k = T(Q_{i_k}\hat{\theta})'\left\{Q_{i_k}B_s^k\left[(B_s^k)'(\hat{\Gamma} \otimes \hat{\Sigma}_a^{-1})B_s^k\right]^{-1}(B_s^k)'Q_{i_k}'\right\}^{-1}(Q_{i_k}\hat{\theta}), \tag{4.6}$$

where $\hat{\Gamma}$ and $\hat{\Sigma}_a$ are consistent estimators of Γ and Σ_a, respectively. Then, under H_0^k in (4.1), λ_k has an asymptotic χ^2 distribution with np degrees of freedom.

Proof. (a) By Proposition 5.3 in Lütkepohl [23], we have

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N\left(\mathbf{0}, B_s^k\left[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k\right]^{-1}(B_s^k)'\right) \quad \text{as } T \to \infty.$$

So it is clear that

$$\sqrt{T}(Q_{i_k}\hat{\theta} - Q_{i_k}\theta) \xrightarrow{d} N\left(\mathbf{0}, Q_{i_k}B_s^k\left[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k\right]^{-1}(B_s^k)'Q_{i_k}'\right) \quad \text{as } T \to \infty.$$

Since we know $[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k]^{-1}$ is nonsingular (thus positive definite) by Proposition 5.1 in Lütkepohl [23], it is clear that $Q_{i_k}B_s^k[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1})B_s^k]^{-1}(B_s^k)'Q_{i_k}'$ is positive semidefinite. Therefore, it suffices to show that it is a nonsingular matrix. Let us now consider the backward search method and, for simplicity, assume R_s^k in (4.2) can be written as

$$R_s^k = \left(\begin{array}{c|c} \mathbf{I}_{(s-1)np} & \mathbf{0}_{(s-1)np \times [K^2p+K-(s-1)np]} \end{array}\right).$$

Then, B_s^k in (4.3) can be written as

$$B_s^k = \begin{pmatrix} \mathbf{0}_{(s-1)np \times [K^2p+K-(s-1)np]} \\ \mathbf{I}_{K^2p+K-(s-1)np} \end{pmatrix}.$$

Since each row in Q_{i_k} is linearly independent of the rows in R_s^k, Q_{i_k} can be written as

$$Q_{i_k} = \left(\begin{array}{c|c} \mathbf{0}_{np \times [(s-1)np]} & Q^* \end{array}\right),$$

where Q^* is an $np \times [K^2p + K - (s-1)np]$ matrix and $rank(Q^*) = rank(Q_{i_k}) = np$. Upon noting that

$$Q_{i_k}B_s^k = \left(\begin{array}{c|c} \mathbf{0}_{np \times [(s-1)np]} & Q^* \end{array}\right)\begin{pmatrix} \mathbf{0}_{(s-1)np \times [K^2p+K-(s-1)np]} \\ \mathbf{I}_{K^2p+K-(s-1)np} \end{pmatrix} = Q^*,$$

we then have $rank(Q_{i_k} B_s^k) = rank(Q^*) = np$. Further, by Cholesky decomposition we have

$$\left[(B_s^k)'(\Gamma \otimes \Sigma_a^{-1}) B_s^k\right]^{-1} = LL',$$

where L is a full-ranked lower triangular matrix. Therefore, $rank(Q_{i_k} B_s^k L) = np$, and thus $rank([Q_{i_k} B_s^k L][Q_{i_k} B_s^k L]') = rank(Q_{i_k} B_s^k LL'(B_s^k)' Q'_{i_k}) = np$. Since now $Q_{i_k} B_s^k [(B_s^k)'(\Gamma \otimes \Sigma_a^{-1}) B_s^k]^{-1} (B_s^k)' Q'_{i_k}$ has a full rank np, it is a nonsingular matrix and the proof is complete. Note that the proof for the forward search method is quite similar, and is therefore not presented for brevity.
(b) Since $Q_{i_k} B_s^k [(B_s^k)'(\Gamma \otimes \Sigma_a^{-1}) B_s^k]^{-1} (B_s^k)' Q'_{i_k}$ is nonsingular by (a), the Wald statistic λ_k in (4.6) is well defined. In addition, $\hat{\Gamma}$ and $\hat{\Sigma}_a$ are consistent estimators of Γ and Σ_a, respectively. So, the required result follows directly by Proposition C.15(5) in Lütkepohl [23]. □

Note that H_0^k in (4.1) is rejected if λ_k in (4.6) is too large, or, conversely, H_0^k in (4.1) is not rejected if λ_k in (4.6) is too small. Thus, at stage s of the backward search, $Y_{i_k,t} \in V_Y^{(s-1)}$ is considered as a potential unimportant variable if the Wald statistic λ_k is small (or its p-value is large). Once all the unimportant variables are identified, the backward search will exclude the "most unimportant one" (associated with λ_k having the largest p-value) from $V_Y^{(s-1)}$. On the other hand, at stage s of the forward search, $Y_{i_k,t} \in \bar{V}_Y^{(s-1)}$ is considered as a potential important variable if the Wald statistic λ_k is large (or its p-value is small). Once all the important variables are identified, the forward search will add the "most important one" (associated with λ_k having the smallest p-value) into $V_Y^{(s-1)}$.

In addition to the ways of developing the minimal important sets, the backward and forward search also have different features in running the multiple hypothesis testing procedure. For instance, if a simplified causal relationship does exist, then it is more difficult for the backward search to reject the null hypothesis at early stages. This is due to the fact that, with less constraints on the VAR coefficients at early stages, the variance of the generalized least squares estimator $\hat{\theta}$ becomes larger, thus resulting in a smaller test statistic λ_k in (4.6). On the other hand, if a simplified causal relationship does exist, then it is easier for the forward search to reject the null hypothesis at early stages. This is due to the fact that, with more constraints on the VAR coefficients at early stages, the variance of $\hat{\theta}$ becomes smaller, thus resulting in a larger test statistic λ_k in (4.6).

Remark 5. Recall that to identify if a particular variable is an important/unimportant one, we need to examine the corresponding row/column in the matrix $A_{XY,j}^{(h)}$. If $h > 1$, this has to be validated under a "nonlinear" constraint of the VAR coefficients (e.g., $h = 2$ refers to a quadratic constraint), which may not be represented as the form in (4.2). Therefore, there may not exist a matrix B_s in (4.3) so that the asymptotic results in Theorem 4.1 can be established. The readers can refer to Lütkepohl [23] for some examples.

4.3. Control of type I error

It is now clear that searching the minimal important set S_Y or S_X corresponds to a multiple testing procedure, where the fundamental problem is to control the so-called familywise error rate (FWER). Since the constrained hypothesis tests described in Section 4.2 are correlated, it is natural to consider the classical Bonferroni procedure (Miller [25]). However, as the number of variables becomes large, both the backward and forward search procedures may still require a considerably large number of individual tests. This will force the Bonferroni procedure to select a rather conservative bound on the type I error rate of each individual test, thus resulting in a reduction in the power of the test. Note that there exists a fairly rich literature on the improvement of the Bonferroni procedure (Šidák [31, 32], Simes [33], Hochberg [14]). Among all the available procedures, here we utilize the one called the *Holm's step-down procedure* (Holm [16]) due to the following reasons: (i) it does not require any restriction on the joint distribution of the test statistics (i.e., it allows dependence among the null hypotheses); and (ii) it controls the same FWER and is shown to be uniformly more powerful than the Bonferroni procedure. We now briefly introduce how the Holm's step down procedure is used to perform the multiple testing at each stage of our search method.

The Holm's Step-down Procedure. Denoting by $\pi_{(1)} \leq \pi_{(2)} \leq \cdots \leq \pi_{(m-s+1)}$ the ordered p-values and by $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(m-s+1)}$ the corresponding null hypotheses at the sth stage of our search method. For a given level of significance $0 < \alpha^* < 1$, let

$$k = \arg \min_{i=1,\ldots,m-s+1} \left\{ \pi_{(i)} > \frac{\alpha^*}{m-s+2-i} \right\}. \tag{4.7}$$

The Holm's step-down procedure rejects the null hypotheses $H_0^{(1)}, \ldots, H_0^{(k-1)}$ and accepts $H_0^{(k)}, \ldots, H_0^{(m-s+1)}$. If such k does not exist, then all the null hypotheses are rejected.

Another important characteristic of our search procedure is that it can possibly stop at any stage, viz., the number of tests required to complete the search procedure is "random". This also implies that the FWER of all the required tests for finding the minimal important set is a random quantity. Therefore, our goal is to select the value of α^* in (4.7) so that the "expected FWER" does not exceed a preset control level α. The detailed steps for finding α^*, the minimal important set, and the required theoretical basis are described below.

Theorem 4.2. *Suppose the backward/forward search is used to estimate the minimal important set S_Y and only one variable is possibly excluded/included at each stage s based on the Holm's step-down procedure. There exists an $\alpha^* \in (0, \alpha)$ such that if α^* is chosen as the bound for the FWER of all tests in each search stage s, then $E_{\alpha^*}(FWER) \leq \alpha$.*

Proof. Let FWER_s denote the familywise error rate of all the tests at stage s; then, by Holm's step-down procedure, we have $\text{FWER}_s \leq \alpha^*$ for all $s =$

$1, 2, \ldots, m$. Since the search can possibly stop at any stage, given any $0 < \alpha^* < \alpha$, the expected FWER of the whole search procedure is expressed as

$$E_{\alpha^*}(\text{FWER}) = \sum_{s=1}^{m} (\text{FWER}^s) p_s, \tag{4.8}$$

where p_s represents the probability that the search stops at stage s, and FWER^s denotes the familywise error rate based on all the individual tests up to stage s, $s = 1, 2, \ldots, m$. Since it is clear that $\text{FWER}^s \leq \sum_{i=1}^{s} \text{FWER}_i \leq s\alpha^*$, we then have

$$E_{\alpha^*}(\text{FWER}) \leq \alpha^* \sum_{s=1}^{m} s p_s. \tag{4.9}$$

Note that $1 \leq \sum_{s=1}^{m} s p_s \leq m$, since $\sum_{s=1}^{m} s p_s$ represents the expected number of required search stages. Therefore, we have $E_{\alpha^*}(\text{FWER}) \leq \alpha$ by simply choosing $\alpha^* = \alpha/m$. □

Theorem 4.2 indicates that, there always exists an $\alpha^* \in (0, \alpha)$ such that if the bound on the FWER of the tests at each search stage is chosen as α^*, then the expected FWER for estimating the minimal important set S_Y will not exceed the preset level α. Note that the value of α^* satisfying $E_{\alpha^*}(\text{FWER}) \leq \alpha$ is clearly not unique, and the best choice is obviously the "least conservative" one, which is given by

$$\bar{\alpha}^* = \sup\{\alpha^* \in (0, \alpha) : \alpha^* \sum_{s=1}^{m} s p_s \leq \alpha\}. \tag{4.10}$$

Therefore, if one knows the probabilities p_s associated with the value of α^* placed in (4.7), then $\bar{\alpha}^*$ can be found by performing a numerical search over the interval $(0, \alpha)$.

Recall that the error terms in the VAR model are assumed to be independent for different time indices (see Section 4.2). Therefore, the bootstrap method can be utilized to estimate the probabilities p_s for any given α^*. To see how the bootstrap procedure works, suppose the least squares estimates of the primary VAR(p) coefficients are solved based on T periods of observations, say, $\{W_t, W_{t-1}, \ldots, W_{t-T+1}\}$, and the associated noise vectors are denoted by $\{a_t, a_{t-1}, \ldots, a_{t-T+1}\}$. The bootstrap sample of T periods, denoted by $\{W_t^*, W_{t-1}^*, \ldots, W_{t-T+1}^*\}$, is generated by the following mechanism:

$$W_t^* = \hat{c} + \sum_{j=1}^{p} \hat{A}_j W_{t-j} + a_t^*, \tag{4.11}$$

where $\hat{c}$ and $\hat{A}_j$ are the least squares estimates of the VAR coefficients based on $\{W_t, W_{t-1}, \ldots, W_{t-T+1}\}$, and a_t^* is the noise resampled from $\{a_t, a_{t-1}, \ldots, a_{t-T+1}\}$. Thus, given a value of $\alpha^* \in (0, \alpha)$, one can estimate the probabilities p_s by performing the search procedure a large number of times, where each time the search is conducted based on a bootstrap sample generated by (4.11). In concluding, the detailed steps for determining whether or not $E_{\alpha^*}(\mathrm{FWER}) \leq \alpha$ are summarized below.

Algorithm 1 (Checking if $E_{\alpha^*}(\mathrm{FWER}) \leq \alpha$ by bootstrap sampling)**.**

Step 1: Select a particular value α^* from the interval $(0, \alpha)$, set $N_s = 0$ for $s = 1, \ldots, m$.

Step 2: Find the least squares estimates of the VAR(p) coefficients based on the observations $\{W_t, W_{t-1}, \ldots, W_{t-T+1}\}$, denote the associated noise vectors by $\{a_t, a_{t-1}, \ldots, a_{t-T+1}\}$.

Step 3: Obtain a bootstrap sample $\{W_t^*, W_{t-1}^*, \ldots, W_{t-T+1}^*\}$ by using the mechanism in (4.11). Perform the backward/forward search procedure, wherein the bound on the FWER of all the tests at each stage is chosen as α^*.

Step 4: If the search stops at stage s based on the Holm's step-down procedure, then set $N_s = N_s + 1$.

Step 5: Repeat Steps 3 and 4 N times (where N is chosen to be a large number), let $\hat{p}_s = N_s/N$, $s = 1, \ldots, m$.

Step 6: If $\alpha^* \sum_{s=1}^m s\hat{p}_s \leq \alpha$, then we conclude that $E_{\alpha^*}(\mathrm{FWER}) \leq \alpha$.

Note that by utilizing Algorithm 1 along with an adequate numerical search method over the interval $(0, \alpha)$, we can approach the value of $\bar{\alpha}^*$ without much computational cost. For example, a simple grid search should work well for this one-dimensional search problem. Once $\bar{\alpha}^*$ is obtained, the minimal important set S_Y can be estimated by performing the backward/forward search based on the primary observed data $\{W_t, W_{t-1}, \ldots, W_{t-T+1}\}$, while the expected FWER is shown to be satisfactorily controlled. For illustrative purpose, the detailed steps of the "backward search" and "forward search" for finding $\tilde{S}_Y$ are summarized below in Algorithms 2 and 3, respectively.

Algorithm 2 (Finding $\tilde{S}_Y$ by the backward search)**.**

Step 1: Set $V_Y^{(0)} = \{Y_{1,t}, \ldots, Y_{m,t}\}$ and $s = 1$.

Step 2: Compute respectively the p-values of the $(m-s+1)$ Wald statistics in (4.6) under the corresponding null hypotheses in (4.1), denote the ordered p-values by $\pi_{(1)}, \ldots, \pi_{(m-s+1)}$ and let $k = \arg\min_{i=1,\ldots,m-s+1}\{\pi_{(i)} > \frac{\bar{\alpha}^*}{m-s+2-i}\}$.

Step 3: Let $Y_{i_{(m-s+1)},t}$ be the potential unimportant variable associated with the null hypothesis $H_0^{(m-s+1)}$ (i.e., the one having the largest p-value). If k exists, set $V_Y^{(s)} = V_Y^{(s-1)} \setminus \{Y_{i_{(m-s+1)},t}\}$ and $s = s + 1$, go to Step 2; otherwise stop the search and set $\tilde{S}_Y = V_Y^{(s-1)}$.

Algorithm 3 (Finding $\tilde{S}_Y$ by the forward search)**.**

Step 1: Set $V_Y^{(0)} = \emptyset$ and $s = 1$.

Step 2: Compute respectively the p-values of the $(m-s+1)$ Wald statistics in (4.6) under the corresponding null hypotheses in (4.1), denote the ordered p-values by $\pi_{(1)}, \ldots, \pi_{(m-s+1)}$ and let $k = \arg\min_{i=1,\ldots,m-s+1}\{\pi_{(i)} > \frac{\bar{\alpha}^*}{m-s+2-i}\}$.

Step 3: Let $Y_{i_{(1)},t}$ be the potential important variable associated with the null hypothesis $H_0^{(1)}$ (i.e., the one having the smallest p-value). If $k \neq 1$, set $V_Y^{(s)} = V_Y^{(s-1)} \cup \{Y_{i_{(1)},t}\}$ and $s = s+1$, go to Step 2; otherwise stop the search and set $\tilde{S}_Y = V_Y^{(s-1)}$.

It is noted that the estimated minimal important sets $\tilde{S}_Y$ and $\tilde{S}_X$ obtained by the backward and forward search may be different. However, they should be reasonably close to the "true minimal important sets". From the viewpoint of implementation, both search methods are computationally feasible and can maintain a certain level of accuracy in deriving the trimmed Granger causality.

4.4. The power of the test

Let us consider the ordered alternative hypotheses at the sth stage of the search procedure:

$$H_a^{(1)} : Q_{i_{(1)}}\theta = \theta_1,\ H_a^{(2)} : Q_{i_{(2)}}\theta = \theta_2, \ldots,\ H_a^{(m-s+1)} : Q_{i_{(m-s+1)}}\theta = \theta_{m-s+1}, \tag{4.12}$$

where the elements of each vector θ_k, $k = 1, \ldots, m-s+1$, satisfy (i) the linear constraint in (4.2) (for the backward search) or (4.4) (for the forward search); and (ii) $A_{XY,j}^{(1)}(:, i_k) \neq \mathbf{0}$ for at least one $j \in \{1, \ldots, p\}$. The power of the kth ordered test against $H_a^{(k)}$ can be computed based on the following theorem.

Theorem 4.3. *Consider the stationary VAR(p) model in (2.1) and let $\theta = vec(A_1, \ldots, A_p)$. Suppose we would like to estimate the minimal important set S_Y based on T periods of observations. At the sth stage of the search procedure, let $\hat{\theta}$ be the generalized least squares estimator of θ under the constraint in (4.2) (for the backward search) or (4.4) (for the forward search). If the value of T is large, then the power of the kth ordered test against $H_a^{(k)}$ is given by*

$$Power_k = P(reject\ H_0^{(k)} \mid H_a^{(k)}) \approx P\left(\chi^2(np, \hat{\gamma}_k) > \chi^2_{1-\bar{\alpha}_k^*}(np)\right), \tag{4.13}$$

where $\chi^2(np, \hat{\gamma}_k)$ is a non-central chi-squared random variable with np degrees of freedom and non-centrality parameter

$$\hat{\gamma}_k = T\theta_k' \left\{ Q_{i_{(k)}} B_s^{(k)} \left[(B_s^{(k)})'(\hat{\Gamma} \otimes \hat{\Sigma}_a^{-1}) B_s^{(k)} \right]^{-1} (B_s^{(k)})' Q_{i_{(k)}}' \right\}^{-1} \theta_k, \tag{4.14}$$

and

$$\bar{\alpha}_k^* = \frac{\bar{\alpha}^*}{m-s+2-k}, \quad k = 1, \ldots, m-s+1.$$

Proof. If T is large, then under $H_a^{(k)}$ we have

$$\sqrt{T} Q_{i_{(k)}} \hat{\theta} \overset{A}{\sim} N\left(\sqrt{T}\theta_k, Q_{i_{(k)}} B_s^{(k)} \left[(B_s^{(k)})'(\Gamma \otimes \Sigma_a^{-1}) B_s^{(k)} \right]^{-1} (B_s^{(k)})' Q_{i_{(k)}}' \right),$$

where "$\overset{A}{\sim}$" means "is approximately distributed as". Thus, by definition, the Wald statistic λ_k in (4.6) is approximately distributed as a non-central chi-squared distribution with np degrees of freedom and non-centrality parameter

$$\gamma_k = T\theta_k' \left\{ Q_{i_{(k)}} B_s^{(k)} \left[(B_s^{(k)})'(\Gamma \otimes \Sigma_a^{-1}) B_s^{(k)} \right]^{-1} (B_s^{(k)})' Q_{i_{(k)}}' \right\}^{-1} \theta_k.$$

Since at each stage some of the coefficients in θ may be unknown or unconstrained, γ_k can be directly estimated by

$$\hat{\gamma}_k = T\theta_k' \left\{ Q_{i_{(k)}} B_s^{(k)} \left[(B_s^{(k)})'(\hat{\Gamma} \otimes \hat{\Sigma}_a^{-1}) B_s^{(k)} \right]^{-1} (B_s^{(k)})' Q_{i_{(k)}}' \right\}^{-1} \theta_k,$$

where $\hat{\Gamma}$ and $\hat{\Sigma}_a$ are the consistent estimates of Γ and Σ_a based on T periods of observations, respectively. Thus, we have

$$\begin{aligned} Power_k = P(reject\ H_0^{(k)} \mid H_a^{(k)}) &= P\left(\lambda_k > \chi^2_{1-\bar{\alpha}_k^*}(np) \mid H_a^{(k)}\right) \\ &\approx P\left(\chi^2(np, \hat{\gamma}_k) > \chi^2_{1-\bar{\alpha}_k^*}(np)\right). \end{aligned}$$

□

5. Numerical results

In this section, we illustrate the proposed hypothesis testing procedures for estimating the trimmed Granger causality with a real example. A simulation study is also carried out to compare their accuracy with that of the Lasso-penalized VAR approach. All the numerical results were obtained by using the software package R (version 2.13.0) and executed on 3.0 GHz AMD Athlon II X2 250 processors with 4 GB of cache under the operating system of Microsoft Windows 7 32-bit Service Pack 1 (SP1).

5.1. A real example

A number of studies have reported strong correlations between international stock markets and indicated the leading role of the markets in Western countries (Copeland and Copeland [6], Jeong [18], Rapach et al. [28]). It is of our interest to identify the important lead/lag linkages between the stock markets of the countries in Asia and Western world. The following are the two groups of stock indices considered in this study:

$X_{1,t}$: Hong Kong Hang-Seng Index (HSI);
$X_{2,t}$: Singapore FTSE Straits Times Index (FTSE STI);
$X_{3,t}$: Bangkok Set Index (BSI);
$X_{4,t}$: Shanghai Synthesis Index (PSI);
$Y_{1,t}$: Germany DAX Index (DAX);
$Y_{2,t}$: Canada S&P/TSX Composite Index (S&P/TSX);
$Y_{3,t}$: Paris SBF 250 Stock Index (SBF250);
$Y_{4,t}$: New York S&P 500 Index (S&P500).

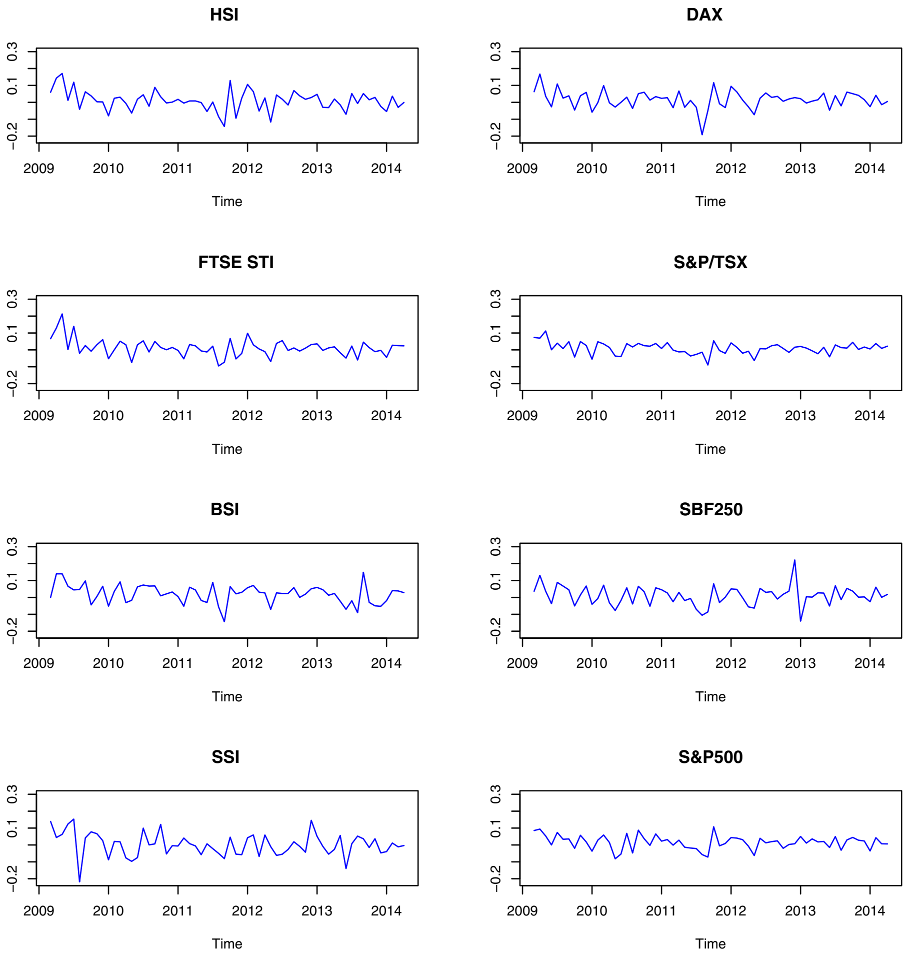

Fig 2. *The time series plots for the growth rates of eight variables recorded monthly from March 2009 to April 2014.*

The data were retrieved from the database of Taiwan Economic Journal (TEJ, `http://www.finasia.biz`), which were collected monthly over the period from March 2009 to April 2014 (i.e., $T = 62$). In order to obtain stationary time series, all the variables were further transformed into the format of "growth rate" (i.e., growth rate = (value in the present period − value in the previous period)/(value in the previous period)). The resulting transformed time series plots are given in Fig. 2.

As shown in Fig. 2, each of the time series plots appears to have fairly constant mean and variance over time. To further validate the property of stationarity, the *Augmented Dickey-Fuller* (ADF) test was applied to each of the eight transformed series. For example, if $Y_{i,t}$ is the transformed series, then a Unit Root test based on the simple assumption of AR(1) model is performed based on 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. As a result, all the p-values for the ADF tests of these eight transformed variables are less than 10^{-4}, which indicates that they can be reasonably treated as stationary time series.

Remark 6. For data that reveal the time-dependent volatilities and/or cross-volatilities, there are other ways that can be used to assess the stationarity of the underlying multivariate time series model. The readers can refer to the work by Aue et al. [2]) and the references therein.

We next establish the desired VAR model based on which we show how to obtain the trimmed Granger causality between $Y_t = (Y_{1,t}, Y_{2,t}, Y_{3,t}, Y_{4,t})'$ and $X_t = (X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t})'$. To select the best order of the VAR model, we first consider four commonly used criteria (Lütkepohl [23]): AIC (Akaike's Information Criterion), HQ (Hannan-Quinn criterion), SC (Schwarz Criterion) and FPE (Final Prediction Error). The definitions of these four criteria are given by

$$\begin{aligned}
\mathrm{AIC}(p) &= \ln\det(\hat{\Sigma}_a) + \frac{2}{T} pK^2, \\
\mathrm{HQ}(p) &= \ln\det(\hat{\Sigma}_a) + \frac{2\ln\ln T}{T} pK^2, \\
\mathrm{SC}(p) &= \ln\det(\hat{\Sigma}_a) + \frac{\ln T}{T} pK^2, \\
\mathrm{FPE}(p) &= \left(\frac{T+Kp+1}{T-Kp-1}\right)^K \det(\hat{\Sigma}_a).
\end{aligned}$$

Note that here $T = 62$, $K = 4 + 4 = 8$, $\hat{\Sigma}_a = T^{-1}\sum_{t=1}^{T} \hat{a}_t \hat{a}_t'$, and $\hat{a}_t$ are the least squares estimates of the error terms. The results, with various choices of order p (say, from 1 to 4), are given in Table 1. As shown in Table 1, most of the criteria (except for the AIC) would suggest the order $p = 1$ that achieves the minimum value. However, the Wald statistic based on the VAR(1) model gives a rather large p-value (say, 0.1528) against the existence of Granger causality

Table 1

Estimation of AIC, HQ, SC and FPE for various choices of the VAR order p. Note that the minimum value of each criterion is highlighted in bold

Order p	AIC(p)	HQ(p)	SC(p)	FPE(p)×10^{24}
1	−54.40	**−53.36**	**−51.70**	**2.43**
2	−54.04	−52.09	−48.94	4.11
3	−53.77	−50.89	−46.26	8.85
4	**−54.87**	−51.07	−44.96	9.20

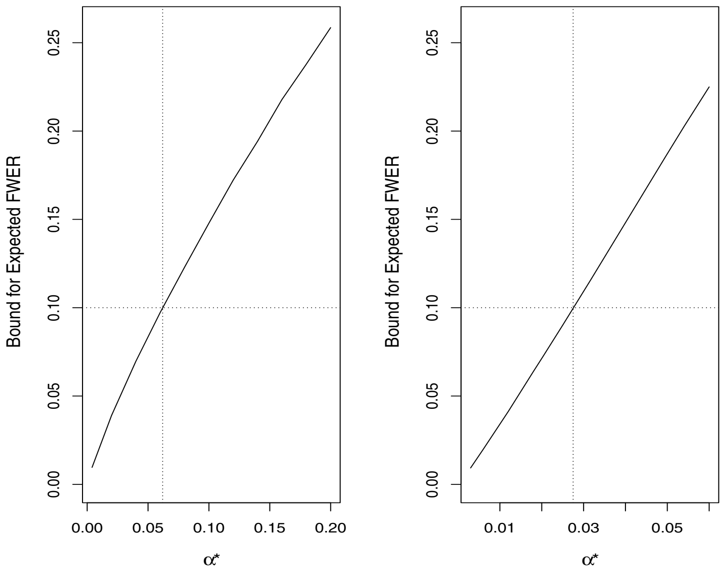

Fig 3. *The estimated bounds for $E_{\alpha^*}(FWER)$ with respect to the α^* values over two selected subintervals of* $(0, \alpha)$ *under the backward (left panel) and forward (right panel) search. Here α is given by* 0.10.

between Y_t and X_t. Therefore, we have decided to choose $p = 2$ for fitting the VAR model, which is suboptimal based on the four criteria but indicates the existence of strong Granger causality between Y_t and X_t (the associated p-value is 0.000125).

Suppose now we choose $c = 1$ (i.e., the one-step-ahead predictor is considered) and $\alpha = 0.10$ (the bound on the expected FWER of obtaining one minimal important set). We first estimate the minimal important set S_Y based on the VAR(2) model. Consider a class of α^* values equally spaced over the interval $(0, \alpha) = (0, 0.10)$ and implement Algorithm 1 for each α^* based on 10^4 bootstrap samples by using the backward and forward search. Since the estimated bound for E_{α^*}(FWER) increases monotonically with α^*, it suffices to examine their relationships over some selected subintervals of $(0, 0.10)$ so as to find $\bar{\alpha}^*$. The results are shown in Fig. 3.

Based on the numerical results associated with Fig. 3, the values of $\bar{\alpha}^*$ for the backward and forward search methods are estimated (by interpolation) as 0.0597 and 0.0275, respectively. Setting $\bar{\alpha}^* = 0.0597$ and 0.0275, the stages of implementing the backward and forward search for obtaining the estimated minimal important set $\tilde{S}_Y$ are detailed in Tables 2 and 3, respectively.

As can be seen from Tables 2 and 3, both search methods stop at stage 3, which result in the same estimated minimal important set $\tilde{S}_Y = V_Y^{(2)} = \{Y_{1,t}, Y_{2,t}\}$. To obtain the estimated minimal important set $\tilde{S}_X$, the stages of implementing the backward and forward search methods are detailed in Tables

TABLE 2

The stage-by-stage result of the "backward search" for obtaining the estimated minimal important set $\tilde{S}_Y$. Here $\alpha = 0.10$, $\bar{\alpha}^* = 0.0597$, $m = 4$, and for each stage the maximum p-value is highlighted in bold

	The associated p-value					
Stage	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$	Action	Important Set
$s = 1$	0.0007	0.0518	0.0820	**0.2119**	Remove $Y_{4,t}$	$V_Y^{(1)} = \{Y_{1,t}, Y_{2,t}, Y_{3,t}\}$
$s = 2$	0.0006	0.0063	**0.0918**	*	Remove $Y_{3,t}$	$V_Y^{(2)} = \{Y_{1,t}, Y_{2,t}\}$
$s = 3$	0.0001	**0.0088**	*	*	Stop	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$

TABLE 3

The stage-by-stage result of the "forward search" for obtaining the estimated minimal important set $\tilde{S}_Y$. Here $\alpha = 0.10$, $\bar{\alpha}^* = 0.0275$, $m = 4$, and for each stage the minimum p-value is highlighted in bold

	The associated p-value					
Stage	$Y_{1,t}$	$Y_{2,t}$	$Y_{3,t}$	$Y_{4,t}$	Action	Important Set
$s = 1$	**0.0003**	0.0230	0.0526	0.2495	Add $Y_{1,t}$	$V_Y^{(1)} = \{Y_{1,t}\}$
$s = 2$	*	**0.0088**	0.1210	0.0529	Add $Y_{2,t}$	$V_Y^{(2)} = \{Y_{1,t}, Y_{2,t}\}$
$s = 3$	*	*	**0.0918**	0.2336	Stop	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$

TABLE 4

The stage-by-stage result of the "backward search" for obtaining the estimated minimal important set $\tilde{S}_X$. Here $\alpha = 0.10$, $\bar{\alpha}^* = 0.0680$, $n = 4$, and for each stage the maximum p-value is highlighted in bold

	The associated p-value					
Stage	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	Action	Important Set
$s = 1$	0.0007	0.0031	**0.3596**	0.0121	Remove $X_{3,t}$	$V_X^{(1)} = \{X_{1,t}, X_{2,t}, X_{4,t}\}$
$s = 2$	0.0002	0.0005	*	**0.0173**	Stop	$\tilde{S}_X = \{X_{1,t}, X_{2,t}, X_{4,t}\}$

TABLE 5

The stage-by-stage result of the "forward search" for obtaining the estimated minimal important set $\tilde{S}_X$. Here $\alpha = 0.10$, $\bar{\alpha}^* = 0.0286$, $n = 4$, and for each stage the minimum p-value is highlighted in bold

	The associated p-value					
Stage	$X_{1,t}$	$X_{2,t}$	$X_{3,t}$	$X_{4,t}$	Action	Important Set
$s = 1$	0.0090	0.1242	0.0385	**0.0040**	Add $X_{4,t}$	$V_X^{(1)} = \{X_{4,t}\}$
$s = 2$	0.0647	0.1240	**0.0528**	*	Stop	$\tilde{S}_X = \{X_{4,t}\}$

4 and 5, respectively. Note that in this case $\bar{\alpha}^* = 0.0680$ and 0.0286 for the backward and forward search, respectively, both of which are estimated based on 10^4 bootstrap samples.

As can be seen from Tables 4 and 5, the estimated important set under the backward search is $\tilde{S}_X = \{X_{1,t}, X_{2,t}, X_{4,t}\}$, while the estimated important set under the forward search is $\tilde{S}_X = \{X_{4,t}\}$. Thus, by choosing $\alpha = 0.10$, the trimmed Granger causality obtained by the backward search is:

$$\{Y_{1,t}, Y_{2,t}\} \underset{(1)}{\rightarrow} \{X_{1,t}, X_{2,t}, X_{4,t}\}.$$

Table 6

The two estimated minimal important sets $\tilde{S}_Y$ and $\tilde{S}_X$ given by the backward and forward search with $\alpha = 0.01, 0.05, 0.10, 0.15$ and 0.20

Choice of α	Backward search	Forward search
0.01	$\tilde{S}_Y = \{Y_{1,t}\}$ $\tilde{S}_X = \{X_{1,t}, X_{2,t}\}$	$\tilde{S}_Y = \{Y_{1,t}\}$ $\tilde{S}_X = \emptyset$
0.05	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$ $\tilde{S}_X = \{X_{1,t}, X_{2,t}\}$	$\tilde{S}_Y = \{Y_{1,t}\}$ $\tilde{S}_X = \emptyset$
0.10	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$ $\tilde{S}_X = \{X_{1,t}, X_{2,t}, X_{4,t}\}$	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$ $\tilde{S}_X = \{X_{4,t}\}$
0.15	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}, Y_{3,t}\}$ $\tilde{S}_X = \{X_{1,t}, X_{2,t}, X_{4,t}\}$	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$ $\tilde{S}_X = \{X_{4,t}\}$
0.20	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}, Y_{3,t}\}$ $\tilde{S}_X = \{X_{1,t}, X_{2,t}, X_{4,t}\}$	$\tilde{S}_Y = \{Y_{1,t}, Y_{2,t}\}$ $\tilde{S}_X = \{X_{4,t}\}$

On the other hand, the trimmed Granger causality obtained by the forward search is:

$$\{Y_{1,t}, Y_{2,t}\} \underset{(1)}{\rightarrow} \{X_{4,t}\}.$$

For comparison purpose, the two estimated minimal important sets given by the backward and forward search, with different choices of α, are shown in Table 6. As can be seen, the estimated important sets obtained by the two different search methods are somewhat different. For example, by choosing $\alpha =$ 0.15 and 0.20, the backward search results in the same estimation of the trimmed Granger causality

$$\{Y_{1,t}, Y_{2,t}, Y_{3,t}\} \underset{(1)}{\rightarrow} \{X_{1,t}, X_{2,t}, X_{4,t}\};$$

by choosing $\alpha = 0.10$, 0.15 and 0.20, the forward search results in the same estimation of the trimmed Granger causality

$$\{Y_{1,t}, Y_{2,t}\} \underset{(1)}{\rightarrow} \{X_{4,t}\}.$$

Note that as the value of α becomes smaller (i.e., a more conservative bound placed on the expected FWER or type I error rate), both the backward and forward search methods result in a rather simplified causal relationship. However, the estimated important set may be "empty" under the proposed search methods (e.g., the $\tilde{S}_X$ obtained by the forward search when $\alpha = 0.01$ and 0.05) - even though the Wald test has proved the existence of Granger causality. These phenomena are caused by the characteristics of both search methods described early in Section 4.3.

To examine the power of the two search methods, at each search stage s we consider the alternative hypothesis $H_a^{(k)} : Q_{i_{(k)}}\theta = \theta_k$ for each individual test, where θ_k is now chosen as $\hat{\theta}$, the ordinary least squares estimate of θ under the constraints in (4.2) and (4.4). Given that $\alpha = 0.10$ and 0.15, the average power of the individual tests at each stage under both search methods for obtaining $\tilde{S}_Y$ and $\tilde{S}_X$ are sketched in Figs. 4 and 5, respectively.

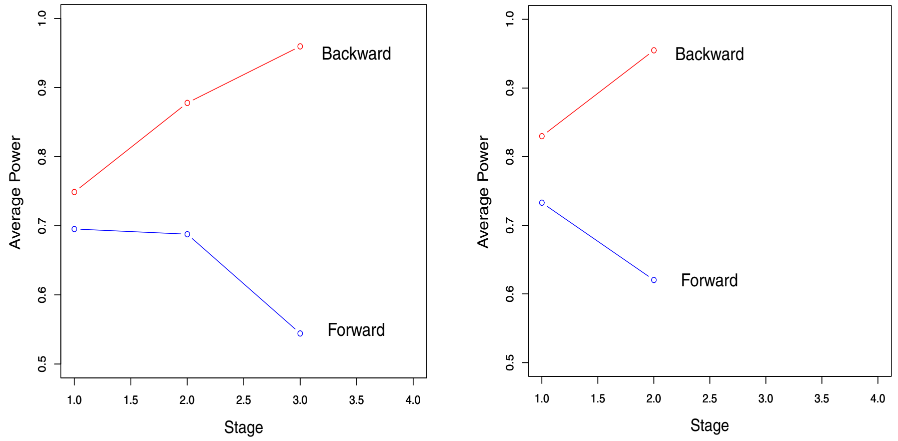

FIG 4. *The average power of the individual tests at each stage of both search methods for obtaining* $\tilde{S}_Y$ *(left panel) and* $\tilde{S}_X$ *(right panel), given that* $\alpha = 0.10$.

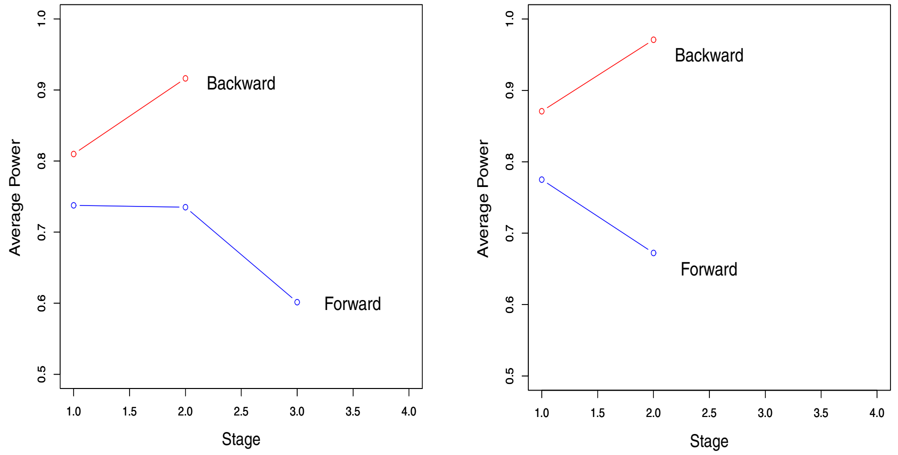

FIG 5. *The average power of the individual tests at each stage of both search methods for obtaining* $\tilde{S}_Y$ *(left panel) and* $\tilde{S}_X$ *(right panel), given that* $\alpha = 0.15$.

Figs. 4 and 5 reveal some interesting findings. First, as the stage moves, the average power increases and decreases respectively under the backward and forward search. This is due to the fact that (i) the bound on the type I error rate of each individual test (i.e., $\bar{\alpha}_k^*$) becomes less strict; and (ii) the estimated non-centrality parameters $\hat{\gamma}_k$ increase (i.e., a shift of the non-central chi-squared distribution to the right) and decrease (i.e., a shift of the non-central chi-squared distribution to the left) respectively under the backward and forward search. Second, for this particular example, the backward search seems to be more powerful (in average sense) than the forward search. Last, as we increase the value of α from 0.10 to 0.15, at each stage the average power of both search methods increases as well. The result is intuitive since less conservative bounds on the type I error rate will increase the possibility of rejecting the null hypotheses.

5.2. A simulation study

Here we perform a simulation study to compare our proposed two search methods and the Lasso-type methods in terms of (i) the accuracy of the estimated two important sets; and (ii) the accuracy of the estimated trimmed Granger causality. Note that two popular Lasso implementations of VAR models are the Lasso-SS and the Lasso-LL method, where the former uses the sum of squared residuals and the later uses minus log likelihood as the loss function, respectively. However, as shown by Davis et al. [7], the performance of these two Lasso-type methods are quite similar under the simulation setup considered in this study. Therefore, numerical results are provided merely for the Lasso-SS method.

Consider the following 6-dimensional VAR(1) model:

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \\ X_{3,t} \\ Y_{1,t} \\ Y_{2,t} \\ Y_{3,t} \end{pmatrix} = \begin{pmatrix} 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.3 & 0 \\ 0.6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \\ X_{3,t-1} \\ Y_{1,t-1} \\ Y_{2,t-1} \\ Y_{3,t-1} \end{pmatrix} + \begin{pmatrix} a_{1,t} \\ a_{2,t} \\ a_{3,t} \\ a_{4,t} \\ a_{5,t} \\ a_{6,t} \end{pmatrix}, \tag{5.1}$$

where $a_t = (a_{1,t}, \dots, a_{6,t})'$ are i.i.d. Gaussian noise with mean zero and covariance matrix

$$\Sigma_a = \begin{pmatrix} \delta^2 & \delta/4 & \delta/6 & \delta/8 & \delta/10 & \delta/12 \\ \delta/4 & 1 & 0 & 0 & 0 & 0 \\ \delta/6 & 0 & 1 & 0 & 0 & 0 \\ \delta/8 & 0 & 0 & 1 & 0 & 0 \\ \delta/10 & 0 & 0 & 0 & 1 & 0 \\ \delta/12 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Suppose now we are interested in finding the trimmed Granger causality between $Y_t = (Y_{1,t}, Y_{2,t}, Y_{3,t})'$ and $X_t = (X_{1,t}, X_{2,t}, X_{3,t})'$ based on the one-step-ahead predictor. By (5.1), the coefficient matrix of interest is then

$$A_{XY,1} = \begin{pmatrix} 0 & 0 & 0 \\ 0.3 & 0 & 0 \\ 0 & -0.3 & 0 \end{pmatrix}.$$

Thus, by choosing $c = 1$, the true important sets are $S_Y = \{Y_{1,t}, Y_{2,t}\}$ and $S_X = \{X_{2,t}, X_{3,t}\}$, respectively, while the trimmed Granger causality is

$$\{Y_{1,t}, Y_{2,t}\} \underset{(1)}{\rightarrow} \{X_{2,t}, X_{3,t}\}.$$

Table 7

The accuracy of different methods for estimating S_Y, S_X and $S_Y \underset{(1)}{\rightarrow} S_X$ with $\delta = 1, 2, 5, 10$, and $\alpha = 0.10, 0.05, 0.01$

Choice of δ	Method	Choice of α	Accuracy of Estimating S_Y	Accuracy of Estimating S_X	Accuracy of Estimating $S_Y \underset{(1)}{\rightarrow} S_X$
1	Backward	0.10	0.636	0.650	0.552
		0.05	0.537	0.521	0.448
		0.01	0.262	0.264	0.206
	Forward	0.10	0.483	0.495	0.398
		0.05	0.439	0.447	0.356
		0.01	0.242	0.252	0.187
	Lasso-SS	–	0.398	0.304	0.200
2	Backward	0.10	0.766	0.781	0.682
		0.05	0.705	0.707	0.633
		0.01	0.510	0.498	0.435
	Forward	0.10	0.665	0.674	0.587
		0.05	0.601	0.611	0.530
		0.01	0.415	0.438	0.363
	Lasso-SS	–	0.326	0.278	0.174
5	Backward	0.10	0.799	0.801	0.724
		0.05	0.770	0.778	0.717
		0.01	0.628	0.625	0.574
	Forward	0.10	0.693	0.710	0.624
		0.05	0.688	0.695	0.613
		0.01	0.527	0.535	0.474
	Lasso-SS	–	0.336	0.272	0.164
10	Backward	0.10	0.811	0.806	0.739
		0.05	0.748	0.750	0.683
		0.01	0.608	0.589	0.531
	Forward	0.10	0.720	0.745	0.660
		0.05	0.661	0.652	0.596
		0.01	0.488	0.507	0.439
	Lasso-SS	–	0.360	0.284	0.170

To assess the accuracy of each method, we simulate the VAR(1) model for 1,000 times by using the R package "mAr" (with the sample size $T = 100$) and record respectively the proportions of times that the estimated important sets and trimmed Granger causality match the true ones (the bias or variation of the estimated VAR parameters is not of particular interest here). For comparison purpose, the order of the VAR model used to perform the backward and forward search is pre-specified by one, while the bound on the type I error rate is chosen to be $\alpha = 0.10$. On the other hand, the Lasso-SS VAR model is estimated by using the R package "fastVAR", with the VAR order pre-specified by one and the tuning parameter selected by 10-fold cross validations. The numerical results are given in Table 7 for various choices of α and δ in Σ_a.

As can be seen from Table 7, both the backward and forward search clearly outperform the Lasso-SS approach, except for the case that a rather tight error bound ($\alpha = 0.01$) and small variability ($\delta = 1$) in the noise term are considered. The improvement on the accuracy becomes more significant as the value of δ

increases (e.g., up to 300% for estimating the trimmed Granger causality). In addition, the proposed two search methods in general have higher accuracy on estimating the important sets as the value of α becomes larger. Such a phenomenon is intuitive since the true coefficient matrix $A_{XY}, 1$ in the VAR model is nonzero (recall that the power of the test increases with α). The reasons why the Lasso-SS method is not able to precisely undertake the characterization of the desired important variables are: (i) spurious non-zero VAR coefficients are produced since the temporal dependence is ignored; (ii) there exists a systematic bias on the estimation of the VAR coefficients due to the shrinkage effect of the Lasso penalty; and (iii) bias accumulates when the entire row or column in the coefficient matrix $A_{XY}, 1$ is estimated. As supported by the numerical evidence, our proposed hypothesis testing procedure nevertheless overcomes the shortcomings of such a pure estimation procedure. Further, by choosing appropriate bounds on the type I error rate, it allows us to better "control" the accuracy on characterizing the desired causal relationship.

6. Concluding remarks

In this paper, we have explained how we could identify the important variables in two groups of time series so that a simplified Granger causal relationship can be presented based on the VAR model. Such a simplified causal relationship, called trimmed Granger causality, allows us to forecast some future quantities by utilizing just a part of variable information. Thus, this work can be viewed as a refined version of the conventional Granger causality test. When the VAR model is specified, explicit conditions are provided for identifying the important variables in both groups of time series. When the parameters of the VAR model are unknown, a multiple hypothesis testing procedure along with two different search algorithms is introduced for estimating the important variables. A simulation study shows that, by choosing appropriate bounds on the type I error rate, our proposed methods significantly outperform the Lasso-type methods (e.g., Lasso-SS and Lasso-LL) in characterizing the correct important variables (or causal relationship). We are currently looking into (i) efficient computational methods for large VAR models and how it would compare with other variable selection methods; and (ii) the hypothesis testing procedure when Granger causality is defined via the h-step-ahead predictor, where $h > 1$ (see Remark 5 for details). The problem in (ii) is challenging since one may have to validate nonlinear constraints on the VAR coefficients in the hypothesis testing procedure, for which the existing theoretical results are rather limited. We hope to report these findings in a future paper.

Acknowledgement

The authors would like to thank the Editor, Associate Editor, and two anonymous reviewers for their valuable comments and suggestions that help improve this manuscript. This work was supported by the research grant NSC 101-2118-M-004-002.

References

[1] Arnold, A., Liu, Y. and Abe, N. (2008). Temporal causal modeling with graphical Granger methods. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 66–75.

[2] Aue, A., Hörmann, S., Horváth, L. and Reimherr, M. (2009). Break Detection in the covariance structure of multivariate time series models. *Ann. Statist.* **37** 4046–4087. MR2572452

[3] Basu, S. and Michailidis, G. (2013). Estimation in high-dimensional vector autoregressive models. Arxiv preprint, arXiv:1311.4175.

[4] Berberian, S. K. (1961). *Introduction to Hilbert Space*. Oxford University Press, New York. MR0137976

[5] Boudjellaba, H., Dufour, J. N. and Roy, R. (1992). Testing causality between two vectors in multivariate autoregressive moving average models. *J. Amer. Statist. Assoc.* **87** 1082–1090. MR1209566

[6] Copeland, M. and Copeland, T. (1998). Lags, lags, and trading in global markets. *Financ. Anal. J.* **54** 70–80.

[7] Davis, R. A., Zang, P. and Zheng, T. (2012). Sparse vector autoregressive modeling. Arxiv preprint, arXiv:1207.0520v1.

[8] Dufour, J. M. and Renault, E. (1998). Short-run and long-run causality in time series theory. *Econometrica* **66** 1099–1125. MR1639415

[9] Fujita, A., Sato, J. R., Garay-Malpartida, H. M., Morettin, P. A., Sogayar, M. C. and Ferreira, C. E. (2007). Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. *Bioinformatics* **23** 1623–1630.

[10] Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Amer. Statist. Assoc.* **77** 304–313. MR0664676

[11] Geweke, J. (1984). *Inference and Causality in Economic Time Series*. Handbook of Econometrics, Vol. 2, North-Holland, Amsterdam, 1101–1144.

[12] Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.

[13] Hacker, R. S. and Hatemi-J, A. (2006). Tests for causality between integrated variables using asymptotic and bootstrap distributions: Theory and application. *Appl. Econ.* **38** 1489–1500.

[14] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126

[15] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1–49. MR0398008

[16] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6** 65–70. MR0538597

[17] Hsiao, C. (1980). Autoregressive modeling and causal ordering of econometric variables. *J. Econ. Dyn. Control* **4** 243–259. MR0671816

[18] Jeong, J. (1999). Cross-border transmission of stock price volatility: Evidence from the overlapping trading hours. *Global Fianc. J.* **10** 53–70.

[19] Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2004). *Applied Liner Regression Models*, 4th ed. MacGraw-Hill, New York.

[20] Luenberger, D. G. (1997). *Optimization by Vector Space Methods*. Wiley, New York. MR0238472

[21] Lütkepohl, H. (1993). *Testing for Causation Between two Variables in Higher Dimensional VAR Models*. In H. Schneeweiss and K. F. Zimmermann (eds), Studies in Applied Econometrics, North-Holland, Amsterdam, 75–91.

[22] Lütkepohl, H. and Burda, M. M. (1997). Modified Wald tests under nonregular conditions. *J. Econometrics* **78** 315–332. MR1453483

[23] Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin. MR2172368

[24] Miller, A. (2002). *Subset Selection in Regression*, 2nd ed., Chapman & Hall/CRC Press, Boca Raton, FL. MR2001193

[25] Miller, R. G. (1981). *Simultaneous Statistical Inference*, 2nd ed., Springer-Verlag, New York. MR0612319

[26] Mosconi, R. and Giannine, C. (1992). Non-causality in cointegrated system: Representation, estimation and testing. *Oxford Bull. Econ. Statist.* **54** 399–417.

[27] Osborn, D. R. (1984). Causality testing and its implication for dynamic econometric models. *Econ. J.* **94** 82–96.

[28] Rapach, D., Strauss, J. and Zhou, G. (2013). International stock return predictability: What is the role of the United States? *J. Finance*, DOI: 10.1111/jofi.12041.

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

[30] Shojaie, A. and Michailidis, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* **26** i517–i523.

[31] Šidák, Z. (1968). On multivariate normal probabilities of rectangles: Their dependence on correlations. *Ann. Math. Statist.* **39** 1425–1434. MR0230403

[32] Šidák, Z. (1971). On probabilities of rectangles in multivariate Student distributions: Their dependence on correlations. *Ann. Math. Statist.* **42** 169–175. MR0278354

[33] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872

[34] Sims, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.

[35] Song, S. and Bickel, P. J. (2011). Large vector auto regressions. Arxiv preprint, arXiv:1106.3915.

[36] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[37] Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69** 297–302. MR0671966