Causal Inference from Slowly Varying Nonstationary Processes
Kang Du and Yu Xiang

Abstract—Causal inference from observational data following the restricted structural causal models (SCMs) framework hinges largely on the asymmetry between cause and effect from the data generating mechanisms, such as non-Gaussianity or non-linearity. This methodology can be adapted to stationary time series, yet inferring causal relationships from nonstationary time series remains a challenging task. In this work, we propose a new class of restricted SCM, via a time-varying filter and stationary noise, and exploit the asymmetry from nonstationarity for causal identification in both bivariate and network settings. We propose efficient procedures by leveraging powerful estimates of the bivariate evolutionary spectra for slowly varying processes. Various synthetic and real datasets that involve high-order and non-smooth filters are evaluated to demonstrate the effectiveness of our proposed methodology.

Index Terms—Causal discovery, nonstationary processes, evolutionary spectra, stationarity test.

I. INTRODUCTION
INFERRING causal relationships from observational data has drawn much attention in recent years [1]–[4], following the pioneering works on structural causal models (SCMs) by Pearl [5]. The main theoretical challenge lies in the identifiability of the causal structure, which is not possible for general SCMs. As a result, various classes of restricted SCMs have been proposed including the linear non-Gaussian acyclic model (LiNGAM) [1], the non-linear additive noise models (ANMs) [2], [6], and the post-nonlinear causal model [3]. The structure identifiability can be proved either exactly [1] or in generic cases [2], [3], [6], and the key to this is to break the symmetry between cause and effect via structural assumptions such as non-Gaussianity or non-linearity.

In light of the ubiquity of time series data, it is appealing to adapt the results for i.i.d. data to stationary time-dependent data. The ANMs have been extended to stationary time series data. In [7], the time series models with independent noise (TiMINo) consider time-invariant functional relationships and i.i.d. noise. Even though the processes generated according to TiMINo are not necessarily stationary, the stationarity of the data is required for the estimation procedure. The well-known Granger causality is designed for vector autoregressive (VAR) models [8] without considering instantaneous effects due to identification issues, while LiNGAM-t [9] incorporates instantaneous effects and non-Gaussian noise. It is worth noting that (bivariate) Granger causality is fundamentally different from the SCM-based approach initiated by [7]. [9], followed in this work. We refer the readers to [10, Chapter 10] and [11] for the vast literature on Granger causality and its variants; also see [12] for adapting Granger causality to locally stationary processes [13]. A bivariate deterministic model via a linear time-invariant filter is studied in [14]. The directed information rate [15], [16] from information theory is defined for bivariate stationary processes (see [17] for its relationship with Granger causality and various estimation methods [18], [19]).

There are several pioneering works on causal inference through the lens of nonstationarity, [20]–[22]. The time-dependent causal model [20] (referred to as TCM in this paper) and CD-NOD [22] model nonstationarity by introducing a surrogate random variable to represent time. TCM deals with time-dependent functional relationships, but the estimation procedure becomes more challenging due to the nonstationarity of the data. The authors in [21] study a linear model, where the coefficients follow the autoregressive models, with additive noise that is uncorrelated in time, and the estimation step relies on nonlinear state-space model estimation procedures. An invariance-based approach is studied in [23] to identify the parents of a target variable $Y_t$, where the linear model that generates $Y_t$ is assumed to be invariant; this is related to the special case of our model with time-invariant filter and stationary noise. However, none of them is built on well-established frameworks for slowly varying nonstationary processes such as evolutionary spectra [24], Wigner-Ville spectral analysis [25], and locally stationary processes [13], from which powerful estimation procedures could be borrowed to greatly facilitate causal discovery tasks.

In this work, we attempt to bridge this gap by proposing a new class of restricted SCMs that allows for causal structure identification in generic cases and can be reliably estimated by leveraging the bivariate evolutionary spectra framework [20], [27]. Our contribution is threefold. First, we focus on a class of processes generated by linear time-varying filters along with stationary Gaussian noise, and develop theoretical results showing that the causal direction is identifiable in generic cases by exploiting the nonstationarity of the data. It is worth stressing that our framework can deal with instantaneous effects, which is an appealing property in comparison with Granger causality. Second, we extend these results to a network setting via a directed acyclic graph (DAG), where the processes are connected through time-varying linear relationships, and the root nodes are assumed to be stationary. The identification result again relies on nonstationarity and this is in contrast to existing works where non-Gaussianity [11]...
or nonlinearity is required for identification. Third, we develop efficient estimation algorithms, leveraging a recent variant of the evolutionary spectra estimate, that perform well on a variety of synthetic and real datasets, including challenging ones with non-smooth and high-order filters.

The paper is organized as follows. In Section II we present our main result on causal identification of a nonstationary bivariate linear model with time-varying coefficients. Various properties of the time-varying lag operator are discussed. In Section III we present our causal inference procedure, building on the bivariate evolutionary spectra estimates and stationarity test. We extend this results to a network setting in Section IV report our experimental results in Section V and summarize our contributions in Section VI.

Preliminary results on the bivariate case have been reported in [29]. In this extended version, (1) we study in greater depth the bivariate case through various concrete settings, and more importantly, (2) we formulate a network setting and develop identification results as well as efficient estimation procedures, supported empirically by the superior performance of our methods in various simulation results.

Due to space limit, we put most of the proofs in the supplemental materials.

A. Notation

Let \( \mathbb{Z}, \mathbb{Z}_{\geq 0}, \) and \( \mathbb{C} \) denote the integers, non-negative integers, and complex numbers, respectively. We use \( \mathbb{Z}_{\geq 0} \) to denote \( \mathbb{Z}_{\geq 0} \cup \{\infty\} \). A sequence of random variables is denoted by \( \{X_t\} \triangleq \{X_t, t \in \mathbb{Z}\} \) with mean function \( \mu_{X,t} = E[X_t] \) and auto-covariance function \( \gamma_{XX}(r,s) = \text{Cov}(X_r, X_s) \). We write \( \{X_t\} \perp \perp \{Y_t\} \) to denote the (statistical) independence between \( \{X_t\} \) and \( \{Y_t\} \), which requires the random vectors \( (X_{t_1}, \ldots, X_{t_n}) \) and \( (Y_{t_1}, \ldots, Y_{t_n}) \) to be independent for any \( n > 0 \) and any sequence \( t_1, \ldots, t_n \in \mathbb{Z} \). Throughout this work, stationary processes are referred to as wide-sense stationary processes. We use the capital Greek letter (e.g., \( \Phi, \Psi, \Theta \)) for the polynomial function and the corresponding lower case (i.e., \( \phi, \psi, \eta \)) for its coefficients. For a matrix \( A \in \mathbb{C}^{n \times n} \), we use \( |A| \) and \( ||A||_p \triangleq \max_{x \neq 0} (||Ax||_p/||x||_p) \), \( p \geq 1 \), to denote its determinant and the matrix norm induced by the \( l_p \) norm, respectively. We use \( \rho(A) \triangleq \max_{1 \leq j \leq n} |\lambda_j| \) to denote the spectral radius of the matrix \( A \), where \( \lambda_1, \ldots, \lambda_n \) are the eigenvalues of \( A \).

II. MODEL IDENTIFIABILITY IN THE BI-VARIATE CASE

A. Linear time-varying filter with additive stationary noise

For a process \( \{X_t\} \), we define the lag operator \( B \) as \( B^j X_t \triangleq X_{t-j}, j \in \mathbb{Z}_{\geq 0} \). Let \( \Phi^p_t(z) \triangleq \sum_{j=0}^p \phi_{t,j} z^j, z \in \mathbb{C} \), denote a time-dependent polynomial function of finite degree \( p \). If \( \Phi^p_t(z) \) is not constantly zero, we require \( \phi_{t,p} \neq 0 \) for some \( t \). For infinite degree \( p = \infty \), we define \( \Phi^\infty_t(z) \triangleq \sum_{j=0}^\infty \phi_{t,j} z^j \), with \( z \in \mathbb{C} \) such that \( |z| \leq 1 \), where the coefficients of \( \Phi^\infty_t(z) \) are assumed to be absolutely summable, i.e., \( \sum_{j=0}^\infty |\phi_{t,j}| < \infty \). Given a polynomial function \( \Phi^p_t(z) \) of degree \( p \in \mathbb{Z}_{\geq 0} \), a time-varying (lag-polynomial) operator of order \( p \) is defined as

\[
\Phi^p_t(B) \triangleq \sum_{j=0}^p \phi_{t,j} B^j, \tag{1}
\]

We call an operator time-invariant if it is not dependent on \( t \). For an operator \( \Phi^p_t(B) \) of finite order \( p \), if there exists an operator \( \Theta^p_t(B), r \in \mathbb{Z}_{\geq 0} \), such that \( \Theta^p_t(B)\Phi^p_t(B) = 1 \), we call \( \Theta^p_t(B) \) the (left) inverse operator of \( \Phi^p_t(B) \), which is denoted by \( (\Phi^p_t(B))^{-1} \).

In this work, we start with a class of bivariate Gaussian processes \( \{X_t, Y_t\} \) that are trend-free \( (\mu_{X,t} = \mu_{Y,t} = 0) \) and follow the following model

\[
Y_t = \Phi^p_t(B)X_t + N_t, \quad p \in \mathbb{Z}_{\geq 0}, \quad \{N_t\} \perp \perp \{X_t\}, \tag{2}
\]

where the noise \( \{N_t\} \) is a stationary process, and we assume that \( \Phi^p_t(B) \) is invertible. The assumption that \( \Phi^p_t(B) \) is invertible (see Lemma I for details) implies that our model always includes the instantaneous effects (i.e., \( \phi_{t,0} \neq 0 \)), which is regarded as a more difficult case compared with the one without instantaneous effects.

**Remark 1:** Note that certain lagged terms of \( Y_t \) are allowed in (2), since (2) can be equivalently written as any invertible time-invariant operator applied to both sides, and the noise remains stationary. We will thus focus on the representation in (2) for simplicity.

**Remark 2:** We do not specify the generating process of \( \{X_t\} \), which is in contrast to the bivariate version of SCMs in [27], [9], [20], [21] where the cause is assumed to be a noise variable. Our bivariate setting is more challenging in that one could use a stationarity test to tell apart the cause from effect if \( \{X_t\} \) is always stationary. In our network setting in Section VI, however, we will have to assume the root nodes are stationary since the problem seems to be intractable otherwise.

We say a backward model exists if there exists \( \Psi^p_t(B), q \in \mathbb{Z}_{\geq 0} \), and a stationary process \( \{\tilde{N}_t\} \) such that

\[
X_t = \Psi^p_t(B)Y_t + \tilde{N}_t, \quad \{\tilde{N}_t\} \perp \perp \{Y_t\}. \tag{3}
\]

The causal direction \( x \to y \) is said to be identifiable if the joint distribution of \( \{X_t, Y_t\} \) does not admit a backward model. Note that a valid backward model requires the coefficients of \( \Psi^p_t(B) \), i.e., \( \psi_{t,i} \) to be absolutely summable.

B. Time-varying operator

In order to establish the identifiability results of our model, we need to first investigate some key properties of the time-varying operator. We say an operator \( \Phi^p_t(B) \) is time-invariant if \( \phi_{t,j} = \phi_{t-1,j} \) holds for all \( j \geq 0 \) and \( t \in \mathbb{Z} \). By applying the operator \( \Phi^p_t(B) \) to \( \{X_t\} \), we obtain

\[
\Phi^p_t(B)X_t = \sum_{j=0}^p \phi_{t,j} X_{t-j}. \tag{4}
\]

Since we focus on Gaussian processes and operators with absolutely summable coefficients, we would like to have any series of the form in (4) to converge even when \( p = \infty \).
address this technical issue, we present the following proposition, and the proof of which is a straightforward extension of the time-invariant case proved in [30] (and we include it in Appendix C for completeness).

**Proposition 1:** Let \( \{X_t\} \) be a sequence of random variables such that \( \sup_t E[|X_t|] < \infty \). If \( \sum_{j=0}^{\infty} |\psi_{t,j}| < \infty \), then

\[
\Psi_t^\infty(B)X_t = \sum_{j=0}^{\infty} \psi_{t,j}X_{t-j}
\]

converges absolutely with probability one. If \( \sup_t E[|X_t|^2] < \infty \), the series converges in mean square to the same limit.

We now discuss the relationship between different operators. First, we say two operators \( \Phi_p(B) \) and \( \Phi_q(B) \), with \( p, q \in \mathbb{Z}_{\geq 0} \), are equivalent if \( \phi_{t,j} = \psi_{t,j} \) holds for all \( j \geq 0 \) and \( t \in \mathbb{Z} \), and we write \( \Phi_p(B) = \Phi_q(B) \). Otherwise, we use \( \not\equiv \) to denote they are not equivalent. To facilitate the analysis, we will make use of an equivalent definition for the rest of the paper. We write \( \Phi_p(B) = \Psi_p(B) \) if

\[
\Phi_p(z) = \Psi_p(z)
\]

holds for \( z \) in some open set \( \mathbb{E} \subset \mathbb{C} \) that contains 0 (see Appendix C for the proof of equivalence).

It can be easily verified that time-varying lag-polynomial operators do not satisfy the commutative property of multiplication in general, i.e.,

\[
\Phi_p(B)\Phi_q(B) \neq \Phi_q(B)\Phi_p(B).
\]

**Remark 3:** For two operators \( \Phi_p(B) \) and \( \Phi_q(B) \) such that \( t_1 \) and \( t_2 \) do not depend on each other, we have \( \Phi_{t_1}(B)\Phi_{t_2}(B) = \Psi_{t_1}(B)\Psi_{t_2}(B) \).

As a consequence of the non-commutative property in (7), \( \Phi_p(B) \) may not be the inverse operator of \( (\Phi_p(B))^{-1} \) in general. It is known that, when \( \Phi_p(B) \) is time-invariant, a necessary and sufficient condition for \( \Phi_p(B) \) to be invertible is

\[
\sum_{j=0}^{p} \phi_{j}z^j \neq 0,
\]

for \( |z| \leq 1 \), which says that the roots of the polynomial in (8) are strictly outside the unit circle. A similar statement is proved in [30]. When \( \Phi_p(B) \) is time-varying and \( \phi_{t,p} \neq 0 \) for all \( t \), a necessary and sufficient condition for the invertibility of \( \Phi_p(B) \) is provided in [31] using Green’s functions, making the evaluation of the condition very challenging. In the following, we provide two sufficient conditions and one necessary condition on the existence of inverse operators that are easy to check. Some examples will be discussed afterwards to illustrate the conditions.

**Lemma 1:** Let \( \Phi_p(B) \) be an operator of finite order \( p \geq 1 \), and we assume that \( \phi_{t,j} \neq 0 \) for some \( j \geq 0 \) for each \( t \).

(i) **Sufficient conditions for \( \Phi_p(B) \) to be invertible.**

The inverse operator \( (\Phi_p(B))^{-1} \) exists if either of the following conditions holds,

(a) \( |\phi_{t,0}| > \sum_{j=1}^{p} |\phi_{t+j,j}| > 0 \),

(b) \( \phi_{t,0} > \phi_{t+1,1} > \ldots > \phi_{t+p,p} \geq 0 \).

(ii) **Necessary condition for \( (\Phi_p(B))^{-1} \) to have a finite order.** \( \Phi_p(B) \) has an inverse operator of finite order \( q \) only if \( \phi_{t,0} \neq 0 \) and

\[
\phi_{t-i,p} = 0, \quad i = 0, 1, 2, \ldots, q.
\]

**Remark 4:** If \( \Phi_p(B) \) is time-invariant, then (11) reduces to \( \phi_{t,p} = 0 \) for all \( t \), which contradicts the definition of \( \Phi_p(B) \) since it requires that \( \phi_{t,p} \neq 0 \) for some \( t \). Thus the inverse operator of \( \Phi_p(B) \) cannot be of finite order in the time-invariant case.

The proof of Lemma 1 is provided in Appendix C. In Section II-D we show that there is a close relationship between condition (9) and a slowly varying condition on the coefficients from the evolutionary spectra framework. The necessary condition in Lemma 1 says that an inverse operator of finite order exists only if \( \phi_{t,p} = 0 \) for infinitely many \( t \). This condition characterizes a class of operators that could be restrictive since it does not contain the time-invariant operators (and recall that the inverse operator of a time-invariant operator is of infinite order). We therefore consider the “complement” of this class to be a more general class of operators.

Now we provide three examples to show that inverse operators exist under the conditions in Lemma 1.

**Example 1.** Consider the first-order operator \( \Phi_1(B) \triangleq 1 + \phi_{t,1}B \), where \( \phi_{1,1} = 1 \) when \( t \) is even, and \( \phi_{1,1} = 0 \) otherwise. One can check that condition (11) holds for all \( q \geq 1 \). Then using (15), it is straightforward to find that

\[
(1 - \phi_{t,1}B)(1 + \phi_{t,1}B) = 1 - \phi_{t,1}\phi_{t-1,1}B^2 = 1.
\]

**Example 2.** Consider the first-order operator \( \Psi_1(B) \triangleq 1 + \psi_{t,1}B \), with \( 0 < |\psi_{t,1}| < 1 \). One can check that condition (11) does not hold for any \( q \geq 1 \) while condition (9) holds immediately. Then by (15), we obtain

\[
(1 + \psi_{t,1}B)^{-1} = 1 + \sum_{j=1}^{\infty} \left(-1\right)^j \sum_{k=1}^{j} \psi_{t-k+1,1}B^j.
\]

**Example 3.** Consider the operator \( \Psi_1(B) \) in Example 2, with \( \psi_{t,0} = 1 \) and \( \psi_{t,1} = 0.5 \cos(t/T) \), \( T \in \mathbb{Z} \). Since \( t/T \) is a rational number for any \( t \in \mathbb{Z} \), we have \( \psi_{t,1} \neq 0 \) and \( |\psi_{t,1}| < 1 \), which implies that (11) does not hold for any \( q \geq 1 \). It then follows from (9) in Lemma 1 that \( \Psi_1(B) \) has an inverse operator of the form (12). This operator was employed in [26] as the transfer function for an open loop system.

C. Identifiability

Recall that for bivariate Gaussian processes, the backward model defined in [3] has to satisfy two constraints: the independence constraint (\( \{\tilde{N}_1\} \perp \{Y_t\} \)) and the stationarity constraint (\( \{\tilde{N}_1\} \) is stationary). Our main theorem characterizes two necessary conditions, corresponding to these two constraints, regarding the existence of a backward model [3]. To illustrate that the constraints for a backward model to exist...
are hard to be satisfied, we provide the identifiability results for the i.i.d. setting in Corollary 1 and 2.

**Theorem 1**: Let \( \{X_t, Y_t\} \) be a bivariate Gaussian process following the model (2) such that

\[
Y_t = \Phi_t^x(B)X_t + N_t,
\]

(13)

where we assume that \( \Phi_t^x(B) \) is invertible. Then a backward model of (13) exists only if the following two conditions are satisfied.

**I. Condition for the independence constraint.** The equation

\[
\gamma_{XX}(t_1, t_2) = H_t^x(B)\alpha(t_1, t_2)
\]

determines a nonempty class of operators \( \mathcal{O} \) such that for any \( H_t^x(B) \in \mathcal{O}, \{Y_t\} \perp \perp \{X_t - H_t^x(B)Y_t\}, \) and \( \{H_t^x(Y_t)\} \) has a unique distribution.

**II. Condition for the stationarity constraint.** Let \( \Theta_t^x(B) \triangleq \Phi_t^x(B)^{-1} \), then there exists an operator \( H_t^x(B) = \sum_{j=0}^s \eta_{t,j}B^j \) in \( \mathcal{O} \) such that

\[
\sum_{j=1}^r \sum_{k=1}^r \eta_{t,j}\theta_{t-j,k}\gamma_{NN}(k-j)
\]

is time-invariant.

**Remark 5**: When \( \{X_i\} \) and \( \{N_i\} \) are i.i.d. Gaussian processes and \( \Phi_t^x(B) \) is time-invariant, \( \{X_t, Y_t\} \) is stationary and our model is reduced to a Gaussian vector autoregressive model with instantaneous effects, which is known to be non-identifiable [9].

The proof of Theorem 1 is provided in Appendix A. As shown in the proof of Theorem 1, the second condition is a consequence of the first one. However, the second condition itself is quite strong in that a combination of time-varying coefficients has to be time-invariant. As a result, Theorem 1 implies that the causal direction is likely to be identifiable in generic cases, which is further supported by our experimental results on both synthetic and real-world datasets in Section V

Note that this is analogous to the identifiability results for the nonlinear ANMs [2] where the backward model only exists under strong conditions. Estimating the time-varying coefficients remains a challenging task. Fortunately, reliable estimation procedures are available for a class of slowly varying processes, called bivariate evolutionary spectra processes [26, 27], based on which we propose a natural causal discovery procedure in Section III

As a consequence of the two constraints, we have the following corollary when both \( \{X_i\} \) and \( \{N_i\} \) are i.i.d. Gaussian processes (see Appendix B for the proof).

**Corollary 1**: If \( \{X_t\} \) and \( \{N_t\} \) are two i.i.d. Gaussian processes with \( \text{E}[X_t^2] = \sigma_{X}^2 \) and \( \text{E}[N_t^2] = \sigma_{N}^2 \), and \( \phi_{t,p} \neq 0 \) for all \( t \), then the coefficients of the operator \( \Psi_t^x(B) \) in (3) are determined by \( \psi_{t,0} = 1/\phi_{t,0} \) and

\[
\psi_{t,i} = -\frac{1}{\phi_{t-i,0}} \left( \min_{j=1}^{\min(p,i)} \psi_{t,j}\phi_{t-j,i} + \frac{\sigma_{N}^2 \psi_{t,p-i}^2}{\sigma_X \phi_{t+p-i,p}} \right)
\]

(14)

for \( i \geq 1 \). A backward model (3) exists only if \( \{\psi_{t,i}\} \) is absolutely summable and \( \text{Var}(N_t) = \sum_{j=1}^{\min(s,r)} \psi_{t,j}\theta_{t-j,i}\gamma_{NN}(0) \) is time-invariant, where \( \{\theta_{t,j}\} \) are the coefficients of the inverse operator of \( \Phi_t^x(B) \) (see equation (15) below).

**Remark 6**: As the SNR \( \sigma_Y^2/\sigma_X^2 \) goes to infinity, the coefficients \( \{\psi_{t,i}\} \), converges to \( \theta_{t,0} = 1/\phi_{t,0} \) and

\[
\theta_{t,i} = \min_{j=1}^{\min(p,i)} \psi_{t,j}\phi_{t-j,i} + \frac{\sigma_{N}^2 \psi_{t,p-i}^2}{\sigma_X \phi_{t+p-i,p}}, \quad i \geq 1,
\]

(15)

which are the coefficients of \( \Theta_t^x(B) \triangleq \Phi_t^x(B)^{-1} \) (see the derivation of (15) in [22] equation (4.10)). Thus, the invertibility of \( \Phi_t^x(B) \) is asymptotically necessary for a backward model to exist as the SNR increases.

In Corollary 1 we show that \( \{\psi_{t,i}\} \) can be solved iteratively, and the variance of \( \{\tilde{N}_t\} \) is written as a combination of \( \{\psi_{t,i}\} \) and \( \{\theta_{t,j}\} \). In general, it could be hard to check whether \( \{\psi_{t,i}\} \) is absolutely summable and whether \( \{\tilde{N}_t\} \) is stationary. To get a concrete sense of the identifiability result, we simplify the setting by letting \( \Phi_t^x(B) \) to be of zero order in the following corollary (see Appendix C for the proof).

**Corollary 2**: Let \( \{X_t\} \) and \( \{N_t\} \) be i.i.d. Gaussian processes with zero means and variances \( \sigma_Y^2 \) and \( \sigma_X^2 \), respectively. Consider the following forward model with \( \phi(t) \neq 0 \),

\[
Y_t = \phi(t)X_t + N_t, \quad \{N_t\} \perp \perp \{X_t\}.
\]

(16)

Then there exists a model as follows,

\[
X_t = \frac{\phi(t)}{\phi^2(t) + \sigma_{N}^2/\sigma_{X}^2}Y_t + \tilde{N}_t, \quad \{\tilde{N}_t\} \perp \perp \{Y_t\},
\]

(17)

where \( \{\tilde{N}_t\} \) is determined by \( \tilde{N}_t = \frac{\sigma_{N}}{\sqrt{\phi^2(t) + \sigma_{N}^2/\sigma_{X}^2}}W_t \), where \( \{W_t\} \) is an i.i.d. process with \( \sigma_{W}^2 = 1 \).

**Remark 7**: Due to the stationarity constraint on \( \{\tilde{N}_t\} \), a backward model exists only if \( |\phi(t)| \) is time-invariant. The noise \( \{\tilde{N}_t\} \) has the form of a stationary process multiplied by a nonnegative function, which belongs to a class of non-stationary processes called the uniformly modulated process (UMP) [24] (see the definition of UMP in Section III).

If the stationary noise assumption is relaxed to be the UMP noise, then a backward model always exists in the setting of Corollary 2. But in the general setting, by replacing \( Y_t \) in (3) with (2), one can write

\[
\tilde{N}_t = (1 - \Psi_t^x(B)\Phi_t^x(B))X_t - \Psi_t^x(B)N_t,
\]

(18)

which is a sum of two independent processes. In generic cases, \( \{\tilde{N}_t\} \) is not only nonstationary but non-UMP. Thus our model is likely to be identifiable even if we consider the UMP noise. This is also supported empirically by our experimental results on synthetic data in Section V.

Since the problem of inferring the causal network of non-stationary processes is well-known to be quite challenging, we assume that there are no confounding factors in this work. However, we provide a toy example to illustrate that nonstationarity can be helpful for avoiding spurious causal relations due to confounders.
Example 4. Consider that $Z_t$ is unobserved and it causes $X_t$ and $Y_t$ (i.e., $X_t \leftarrow Z_t \rightarrow Y_t$) as follows: $X_t = Z_t + N_t$, and $Y_t = a(t)Z_t + N_t$, with $\{Z_t\}, \{N_t\}$, and $\{N_t\}$ being jointly independent i.i.d. Gaussian processes with zero mean and unit variance. In this setting, any inferred causal direction from $X_t$ to $Y_t$ or vice versa will be considered as a spurious causal relation, and we show below that they can be avoided by the nonstationarity introduced through $a(t)$. Similar to the calculations in Corollary 2, one can find the model $X_t = (a(t)/a^2(t) + 1)Y_t + \varepsilon_t$, with $\{Y_t\} \perp \{\varepsilon_t\}$ and $\text{Var}(\varepsilon_t) = 1 + (a^2(t) + 1)^{-1}$, which does not satisfy the stationarity constraint. For the other direction we have $Y_t = (a(t)/2)X_t + \varepsilon_t$, with $\{X_t\} \perp \{\varepsilon_t\}$ and $\text{Var}(\varepsilon_t) = 1 + a^2(t)/2$. Thus one cannot fit a model for either of the two directions.

Our main identifiability results in Theorem 1 provide the necessary conditions for the backward model to exist. For a given bivariate Gaussian process $\{X_t, Y_t\}$ generated according to the forward model, however, it is not straightforward to evaluate whether the set of operators $O$ in Theorem 1 is empty or not. To shed some light on the operators determined by the independence constraint, we consider the identifiability of our model for finite sequences in the next section.

D. Identifiability for finite sequences

Consider a finite sequence of random variables $\{X_t, Y_t\}_{t=T}^{0}$, from the bivariate Gaussian process $\{X_t, Y_t\}$, and we denote $X_t = (X_{t−T}, X_{t−T+1}, \ldots, X_t)^T$ and $Y_t = (Y_{t−T}, Y_{t−T+1}, \ldots, Y_t)^T$. We call the following model a truncated backward model

$$X_t = b_t^\top Y_t + \varepsilon_t, \quad (19)$$

where $b_t \in \mathbb{R}^{(T+1) \times 1}$ and $\varepsilon_t \triangleq X_t - b_t^\top Y_t$. For this model, we weaken the independence constraint in (3) to $Y_t \perp \varepsilon_t$ for each $t$, and the stationarity constraint is on the process $\{\varepsilon_t\}$. In this case, if a truncated backward model exists, $b_t$ is unique and simply given by

$$b_t = \text{Cov}^{-1}(Y_t, Y_t) \text{Cov}(Y_t, X_t), \quad (20)$$

where we assume that $\text{Cov}(Y_t, Y_t)$ is invertible in this section for simplicity. However, $b_t$ is not necessarily in the class of operators $O$ in Theorem 1, since the independence constraint has been weakened. In the following proposition (see Appendix E for the proof), we show that the truncated backward model (19) is a good approximation of the backward model (5) as the length of the sequence $T$ increases.

Proposition 2: Assume that $\inf_{T \geq 0} \lambda_{\min}(\text{Cov}(Y_t, Y_t)) > 0$, where $\lambda_{\min}(A)$ denotes the smallest eigenvalue of $A$. If a backward model (5) exists, then for $T$ large enough, we have

$$\Psi_t^\top \Psi_t Y_t - b_t^\top Y_t = \begin{cases} 0, & 0 < q < T \\ o(1), & q = \infty. \end{cases}$$

Different from Theorem 1, the existence of a backward model (19) only relies on the stationarity constraint, as provided in the following proposition (see Appendix E for the proof).

Proposition 3: A truncated backward model (19) exists only if the following expression is time-invariant

$$\text{Var}(\varepsilon_t) = \text{Var}(X_t) - \text{Cov}(X_t, X_t) \Phi_t^\top \left( \Phi_t \text{Cov}(X_t, X_t) + \text{Cov}(N_t, N_t) \right)^{-1} \Phi_t \text{Cov}(X_t, X_t),$$

where $X_t = (X_t, X_{t−1}, \ldots, X_{t−T−p})^T$, $N_t = (N_t, N_{t−1}, \ldots, N_{t−T})^T$, and $\Phi_t \in \mathbb{R}^{(T+1) \times (T+p+1)}$ is defined by $\Phi_{t,j} = \phi_{t−1, j−1}$. Observe that the only quantity in $\text{Var}(\varepsilon_t)$ that is time-invariant is $\text{Cov}(N_t, N_t)$, thus $\text{Var}(\varepsilon_t)$ is time-varying in generic cases. In the same i.i.d. setting as in Corollary 1, we provide the following corollary to further illustrate the stationarity constraint.

Corollary 3: If $\{X_t\}$ and $\{N_t\}$ are two i.i.d. Gaussian processes with $\text{E}[X_t^2] = \sigma_X^2$ and $\text{E}[N_t^2] = \sigma_N^2$, then a truncated backward model (19) exists only if

$$\text{Var}(\varepsilon_t) = \sigma_X^2 - \sigma_X^2 \Phi_t^\top \left( \Phi_t \text{Cov}(X_t, X_t) + \text{Cov}(N_t, N_t) \right)^{-1} \Phi_t \text{Cov}(X_t, X_t),$$

is time-invariant, where $\varepsilon_t \in \mathbb{R}^{(T+p+1) \times 1}$ is a unit vector with the first element being one.

The proof of Corollary 3 follows from simple algebra. As $\sigma_X^2/\sigma_N^2 \to \infty$, we have

$$\text{Var}(\varepsilon_t) \to \sigma_X^2 - \sigma_X^2 \Phi_t^\top \left( \Phi_t \text{Cov}(X_t, X_t) + \text{Cov}(N_t, N_t) \right)^{-1} \Phi_t \text{Cov}(X_t, X_t),$$

where $\Phi_t \triangleq \Phi_t^\top \left( \Phi_t \text{Cov}(X_t, X_t) + \text{Cov}(N_t, N_t) \right)^{-1} \Phi_t$ is a time-varying projection matrix. Thus it is straightforward to see that $\text{Var}(\varepsilon_t)$ is time-varying in general.

III. Model estimation

A. Causal inference procedure

To simplify the presentation of the estimation procedure in this section, we will adopt an alternative expression of the model (2) without using the lag-polynomial operator. Consider a bivariate process $\{X_t, Y_t\}$, we say the causal direction between $\{X_t\}$ and $\{Y_t\}$ is $x \rightarrow y$ if the following model holds,

$$Y_t = \sum_{u=0}^{\infty} d_t(u) X_{t−u} + N_t, \quad \{X_t\} \perp \{N_t\}, \quad (21)$$

where $\{N_t\}$ is a stationary process, and $\{d_t(u)\}$ is called the time-varying filter. Conversely, if $\{X_t, Y_t\}$ admits the model, $X_t = \sum_{u=0}^{\infty} d_t(u) Y_{t−u} + N_t, \{Y_t\} \perp \{N_t\}$, where $\{N_t\}$ is a stationary process, then we say the causal direction is $y \rightarrow x$. The assumptions on model (21) that allow efficient estimation of $\{d_t(u)\}$ are technical and will be deferred to Section III-D after a brief overview of the evolutionary spectra framework. We now describe our causal inference procedure in Algorithm 1 to test the null hypothesis $H_0 : x \rightarrow y$, and the test for $y \rightarrow x$ can be done in the same manner. Let $p_t^{x \rightarrow y}$ denote the $p$-value from the independence test and $q_t^{x \rightarrow y} = 1$ if the residual is stationary and 0 if nonstationary. Similarly, we obtain $p_t^{y \rightarrow x}$ and $q_t^{y \rightarrow x}$ from the test for $y \rightarrow x$. We accept or
reject $H_0$ by checking the following conditions. For a prefixed $\alpha$, we accept $H_0$ if $p_1^{xy} \geq \alpha$ and $p_2^{yx} \leq \alpha$. (Similarly, we reject $H_0$ if $p_1^{xy} < \alpha$ and $p_2^{yx} > \alpha$.) If $p_1^{xy} \geq \alpha$ and $p_2^{yx} \geq \alpha$, then we rely on the stationarity test: We accept $H_0$ if $q_S^{xy} = 1$ and $q_S^{yx} = 0$ (or reject $H_0$ if $q_S^{xy} = 0$ and $q_S^{yx} = 1$). The causal inference procedure remains undecided for all the other cases.

Here are some comments regarding the implementation details in Algorithm 1. Given a window size $N_F$, the maximal order of model (21) considered by our estimation procedure is $[N_F/2]$ (see Section V for details). The order $p$ can be selected using AIC [33] or BIC [34]. For a similar independence test task, previous works [7], [20] have used a kernel-based independence test developed for i.i.d. data [35] (referred to as HSIC), which may suffer from high false positive rates in certain cases [36]. The estimation of the filter and the stationarity test are based on the evolutionary spectra framework [37] by incorporating the multitaper method as in the univariate case [28], [38] (see more details below).

B. Univariate nonstationary processes

To set the stage, we start with a brief review of the evolutionary spectra framework [24]. Consider a class of nonstationary processes $\{X_t\}$, with $E[X_t] = 0$ and $E[|X_t|^2] < \infty$ for $t \in \mathbb{Z}$, such that

$$X_t = \int_{-\pi}^{\pi} \phi_t(w) dZ(w), \ t \in \mathbb{Z},$$

(22)

for some family $\mathcal{F}$ of functions $\{\phi_t(w)\}$ (defined on $[-\pi, \pi]$ indexed by $t$) and a measure $\mu(w)$, where $Z(w)$ is an orthogonal increment process with $E[|dZ(w)|^2] = d\mu(w)$. If there exists a family of functions $\mathcal{F} = \{\phi_t(w) = e^{iwt}A_t(w)\}$ such that $\{X_t\}$ can be represented as in (22) and for any fixed $w$, the Fourier transform of $h_w(t) \triangleq A_t(w)$ viewed as a function of $t$, denoted by $H_w(v)$, has an absolute maximum at the origin, then $\{X_t\}$ is called an oscillatory process with respect to oscillatory functions $\{e^{iwt}A_t(w)\}$, and the evolutionary spectrum at time $t$ with respect to $\mathcal{F}$ is

$$dF_t(w) = |A_t(w)|^2 d\mu(w).$$

Note that $h_w(t) = 1$ corresponds to the case when $\{X_t\}$ is a stationary process, which leads to $H_w(v) = \delta(v)$, where $\delta(\cdot)$ is the Dirac delta function. To estimate the evolutionary spectral density, Priestley [24] proposed a double-window technique, consisting of a short-time Fourier transform and smoothing. Recently, the bias/variance/resolution tradeoff of a variant of the evolutionary spectra estimate, incorporating the multitaper method [39], is characterized [28]. Interesting methodologies on neural processes can be found in [40].

It is hard to characterize characteristic widths [24], which quantifies the length of a “stable” segment, exactly for semi-stationary processes [41]. However, there is one important class of processes whose characteristic widths can be bounded from below. This class, termed as the uniformly modulated processes (UMP) [24], is of the following form:

$$X_t = c(t)Y_t,$$

(23)

where $Y(t)$ is a stationary process with zero mean and spectral density $f_Y(w)$, and the Fourier transform of $c(t)$ has an absolute maximum at the origin. Thus it follows straightforwardly that

$$X_t = \int_{-\pi}^{\pi} c(t)e^{iwt}dZ(w),$$

where $E[dZ(w)]^2 = dF_Y(w).$ The process introduced in (23) is an oscillatory process since $F_Y = \{c(t)e^{iwt}\}$ is a family of oscillatory functions. The evolutionary spectrum with respect to $\mathcal{F}$ is $f_t(w) = c^2(t)f_Y(w)$.

Algorithm 1 Causal inference procedure (bivariate)

Input: $\{(X_t, Y_t)\}_{t=1}^T$, window size $N_F$

Output: $p$-values from the independence tests and decisions from the stationarity tests

procedure TEST($x \rightarrow y$)

Estimation:

Estimate the filter $\hat{d}_t(u)$

Compute the residuals $\hat{N}_t \triangleq Y_t - \sum_{u=0}^p \hat{d}_t(u)X_{t-u}$

Independence test:

Test whether $\{X_t\} \perp \perp \{\hat{N}_t\}$

Stationarity test:

Test the stationarity of $\{\hat{N}_t\}$

procedure TEST($y \rightarrow x$)

C. Bivariate nonstationary processes

Now we are ready to discuss bivariate processes $\{X_t, Y_t\}$, consisting of two oscillatory processes,$$
X_t = \int_{-\pi}^{\pi} A_{t,x}(w)e^{iwt}dZ_x(w),
Y_t = \int_{-\pi}^{\pi} A_{t,y}(w)e^{iwt}dZ_y(w),$$

where $\{Z_x(w)\}$ with $E[dZ_x(w)]^2 = d\mu_x(w)$ and $\{Z_y(w)\}$ with $E[dZ_y(w)]^2 = d\mu_y(w)$ are two orthogonal increment processes, and $E[dZ_x(w)dZ_y(w)] \triangleq d\mu_{xy}(w)$. The evolutionary cross-spectrum [26] of $\{X_t, Y_t\}$ at time $t$ with respect to $\mathcal{F}_x$ and $\mathcal{F}_y$ is

$$dF_{t,xy}(w) = A_{t,x}(w)A_{t,y}(w)d\mu_{xy}(w).$$

The cross-spectral density of $\{X_t, Y_t\}$ at time $t$ is

$$f_{t,xy}(w) = A_{t,x}(w)A_{t,y}(w)\frac{d\mu_{xy}(w)}{dw}.$$
D. Estimation of the filter

Following the model assumptions in [26], we assume \( \{X_t\} \) and \( \{Y_1\} \) are semi-stationary processes. The filter \( d_t(u) \) is assumed to satisfy the slowing-varying condition [27] described as follows. Let \( D_t(u) \) denote the Fourier transform of \( d_t(u) \) with respect to \( u \). For each \( w \), consider \( D_t(u) \) as a function of \( t \), with (generalized) Fourier transform

\[
D_t(w) = \sum_{\theta=-\infty}^{\infty} e^{i\theta t} L_w(\theta).
\]

If \( |L_w(\theta)| \) attains the maximum at \( \theta = 0 \) for all \( w \), i.e., \( |L_w(0)| \geq |L_w(\theta)| \) for \( \theta \neq 0 \), we say the slowly varying condition is satisfied. Since \( t \) and \( u \) in \( d_t(u) \) are discrete, \( L_w(\theta) \) is a periodic function of \( w \) and \( \theta \) both with period \( 2\pi \). It suffices to define the slowly varying condition in the period \(-\pi < w, \theta \leq \pi \). By leveraging the powerful estimation procedure in [26], we estimate \( d_t(u) \) by

\[
\hat{d}_t(u) = \mathcal{F}_w^{-1} \{ \hat{D}_t(u) \} = \mathcal{F}_w^{-1} \{ \hat{f}_k(x_k(w)) \},
\]

where \( \mathcal{F}_w^{-1} \) denotes the inverse Fourier transform. To examine the empirical performance of our estimation procedure, in Section [X-A] we consider models with different frequencies and with different SNRs in Experiment 1 and non-smooth filters in Experiment 4.

It is natural to ask whether the slow varying condition and either of the two sufficient conditions for invertibility in Lemma 1 can be satisfied simultaneously. We provide two classes of filters to show that this is indeed the case.

Example 5. Consider the filter \( d_t(u) = \delta(u) + \sum_{k=1}^{l} 1/2^k \cos(A_k t) \delta(u - k) \), where \( A_k \in (-\pi, \pi) \), \( l \geq 1 \) and \( \delta(u) \) is the unit impulse function. Since we have \( \sum_{k=1}^{l} 1/2^k \cos(A_k(t+k)) < \sum_{k=1}^{l} 1/2^k < 1 \), the invertibility condition (9) is satisfied. The functions \( D_t(u) \) and \( L_w(\theta) \) in the period \(-\pi < w, \theta \leq \pi \) are given as follows,

\[
D_t(w) = 1 + \sum_{k=1}^{l} \frac{1}{2^k} \cos(A_k t) e^{-ikw},
\]

\[
L_w(\theta) = 2\pi \left( \delta(\theta) + \sum_{k=1}^{l} \frac{e^{-ikw}}{2^{k+1}} (\delta(\theta - A_k) + \delta(\theta + A_k)) \right).
\]

Since \( L_w(\theta) \) is a linear combination of delta functions, it is straightforward to see that the slowly varying condition is satisfied.

Example 6. Consider the filter \( d_t(u) \) with \( d_t(0) = 1 \) and \( \sum_{u=1}^{\infty} |d_t(u)| < \infty \) for \( 1 \leq u \leq p \), for which we can assume that condition (9) or condition (10) holds. Then, we have

\[
L_w(\theta) = 2\pi \delta(\theta) + \sum_{u=1}^{p} \sum_{t=-\infty}^{\infty} d_t(u) e^{-iuw} e^{-i\theta t}
\]

for \(-\pi < w, \theta \leq \pi \), where

\[
\sum_{u=1}^{p} \sum_{t=-\infty}^{\infty} d_t(u) e^{-iuw} e^{-i\theta t} \leq \sum_{u=1}^{p} \sum_{t=-\infty}^{\infty} |d_t(u)| < \infty.
\]

For any \(-\pi < w, \theta \leq \pi \), we have \( |L_w(\theta)| < \infty \) for \( \theta \neq 0 \) and \( |L_w(0)| = \infty \). Thus the slowly varying condition is satisfied.

The examples are in fact more general than they seem to be. Specifically, the filter \( d_t(u) \) in the first example can be generalized in different ways. For example, one can shift and scale the cosine function kernels with some constants, and modify the coefficient \( 1/2^k \). The zero-order term \( d_t(0) = 1 \) for both examples can be replaced by other function forms. To see this, one can multiply a function \( a(t) \) with Fourier transform \( A(w) \) to \( d_t(u) \), then the function \( L_w(\theta) \) is convolved with \( A(w) \). If \( |A(w)| \) is highly concentrated around the zero frequency, then the slowly varying condition on \( L_w(\theta) \) could be preserved after the convolution. Meanwhile, since \( a(t)d_t(u) = |A(0)|d_t(u) + (a(t) - |A(0)|)d_t(u) \), where the first term is dominating, the invertibility condition could hold for \( a(t)d_t(u) \).

Algorithm 2 PSR stationarity test [42]

Input: \{\( f^K_t(w_j) \), \( 1 \leq i \leq I, 1 \leq j \leq J \}\)
Output: accept or reject \( H_0 \)
Compute \{\( W_{ij} \), \( 1 \leq i \leq I, 1 \leq j \leq J \)\}
Test \( S_T/n^2 \sim \chi^2_{I(J-1)} \)
if significant then
    reject \( H_0 \)
else
    Test \( S_T/n^2 \sim \chi^2_{I-1} \)
    if significant then
        reject \( H_0 \)
    else
        accept \( H_0 \)

E. Stationarity test

In order to verify the stationarity of the residual processes \( N_t = Y_t - \sum_{u=0}^{\infty} d_t(u)X_{t-u} \) as in Algorithm 1, we make use of an improved version of the original Priestley and Subba Rao stationarity test (PSR test) [42] by incorporating the multitaper method [28, 38] to obtain \{\( f^K_t(w_j) \), \( 1 \leq i \leq I, 1 \leq j \leq J \)\} in Algorithm 2. The null hypothesis \( H_0 \) is “\( \{X_t\} \) is stationary”. Consider a semi-stationary process \( \{X_t, 0 \leq t \leq T\} \), let \( f_t(w) \) denote its evolutionary spectral density and \( f^K_t(w) \) denote the multitaper estimator with \( K \) tapers and \( N \) as the length of the sample records. For \( i \in \{1, \ldots, I\} \), with \( I = [T/N] \), and \( j \in \{1, \ldots, J\} \), with \( J = [(N+1)/(K+1)] \), let \( W_{ij} = \log f^K_t(w_j) + \psi(k) + \log K \), where \( \psi(k) \) is the digamma function. The stationarity test is based on applying the two-way analysis of variance (ANOVA) test to \( \{W_{ij}\} \). Let \( W_j = (1/I) \sum_{i=1}^{I} W_{ij} \), \( W_{i} = (1/J) \sum_{j=1}^{J} W_{ij} \), and \( W_{ij} = (1/IJ) \sum_{i=1}^{I} \sum_{j=1}^{J} W_{ij} \). The following statistics are needed: between time variance \( S_T = J \sum_{i=1}^{I} (W_{i} - W.)^2 \); between
frequencies variance $S_F = I \sum_{j=1}^d (W_j - W_\cdot)^2$; interaction and residual variance $S_{I+R} = \sum_{i=1}^d \sum_{j=1}^d (W_{ij} - W_i - W_j + W_\cdot)^2$. The algorithm is described in Algorithm 2 where testing $S_{I+R}/\sigma^2 \sim \chi^2_{(l-1)(l-1)}$ is essentially a UMP test (see details from [42]).

IV. Extension to the Network Setting

Our bivariate model shows how nonstationarity can be leveraged to identify causal relations of a pair of processes. In this section, we go beyond the bivariate setting and exploit nonstationarity for identifying the underlying DAG of a set of Gaussian processes.

For a DAG $G$ with nodes $V = \{1, \ldots, N\}$, we use $PA(j)$ for the set of parents $j \in V$, and use $ND(j)$ for the set of non-descendents, defined as the set of nodes such that no path exists from $j$ to any $k \in ND(j)$.

A. Model identifiability

Consider an $N$ dimensional Gaussian process $\{X_t\} = \{\{X_1^t\}, \ldots, \{X_N^t\}\}$ generated by the following SCM corresponding to a DAG $G$ with nodes $V = \{1, \ldots, N\}$,

$$X_t^j = \sum_{k \in PA(j)} \Phi_t^{k \to j}(B)X_t^k + N_t^j, \quad j \in V, \quad (24)$$

where the noise processes $\{N_t^j\}, j \in V$, are jointly independent and stationary. The maximal order of the operators $\Phi_t^{k \to j}(B)$’s is $p$. The DAG structure implies that the variables over $V$ can be arranged in a causal ordering, such that no later variable causes any earlier variable. It is noteworthy that when $N = 2$, the model (24) reduces to a bivariate model with a stationary process being the cause. On the other hand, in our bivariate model, the generating process of the cause $\{X_t\}$ is not specified. Thus the model (24) is an extension of a special case of (2) to the network setting.

As a consequence of the time-varying operators in (24), $\{X_t\}$ is a set of stationary/nonstationary processes. Then a natural question is: What kind of nonstationarity is needed for identifying the DAG $G$? First, let us start with the following example to show that the time-varying operators may not lead to nonstationarity.

Example 7. Let $Y_t = \Phi_t^2(B)X_0$, where $\{X_t\}$ is i.i.d. and $\Phi_t^2(B)$ is defined by $\phi_{t,0} = (-1)^t$ and $\phi_{t,1} = (-1)^t - 1$. Then $\{Y_t\}$ is stationary since $E[Y_t] = (-1)^t \mu_X + (-1)^{t-1} \mu_X = 0$, $\text{Cov}(Y_t, Y_s) = 2\sigma_X^2$ for $s = t$, $\text{Cov}(Y_t, Y_s) = \sigma_X^2$ for $|t-s| = 1$, and $\text{Cov}(Y_t, Y_s) = 0$ for $|t-s| \geq 2$.

This example can be easily generalized by applying any time-invariant operator to $Y_t$. As a result, simply relying on time-varying operators in (24) may not lead to nonstationarity in some non-generic cases. In order for the complete graph $G$ to be identifiable, we need the following assumption.

Assumption 1: A process $\{X_t^j\}$ remains nonstationary if one conditions on $\{X_t^j = 0\}$, where $PA(j) \not\subseteq S \subseteq ND(j) \setminus j$.

In general, Assumption 1 is satisfied when all operators in (24) are time-varying and chosen generically. In the multivariate setting, the restricted ANM formulation [6] assumes that the model $X_j = f_j(XPA(j)) + N_j$ belongs to an identifiable class (see [6] Condition 19 and Definition 27) for details), which puts constraints on $f_j(\cdot)$ when evaluated at some $xPA(j)$'s. In particular, for Gaussian noise $N_j$, the function $f_j(\cdot)$ needs to remain nonlinear for some $xPA(j)$'s and for all $k$ (see [6] Corollary 31). This is similar in spirit to how we require the nonstationarity to exist when conditioning on $\{X_t^j = 0\}$.

Let $R \subseteq V$ denote the set of root nodes in $G$ (i.e., all nodes $j$’s such that $PA(j) = \emptyset$). The identifiability of the graph $G$ is built on the following lemma and the notion of causal ordering.

Lemma 2: Under Assumption 1, a process $\{X_t^j\}, j \in V$, is stationary if and only if $j \in R$.

Proof: $\implies$: For any $j \in R$, the process $\{X_t^j\}$ is determined by $X_t^j = N_t^j$ and thus it is stationary. $\iff$: If no process (i.e., the empty set) is condition on, then Assumption 1 implies that $\{X_t^j\}$ is nonstationary if $PA(j) \neq \emptyset$ (i.e., $j \notin R$). Therefore, any process $\{X_t^j\}$ is stationary if and only if $j \in R$.

Theorem 2: Under Assumption 1, the graph $G$ entailed in (24) is identifiable.

Proof: First, we classify the nodes of $V$ into $K$ classes $\{V^1, \ldots, V^K\}$ such that there is no directed path $j \rightarrow \cdots \rightarrow i$ between $i \in V^k$ and $j \in V^l$ for $1 \leq k \leq l \leq K$. In other words, a node in an earlier class will have a higher causal ordering than a node in a later class. Secondly, we show that the parents of each node $j \in V$ are identifiable, which implies that the whole graph is identifiable.

For the first part, let $V^1$ be the set of root nodes $R$ that is identifiable by Lemma 2. For $k \geq 2$, the nodes will be classified according to the following rule: a node $i \in V \setminus \bigcup_{l=1}^{k-1} V^l$ (i.e., $i$ in the set of unclassified nodes) will be put into $V^k$ if $\{X_t^i\}$ is stationary when conditioning on $\{X_t^{S^k} = 0\}$, $S = \bigcup_{l=1}^{k-1} V^l$.

Now, we prove that this rule can successfully put every node $j \in V$ into some $V^k$, $k \in \{1, \ldots, K\}$, i.e., there exists $K > 0$ such that $V = \bigcup_{k=1}^K V^k$. If the nodes are classified one by one according to a causal ordering, conditioning on the nodes with higher causal orderings is equivalent to conditioning on the parents, by the Markov property of structural causal models [6]. Thus, every node can be put into some $V^k$ since $\{X_t^j\}$ is stationary when conditioning on $\{X_t^{PA(j)} = 0\}$ by our stationarity noise assumption. It turns out the nodes can be only classified according to a causal ordering. To see this, note that according to Assumption 1 it is not possible that the first time a node $j \in V$ is put into some $V^k$ without having $PA(j)$ contained in the previous classes. Thus, the claimed classification procedure is valid and the ordering of the classes is consistent with any causal ordering of $V$.

For the second part, Assumption 1 implies that $PA(i)$ of $i \in V^k$ is the smallest set $S \subseteq \bigcup_{l=1}^k V^l$ such that $\{X_t^i\}$ is stationary when $\{X_t^{S} = 0\}$ is conditioned on. Therefore, since the parents of each node $i \in V$ are identified, the graph $G$ is identifiable.

Remark 8: If Assumption 1 is not satisfied for every $\{X_t^j\}, j \in V$, Lemma 2 will fail to hold. The first part of
the procedure, described in the proof of Theorem 2 may put a node \( j \) in a class that is prior to the true class that it belongs to; the second part will choose a subset of \( PA(j) \) to be the parents of \( j \). As a consequence, the true graph can only be partially identified. In the case when \( \{X_t^j\} \)'s are all stationary, no arrows will be found.

B. Model estimation

In Section III-D, we have described an estimation procedure of the time-varying filter for the bivariate model \((21)\). However, the estimation of time-varying filters for general multivariate models under the evolutionary spectra framework remains an open problem. Our causal inference procedure for the network setting is motivated by the following observation. Let \( AN(j) \) denote the set of ancestors of the node \( j \), i.e., all nodes \( k \)'s such that there exists a path from \( k \) to \( j \). By replacing each \( X_t^k \) in \((24)\) with the corresponding structural equation iteratively, we obtain an equivalent representation of model \((24)\):

\[
X_t^j = \sum_{k \in AN(j)} \Psi_t^{k \rightarrow j}(B)N_t^k + N_t^j, \tag{25}
\]

where

\[
\Psi_t^{k \rightarrow j}(B) = \sum_{(k,v_1,v_2,\ldots)} \Phi_t^{k \rightarrow v_1}(B)\Phi_t^{v_1 \rightarrow v_2}(B)\ldots \Phi_t^{v_d \rightarrow j}(B)
\]

with \((k,v_1,v_2,\ldots)\) denoting any path of any length \( d + 1 \) from \( k \) to \( j \). Note that the operator \( \psi_t^{k \rightarrow j}(B) \) in \((25)\) and the operator \( \Phi_t^{k \rightarrow j}(B) \) in \((24)\) are equal for each \( j \) and \( k \in PA(j) \); we use \( d_t^{k \rightarrow j}(u) \) to denote \( \psi_t^{k \rightarrow j}(B) \) in our algorithm for the network setting as in Algorithm 1. Since \( X_t^j \) is written as a time-dependent linear combination of jointly independent variables, we estimate each filter in \((25)\) using the pairwise procedure described in Section III-D, which turns out to perform well empirically. While \( \{N_t^k\} \) is not observed if \( k \) is not a root node, we shall see later that our algorithm naturally provides estimates of the residuals.

Based on model \((25)\), our algorithm first identifies the ancestors of a node \( j \). Then the task is to identify the parents of \( j \) given its ancestors. Implied by Assumption 1, \( PA(j) \) is the smallest set \( Q \subseteq AN(j) \) such that \( \{X_t^j\} \) is stationary when \( \{X_t^k = 0\} \subseteq Q \), are conditioned on. However, such conditioning is hard to evaluate in practice. To introduce our procedure for identifying \( PA(j) \) (i.e., Procedure 2), we need the following assumption, which is again satisfied in generic cases. With this assumption in place, we will be able to show the correctness of Procedure 2 in one proposition below.

**Assumption 2:** For any \( Q \subseteq AN(j) \) such that \( Q \neq PA(j) \), the equation

\[
W_t^j = X_t^j - \sum_{k \in Q} \Psi_t^{k \rightarrow j}(B)X_t^k \tag{26}
\]

determines a nonstationary process \( \{W_t^j\} \).

**Proposition 4:** For any \( Q \subseteq AN(j) \), the process \( \{W_t^j\} \) determined by \((26)\) is stationary if and only if \( Q = PA(j) \)

It is straightforward to see that we obtain \( W_t^j = N_t^j \) when \( Q = PA(j) \) in \((24)\), using \( \Psi_t^{k \rightarrow j}(B) = \Phi_t^{k \rightarrow j}(B) \) and \((24)\). Thus \( \{W_t^j\} \) is stationary. The other direction is a direct consequence of Assumption 2.

**Algorithm 3** Causal inference procedure (network)

**Input:** \( N \) time series \( \{x_t^j\} \), \( j \in J \)

**Output:** \( PA_j = \{i : \{i \} \text{ is UMP}\} \)

**Initialization:** \( S = V \) \( A = 0_{N \times N} \) \( \{N_t^j\} = \{x_t^j\} \)

while \( S \neq \emptyset \) do

\[ i^* = \text{MinStationary}(\{N_t^j, i \in S\}) \]

if \( i^* = \emptyset \) then break

\[ S \leftarrow S \setminus i^* \]

for \( j \in S \) do

\[ \text{Initialization:} \ c_t^j = 0_{T \times 1}, AN_j = \emptyset \]

\[ \text{for } k \in S^c \text{ do} \]

\[ \text{Estimate the filter } d_t^{k \rightarrow j}(u) \]

if \( \{x_t^j\} \not\subseteq \{N_t^k\} \) then

\[ AN_j \leftarrow AN_j \cup k \]

\[ c_t^j \leftarrow c_t^j + \sum_u d_t^{k \rightarrow j}(u)N_t^{k \rightarrow u} \]

\[ n_t^j = x_t^j - c_t^j \]

\[ j^* = \text{MinStationary}(\{n_t^j, i \in S\}) \]

if \( j^* \neq \emptyset \) then

\[ \{N_t^j\} = \{n_t^j\} \]

\[ PA_{j^*} = \text{SelectParents}(AN_{j^*}, \{x_t^j\}, \{d_t^{k \rightarrow j}\}) \]

\[ A(i,j^*) = 1, \forall i \in PA_{j^*} \]

end for

end while

end

**Procedure 1** MinStationary

**Input:** \( N \) time series \( \{x_t^j\} \), \( j \in J \)

**Output:** \( j^* \in J \)

\[ U = \{j \in J : \{x_t^j\} \text{ is UMP}\} \]

if \( U \neq \emptyset \) then

\[ j^* = \arg\min_{j \in U} ST(\{x_t^j\}) \]

else \( j^* = \emptyset \)

end

**Procedure 2** SelectParents

**Input:** \( AN_{j^*}, \{x_t^j\}, \{d_t^{k \rightarrow j}\} \)

**Output:** \( PA_{j^*} \)

for each \( Q \subseteq AN_{j^*} \) do

\[ \hat{W}_t(Q) = x_t^j - \sum_{m \in Q} \sum_u d_t^{m \rightarrow j}(u)x_t^m \]

\[ PA_{j^*} = \text{MinStationary}(\{\hat{W}_t(Q)\}) \]

end for

end

Our Algorithm 3 follows the main idea of the way we prove Theorem 2. In each iteration of the while loop, the task is to first identify the ancestors of one node in \( S \) and then select the parents from the ancestors, where \( S \) contains the nodes whose parents are unknown and \( S^c \) denotes the complement of \( S \). The order that the nodes leave the set \( S \) is a causal ordering. We will obtain an estimate of the residuals \( \{N_t^j\} \) if the ancestors of \( j \) are contained in \( S^c \). Later, the estimated residuals will be used to estimate the filters. There are three places in the algorithm where we need to select the time series that minimizes a certain stationarity measure, which is carried out in Procedure 1. Specifically, we use the UMP test (i.e., the interaction and residual variance \( S_{I+R} \) in Algorithm 2).
as a prescreening step and then compute the between time variance $S_T$ in Algorithm 2 to quantify the stationarity of the time series.

For the independence test between $\{x_t^j\}$ and $\{\hat{N}_t^k\}$, one can use the kernel independence test for random processes [36], which can be computationally demanding. An efficient approximation is to test whether $(1/T)|\mu_t| = \sum_t d_t^{i-j}(u) < a$ for $u = 1, \ldots, q$. In practice, when our model assumptions are violated, one can test the joint independence of the estimated residuals $\{\hat{N}_t^i\}, i \in V$, at the end of the algorithm. In Section V, this step is omitted since our algorithm is applied to the data generated by model (24).

Remark 9: From one synthetic data experiment, Experiment 5 in Section VI we observe that the selected parents $PA_j$ in Procedure 2 may be empty in certain cases due to the estimation procedure. In such cases, one could replace $PA_j$ with $AN_j$ in Procedure 2 resulting in additional edges in the inferred graph (which is a subgraph of the transitive closure of $G$ [43]). It is worth noting that the additional edges will not affect the causal ordering of nodes.

V. EXPERIMENTS

For all data sets, we use $x$ and $y$ to denote the true cause and effect, respectively. For the independence test, we use the default configuration of HSICp [36]. The significance level is denoted by $\alpha$ for the stationarity test, UMP test (i.e., the test on $S_{I+R}$ in algorithm 2) and the independence test, and we take $\alpha = 0.05$ for the independence test throughout this section. For the multitaper method, finding the optimal window size is notoriously hard even for stationary processes. We thus set the window size $N_F$ to be 128 for synthetic data, and the robustness of $N_F$ is tested in Experiment 1 as well as the real data simulations. In the synthetic experiments, since the true order $p$ is less than the maximum order $[N_F/2]$, i.e., the true model is in the model class, we adopt BIC for order selection since it is consistent. For real data, we test both AIC and BIC. The length of the processes $N$ is fixed to 2048 for all the synthetic data. We compare with TiMINo-linear [7], TCM [20], LiNGAM-t [2] and Granger causality [8].

A. Synthetic Data

Experiment 1: First-order Models. We consider first-order models from (26).

$$Y_t = X_t + a(t)X_{t-1} + \epsilon_t, \quad 0 \leq t \leq T - 1, \quad (27)$$

where $\{X_t\}$ is a UMP defined by $X_t = b(t)Z_t$, with $b(t)$ being a Gaussian kernel $\mathcal{N}(\mu_0, \sigma_0^2)$, and $\sigma_0 = 0.2T$, with the same ratios to $T$ as in (26). The process $\{Z_t\}$ is defined by a second-order AR model, $Z_t = 0.8Z_{t-1} - 0.4Z_{t-2} + \epsilon_t$, in which $\{\epsilon_t\}$ is a white Gaussian noise with $\epsilon_t \sim \mathcal{N}(0, 100^2)$. The stationary noise process\footnote{The only difference between the model in (27) and that in (26) is that the latter considers $\{N_t\}$ to be a UMP process.} $\{N_t\}$ is defined by $N_t = 0.8N_{t-1} - 0.16N_{t-2} + \epsilon_t$, where $\{\epsilon_t\}$ is a white Gaussian noise with $\epsilon_t \sim \mathcal{N}(0, \sigma_N^2)$.

1) Different frequencies: We first test how the window size $N_F$ and the frequency of the cosine function $a(t) = 0.5 \cos(t/L)$ affect the performance of our method. Let $\sigma_N = 25$, $L \in \{25, 50, 100, 200, 400\}$, and $N_F \in \{32, 64, 128, 256, 512\}$. For each set of parameters, we test 100 models. Fig. 1 shows that our method performs well for cosine functions with low frequencies ($L \geq 100$) regardless of the choice of $N_F$. For high frequencies (i.e., when $L$ is small), our method performs well only when $N_F$ is small. This aligns with the intuition that small $N_F$ can help reveal more high-frequency components.

2) Different SNRs: We now examine how sensitive our method is with respect to the SNR level. We use the parameter $\sigma_N$ to control the SNR level. Let $a(t) = 0.5 \cos(t/200)$. For each $\sigma_N$ in $\{5, 10, 15, 20, 25, 40, 50, 75, 100\}$, we test 100 models. For $\sigma_N = 25$, one can tell from Fig. 2a that the residuals of $x \rightarrow y$ are more likely to be nonstationary than the residuals of $y \rightarrow x$. Overall, the percentage of identifying the correct directions is above 80% for different SNRs. Though, Fig. 2b shows that the estimated cosine functions are noisier when the SNR is lower (i.e., when $\sigma_N$ is larger). Note that even for a fixed $\sigma_N$, the SNR changes over time with $\sigma(x)$ (Fig. 2a). The estimated functions are noisier at the start and end of the time range. This suggests that our method is relatively robust with respect to different SNR levels. Since changing the SNR is equivalent to scaling the amplitude of the filter differently, this experiment examines our method under different causal strengths.

Experiment 2: High-order Models. In this experiment, we focus on high-order models with smooth filters. We generalize model (27) to incorporate higher orders $p$ which is generate from Unif$\{1, \ldots, 5\}$. Specifically, we have

$$Y_t = c(t) \sum_{k=0}^p a_k(t)X_{t-k} + \epsilon_t, \quad 0 \leq t \leq T - 1, \quad (28)$$

where $c(t) = 1 + A \cos(t/L)$, $A \sim \text{Unif}[0.05, 0.2]$, $L \sim \text{Unif}[400, 800]$, $a_0(t) = 1$, $a_k(t) = g_k(f(P(t/T))) + S(t)$, $k \geq 1$, in which $S(t) = \sin(t - a)/b$, $\sin(t) \triangleq \sin(\pi t)/\pi t$, $a \sim \text{Unif}[300, 1500]$, $b \sim \text{Unif}[400, 800]$, and $P(t)$ is a polynomial function of degree $q \sim \text{Unif}[1, \ldots, 6]$.
Fig. 2: Synthetic data A2. (a-1) The standard deviation of \( \{X_t\} \), (a-2) The estimated residuals of \( x \rightarrow y \) when \( \sigma_N = 25 \), (a-3) The estimated residuals of \( y \rightarrow x \) when \( \sigma_N = 25 \). (b) Estimated \( \alpha(t) \)'s for different \( \sigma_N \). The red lines denote the ground truth, and the blue lines denote the estimate.

with roots sampled from \( \text{Unif}[-1, 1] \). The functions \( f \) and \( g_k \)'s are scaling functions defined as \( f(P(t)) = P(t)/\max_x(P(t)) \) and \( g_k(a_k(t)) = c_k a_k(t)/\max_x(a_k(t)) \), with \( c_k = (1/1.5)^k \). The process \( \{X_t\} \) is defined by \( X_t = Z_t + \sum_{k=1}^2 b_k(t)Z_{t-k} \), where \( b_k(t) \)'s are generated by the same way as \( a_k(t) \)'s. The stationary processes \( \{Z_t\} \) and \( \{N_t\} \) are each generated by the ARMA model, \( W_t = \Phi^2(B)W_t + \Theta^2(B) \varepsilon_t \), where \( \Phi^2(B) = (d_1 + d_2)B - d_1 d_2 B^2 \), \( \Theta^2(B) = 1 + (d_3 + d_4)B + d_3 d_4 B^2 \), and \( d_1, d_2, d_3, d_4 \sim \text{Unif}[-0.6, -0.1] \cup [0.1, 0.6] \). Note that \( 1/d_1 \) and \( 1/d_2 \) are the roots of the polynomial \( 1 - \Phi^2(z) \), and \( 1/d_3 \) and \( 1/d_4 \) are the roots of the polynomial \( \Theta^2(z) \). Since the roots are all strictly outside the unit circle, the randomly generated ARMA model has a unique stationary solution that is causal \([30]\). The white Gaussian noise \( \{e_t\} \) has variances \( \sigma_Z^2 = 100^2 \) and \( \sigma_N^2 = 25^2 \) for \( \{Z_t\} \) and \( \{N_t\} \), respectively. For \( \alpha = 0.01 \) and \( \alpha = 0.05 \), we test 1000 randomly generated models.

In Table I we classify the undecided cases into three categories: (1) The independence tests are significant for both directions; (2) The independence tests are not significant and the residual processes are stationary; (3) The independence tests are not significant and the residual processes are non-stationary. Both TiMINO \([7]\) and TCM \([20]\) remain undecided since the p-values are too small for both directions. Granger causality infers the correct (or wrong) direction for 41\% (or 2.8\%) of the models. LiNGAM-t infers the correct (or wrong) direction for 69\% (or 31\%) of the models.

| \( \alpha \) for the stationarity test | 0.01 \( x \rightarrow y \) | 0.05 \( x \rightarrow y \) | 0.01 \( y \rightarrow x \) | 0.05 \( y \rightarrow x \) |
|---|---|---|---|---|
| both not independent | 11.7 | 10.8 | 0.6 | 0.9 |
| both stationary | 1.0 | 0.4 | 0.6 | 0.9 |
| both nonstationary | 2.6 | 5.1 | 0.6 | 0.9 |

**Experiment 3: High-order Models with Gaussian/non-Gaussian UMP noise.** We replace the noise process in Experiment 2 with a UMP defined as \( N_t = g(t)W_t \), where \( g(t) = \exp((t - T/2)^2/(2\sigma_W^2)) \), \( \sigma_g \sim \text{Unif}[0.4T, 0.8T] \). \( \{W_t\} \) is simulated from the following three models: (1) the randomly generated ARMA model in Experiment 2 (i.e., \( \{Z_t\} \)) with \( \sigma_e = 25 \); (2) i.i.d. uniform with \( W_t \sim \text{Unif}[-30, 30] \); (3) \( W_t = 15V_i \), with \( \{V_i\} \) being a sequence of i.i.d. variables following the student’s t-distribution with degrees of freedom 5. We use a variant of Algorithm \([1]\) that replaces the stationarity test with a UMP test. We sample 1000 models from \( \{Z_t\} \) for each case of the UMP noise. Our method works well for the three cases (see Table I). In particular, the percentage of inferring the wrong direction is below 1\%. LiNGAM-t tends to infer more wrong directions and Granger causality performs worse than a random guess (i.e., 50\%). TCM and TiMINO-linear remain mostly undecided.

| \( \alpha(t) \) | 0.01 \( x \rightarrow y \) | 0.05 \( x \rightarrow y \) | 0.01 \( y \rightarrow x \) | 0.05 \( y \rightarrow x \) |
|---|---|---|---|---|
| both not independent | 11.7 | 10.8 | 0.6 | 0.9 |
| both stationary | 1.0 | 0.4 | 0.6 | 0.9 |
| both nonstationary | 2.6 | 5.1 | 0.6 | 0.9 |

**Experiment 4: Models with Non-smooth Functions.** We have demonstrated the performance of our method for a large class of smooth filters in Experiment 2, and we now examine a class of non-smooth functions. Consider model (27) with \( \alpha(t) = 0.5 \text{tri}(t/200; b), 0 \leq b < 1 \), which is a triangle wave function. The function \( \text{tri}(t; b) \) in one of its period \([0, 2\pi]\) is defined as

\[
\text{tri}(t; b) = \begin{cases} 
1 - \frac{4}{\pi}t, & 0 \leq t \leq 2\pi, b = 0 \\
\frac{1}{2\pi b - 2\pi b_0} + \frac{1}{\pi b_0}, & 0 \leq t \leq 2\pi, b \neq 0 \
2\pi b < t \leq 2\pi, b \neq 0
\end{cases}
\]

where \( (2\pi b, 1) \) is a vertex of the triangle function that changes with \( b \). The function \( \text{tri}(t; b) \) is a right triangle when \( b = 0 \), which leads to a discontinuous point at \( t = 0 \). Let \( \sigma_N = 25 \). For each \( b \) in \( \{0, 0.125, 0.25, 0.375, 0.5\} \), we test 100 models. Table III shows that our method performs well except for the case when the triangle wave function has discontinuous points.

| \( b \) | \( x \rightarrow y \) | \( y \rightarrow x \) |
|---|---|---|
| 0 | 33 | 4 |
| 0.125 | 89 | 2 |
| 0.25 | 86 | 2 |
| 0.375 | 86 | 2 |
| 0.5 | 87 | 3 |

**Experiment 5: Network Setting.** We consider randomly generated DAGs with number of nodes \( N_G \sim \text{Unif}\{2, \ldots, 5\} \). Each edge is included with probability 0.6. In model \([28]\),
let \( \{ N^t \} \) be an i.i.d. Gaussian process with zero mean and variance \( \sigma^2 \sim \text{Unif}[5, 10] \). Each time-varying filter is defined in the same way as the bivariate model \( \mathcal{M} \), where \( a_k(t) \) has order 2, \( c(t) = 1 + A \cos(t/L) \), with \( A \sim \text{Unif}[0.5, 2] \) and \( L \sim \text{Unif}[400, 800] \). We test Algorithm 5 with \( \alpha = 0.01 \) and \( \alpha = 0.05 \) for the UMP test, respectively. To approximately test the independence of \( \{ x^t \} \) and \( \{ N^t \} \) in Algorithm 5, we test whether \( |\mu| = (1/T) \sum \sigma^{-1}(t) \) is smaller than \( a \), \( a = 0.15 \), for \( u = 1, \ldots, 3 \). We test the sensitivity of our method with respect to parameter \( a \) for \( a = \{0.1, 0.125, 0.15, 1.75\} \) when \( \alpha = 0.05 \).

The results are classified into three categories: (1) the inferred graph is correct; (2) the inferred graph is a proper subgraph of the correct graph (i.e., correct graph with missing edges); (3) all other cases (i.e., correct graph with additional edges or correct graph with both missing and additional edges). We refer to the proper subgraph as p-subgraph in Table IV. When a proper subgraph is inferred, it means that our method tends to remain conservative. Our method mostly infers either the correct graph or a proper subgraph of the correct graph. Both Granger causality and LiNGAM-t give a large percentage of wrong graphs that include wrong edges. TiMINo remains mostly undecided. Overall, our method is relatively robust with respect to \( a \) since the results mainly fall into the first two categories (see Table IV). Our method infers more cases of wrong edges as \( a \) gets smaller, thus we suggest using \( a \geq 0.1 \).

**Table IV: Experiment 5 results (%)**

|                | Ours  | Granger | LiNGAM-t |
|----------------|-------|---------|----------|
| correct graph  | 91.6  | 81.7    | 32.4     | 13.0     |
| p-subgraph     | 5.4   | 14.2    | 14.1     | 9.5      |
| others         | 3.0   | 4.1     | 53.5     | 77.5     |

**Table V: Sensitivity test of \( \alpha \) (%)**

| \( \alpha \)   | 0.1  | 0.125 | 0.15 | 0.175 |
|-----------------|------|-------|------|-------|
| correct graph   | 74.7 | 77.2  | 81.7 | 82.6  |
| p-subgraph      | 14.2 | 16.5  | 14.2 | 13.6  |
| others          | 11.1 | 6.3   | 4.1  | 3.8   |

### B. Real Data

Let \( \alpha = 0.05 \). The only preprocessing needed for our method is detrending. Since the first two data sets are too short in length for HSICp, we infer the causal directions using HSIC. In all experiments, we fix \( N_F = 64 \) but the p-values remain to be similar for \( N_F = 32 \) or 128. It turns out that the p-values are similar under AIC or BIC. We set the maximum order for the compared methods (i.e., TiMINo, TCM, LiNGAM-t, Granger causality) to be 10. In all three experiments below, TCM \( \mathcal{M} \) remains undecided.

**Experiment 6: Input Gas Rate vs. Output CO2.** The Gas Furnace data set \( \mathcal{M} \) has two variables: the input gas rate \( x \) and the output CO2 \( y \). Our method with HSIC yields the correct direction with p-values \( p_{x \rightarrow y} = 0.0895 \) and \( p_{y \rightarrow x} = 5.511 \cdot 10^{-11} \). TiMINo and Granger causality correctly infer the direction \( x \rightarrow y \) \( \mathcal{M} \), whereas LiNGAM-t gives the wrong direction.

**Experiment 7: Duration vs. Time Interval.** The Old Faithful data set \( \mathcal{M} \) contains two variables observed from the Old Faithful geyser: the duration of an eruption \( x \) and the time interval before the next eruption \( y \). As mentioned in \( \mathcal{M} \), the data is not collected with fixed time resolution, but we treat the data as time series. Our method equipped with HSIC leads to p-values \( p_{x \rightarrow y} = 0.0777 \) and \( p_{y \rightarrow x} = 9.364 \cdot 10^{-10} \). TiMINo and LiNGAM-t infer the correct direction \( \mathcal{M} \). Granger causality gives the wrong direction.

**Experiment 8: Breathing Pattern vs. Heart Rate.** The modulation of heart rate by the breathing pattern is known as respiratory sinus arrhythmia (RSA) \( \mathcal{M} \). This phenomenon is widely observed, especially among young and healthy individuals. We use the data of sleeping patients \( \mathcal{M} \) to verify the causal relationship between breathing pattern and heart rate. This dataset contains two variables: chest volume \( x \) and heart rate \( y \); the interaction between \( x \) and \( y \) is potentially complicated. Nevertheless, the mechanism of respiratory sinus arrhythmia (RSA) is often understood to be the modulation of heart rate by the breathing rhythm, and the degree of RSA increases at rest and during sleep \( \mathcal{M} \). Thus it is reasonable to take the causal direction from \( x \) to \( y \) as the underlying truth. The challenging parts of the data are the nonstationarity and the seasonality of the data. Taking a segment of length 2048 from the data (i.e., samples \( 2000 \sim 4047 \) of data set \( B_1 \), our method yields p-values \( p_{x \rightarrow y} = 0.0610 \) and \( p_{y \rightarrow x} = 0.004 \). Since the stationarity tests are always significant, we conclude the causal directions based only on HSICp. The results show that our method gives the correct direction. TiMINo remains undecided due to small p-values \( (< 10^{-10}) \) for both directions, and this might be because it requires the stationarity of the data for the estimation procedure. Granger causality infers the correct direction, while LiNGAM-t gives the wrong direction.

**VI. DISCUSSION**

For linear Gaussian models, causal structures are generally nonidentifiable. By introducing nonstationarity to linear Gaussian models through time-varying filters, we establish identifiability results for both the bivariate and network settings, where we consider both finite and infinite sequences. For causal inference in the network setting, we propose a novel estimation procedure for time-varying filters in the multivariate case. From a technical perspective, our work provides two sufficient conditions and one necessary condition for the invertibility of time-varying lag operators, and we draw new connections between the slowing varying condition and invertibility. Compared with previous works, our modeling of nonstationary processes and our causal inference procedures are well grounded in the evolutionary spectra framework. The empirical performance of our method compares favorably with existing methods on various real and synthetic datasets, and the robustness has been examined in terms of weakened causal strength, non-Gaussian noise, or non-smooth filters.
Several future directions and extensions are possible and we briefly discuss them as follows. It is of importance to provide theoretical guarantees for models where the stationary noise assumption is relaxed. To this end, we have made an attempt to handle the nonstationary UMP noise theoretically, which is also supported by our empirical experiments. Another fundamental direction is to investigate the impact of confounders on our setting. For instance, one interesting question is to characterize the conditions on nonstationarity (beyond Example 4) that can help avoid spurious causal relations due to confounders.

VII. ACKNOWLEDGEMENT

The authors would like to thank Jie Ding and Kun Zhang for inspiring discussions.

REFERENCES

[1] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, “A linear non-Gaussian acyclic model for causal discovery,” *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

[2] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, “Nonlinear causal discovery with additive noise models,” in *Advances in Neural Information Processing Systems*, 2009, pp. 689–696.

[3] K. Zhang and A. Hyvärinen, “On the identifiability of the post-nonlinear causal model,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 647–655.

[4] J. Peters and P. Bühlmann, “Identifiability of Gaussian structural equation models with equal error variances,” *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[5] J. Pearl, “Causality: Models, reasoning and inference,” Cambridge, UK: Cambridge University Press, 2000.

[6] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, “Causal discovery with continuous additive noise models,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2009–2053, 2014.

[7] J. Peters, D. Janzing, and B. Schölkopf, “Causal inference on time series using restricted structural equation models,” in *Advances in Neural Information Processing Systems*, 2013, pp. 154–162.

[8] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[9] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, “Estimation of a structural vector autoregression model using non-Gaussianity,” *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[10] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*, 2017.

[11] A. Shoaie and E. B. Fox, “Granger causality: A review and recent advances,” *Annual Review of Statistics and Its Application*, vol. 9, pp. 289–319, 2022.

[12] Y. Liu, M. Taniguchi, and H. Ombao, “Statistical inference for local granger causality,” arXiv preprint arXiv:2103.00209, 2021.

[13] D. Dahlhaus, “On the Kullback-Leibler information divergence for locally stationary processes,” *Stochastic Processes and their Applications*, vol. 62, pp. 139–168, 1996.

[14] N. Shajarisales, D. Janzing, B. Schölkopf, and M. Bassever, “Telling cause from effect in deterministic linear dynamical systems,” in *International Conference on Machine Learning*, 2015, pp. 285–294.

[15] J. Massey et al., “Causality, feedback and directed information,” in *Proc. Int. Symp. Inf. Theory Appl.* (ISITA-90). Citeeseer, 1990, pp. 303–305.

[16] G. Kramer, *Directed information for channels with feedback*. Citeeseer, 1998.

[17] P.-O. Amblard and O. J. Michel, “The relation between granger causality and directed information theory: A review,” *Entropy*, vol. 15, no. 1, pp. 113–143, 2013.

[18] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, “Estimating the directed information to infer causal relationships in ensemble neural spike train recordings,” *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[19] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, “Universal estimation of directed information,” *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[20] B. Huang, K. Zhang, and B. Schölkopf, “Identification of time-dependent causal model: A Gaussian process treatment,” in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[21] B. Huang, K. Zhang, M. Gong, and C. Glymour, “Causal discovery and forecasting in nonstationary environments with state-space models,” in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 2901–2910.

[22] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, “Causal discovery from heterogeneous/nonstationary data,” *Journal of Machine Learning Research*, vol. 21, no. 89, pp. 1–53, 2020.

[23] N. Pfister, P. Bühlmann, and J. Peters, “Invarient causal prediction for spatio-temporal data,” *Journal of the American Statistical Association*, vol. 114, no. 527, pp. 1264–1276, 2019.

[24] M. B. Priestley, “Evolutionary spectra and non-stationary processes,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 153–166, 1973.

[25] T. S. Rao and H. Tong, “A test for time-dependence of linear open-loop systems,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 235–250, 1972.

[26] Y. Xiang, J. Ding, and V. Tarokh, “Estimation of the evolutionary spectra with application to stationarity test,” *IEEE Transactions on Signal Processing*, vol. 67, no. 5, pp. 1353–1365, 2019.

[27] K. Du and Y. Xiang, “Causal inference using linear time-varying filters with additive noise,” in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 896–901.

[28] R. A. Davis and P. J. Brockwell, *Time Series: Theory and Methods*. Springer-Verlag, 1987.

[29] M. Hallini, “Non-stationary q-dependent processes and time-varying moving-average models: invertibility properties and the forecasting problem,” *Advances in applied probability*, pp. 170–210, 1986.

[30] N. Abdubbo and M. Priestley, “On the prediction of non-stationary processes,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 29, no. 3, pp. 570–585, 1967.

[31] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in *Proceedings of the 2nd International Symposium on Information Theory*, 1973, p. 267–281.

[32] G. Schwarz et al., “Estimating the dimension of a model,” *Annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[33] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, “A kernel statistical test of independence,” in *Advances in neural information processing systems*, 2008, pp. 585–592.

[34] K. Chwialkowski and A. Gretton, “A kernel independence test for random processes,” in *International Conference on Machine Learning*, 2014, pp. 1422–1430.

[35] D. B. Percival and A. T. Walden, *Spectral Analysis for Univariate Time Series*. Cambridge University Press, 2020, vol. 51.

[36] W. Constantine and D. Percival, “Fractal: fractal time series modeling and analysis,” *R package version*, 2011.

[37] D. J. Thomson, “Spectrum estimation and harmonic analysis,” *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[38] A. Rupasinghe and B. Babadi, “Multitaper analysis of semi-stationary spectra from multivariate neuronal spiking observations,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 4382–4396, 2020.

[39] G. Mélard and A. H.-d. Schutter, “Contributions to evolutionary spectral analysis,” *Journal of Time Series Analysis*, vol. 10, no. 1, pp. 41–63, 1989.

[40] M. B. Priestley and T. S. Rao, “A test for non-stationarity of time-series,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 140–149, 1969.

[41] J. Bang-Jensen and G. Z. Gutin, *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.

[42] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[43] A. Azzalini and A. W. Bowman, “A look at some data on the old faithful geyser,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 39, no. 3, pp. 357–365, 1990.

[44] R. A. Davis and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: [http://archive.ics.uci.edu/ml]

[45] J. A. Hirsch and B. Bishop, “Respiratory sinus arrhythmia in humans: how breathing pattern modulates heart rate,” *American Journal of
Physiology-Heart and Circulatory Physiology, vol. 241, no. 4, pp. H620–H629, 1981.

[48] D. R. Rigney, A. L. Goldberger, W. C. Ocasio, and Y. Ichimaru, “Multi-channel physiological data: description and analysis (data set B),” in Time Series Prediction: Forecasting the Future and Understanding the Past, vol. 15. MA: Addison-Wesley, 1993, pp. 105–129.

[49] J. Hayano, F. Yasuma, A. Okada, S. Mukai, and T. Fujinami, “Respiratory sinus arrhythmia: a phenomenon improving pulmonary gas exchange and circulatory efficiency,” Circulation, vol. 94, no. 4, pp. 842–847, 1996.

[50] R. A. Horn and C. R. Johnson, Matrix analysis. Cambridge university press, 2012.

[51] E. Key and H. Volkmer, “A note on the spectral radius of a product of companion matrices,” The Electronic Journal of Linear Algebra, vol. 27, 2014.

[52] P. Bauer, M. Mansour, and J. Duran, “Stability of polynomials with time-variant coefficients,” IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 40, no. 6, pp. 423–426, 1993.
APPENDIX A
PROOF OF THEOREM 1

Proof: Suppose there exists a backward model of the form
\[ X_t = \Psi_t^q(B)Y_t + \tilde{N}_t, \quad \{\tilde{N}_t\} \perp \{Y_t\}, \quad q \in \mathbb{Z}_{\geq 0}, \tag{29} \]
where \{\tilde{N}_t\} is stationary.

(I) First, we show the condition for the independence constraint. Let \( O \) denote the class of operators such that \( \{Y_t\} \perp \{X_t - H_t^*(B)Y_t\} \) for any \( H_t^*(B) \in O, s \in \mathbb{Z}_{\geq 0}. \) Note that \( O \) is nonempty since \( \Psi_t^q(B) \in O. \) For any \( H_t^*(B) \in O, \) we obtain the model
\[ X_t = H_t^*(B)Y_t + W_t, \tag{30} \]
where \( \{W_t\} \) is defined by \( W_t = X_t - H_t^*(B)Y_t. \) We replace \( Y_t \) in (30) by that in (13), and have
\[ W_t = (1 - H_t^*(B)\Phi_t^p(B))X_t - H_t^*(B)N_t. \tag{31} \]

Now, we prove that the Gaussian process \( \{H_t^*(B)Y_t\} \) has the same distribution for any \( H_t^*(B) \in O, \) which is equivalent to prove that \( \{H_t^*(B)Y_t\} \) and \( \{\Psi_t^q(B)Y_t\} \) have the same covariance function. According to the independence constraint, we know that both \( \{W_t\} \) and \( \{\tilde{N}_t\} \) are independent of \( \{Y_t\}, \) which implies that \( \{W_t - \tilde{N}_t\} \perp \{Y_t\}. \) Then, subtracting (29) from (30) yields \( W_t - \tilde{N}_t = (\Psi_t^q(B) - H_t^*(B))Y_t. \)

It follows from Lemma 4 that the covariance function of \( \{\Psi_t^q(B) - H_t^*(B)\}Y_t \) is zero. Thus, we have shown that \( \{H_t^*(B)Y_t\} \) has the same distribution for any \( H_t^*(B) \in O. \)

Moreover, we provide an explicit characterization of the operators in \( O \) as follows. Since two Gaussian processes are independent if and only if their cross-covariance function equals to zero, from (13) and (31), we have
\[
\text{Cov}(Y_t, W_t) = \text{Cov}(\Phi_t^p(B)Y_t + N_t, \Phi_t^p(B)X_t - H_t^*(B)N_t)
= H_t^*(B)(1 - H_t^*(B)\Phi_t^p(B))\gamma_{XX}(t_1, t_2)
- H_t^*(B)\gamma_{YN}(t_2 - t_1) = 0. \tag{32}\]

Since we assume \( \Phi_t^p(B) \) is invertible, (32) can be written as
\[
(1 - H_t^*(B)\Phi_t^p(B))\gamma_{XX}(t_1, t_2)
= H_t^*(B)(\Phi_t^p(B))^{-1}\gamma_{YN}(t_2 - t_1), \tag{33}\]
which can be further simplified as
\[
\gamma_{XX}(t_1, t_2) = H_t^*(B)(\Phi_t^p(B))^{-1}\gamma_{YN}(t_2 - t_1). \tag{34}\]

Therefore, we have shown that (32) determines a class of operators in \( O \), simply because that (34) is equivalent to the independence of \( \{Y_t\} \) and \( \{W_t\}. \)

(II) Now we move on to prove the condition for the stationarity constraint. Since when \( H_t^*(B) = \Psi_t^q(B), \) we obtain \( \{W_t\} = \{N_t\} \) in (30). Thus, there exists \( H_t^*(B) \in O \) such that \( \{W_t\} = \{X_t - H_t^*(B)Y_t\} \) is stationary. The stationarity of \( \{W_t\} \) implies that its covariance function is a function of \( t_2 - t_1. \) From (31) and (33), we have
\[ \text{Cov}(W_{t_1}, W_{t_2}) = (1 - H_{t_1}^*(B)\Phi_{t_1}^p(B))(1 - H_{t_2}^*(B)\Phi_{t_2}^p(B))\gamma_{XX}(t_1, t_2)
+ H_{t_1}^*(B)H_{t_2}^*(B)\gamma_{NN}(t_2 - t_1)
= H_{t_2}^*(B)(\Phi_{t_1}^p(B))^{-1}\gamma_{NN}(t_2 - t_1). \tag{35} \]

Let \( t_1 = t_2 = t \) in (35) and \( \Theta_t^q(B) = (\Phi_t^p(B))^{-1}, \) we obtain the variance of \( \{W_t\} \) as
\[
\text{Var}(W_t) = \sum_{j=1}^{p} \sum_{k=1}^{r} \eta_{j,k} \Theta_{t-j-k} \gamma_{NN}(k - j), \tag{36}\]
which is time-invariant.

APPENDIX B
PROOF OF COROLLARY 1

Proof: Note that \( E[X_iY_{i-s}] = \sigma_X^2 1_{i=s} \) since \( \{X_t\} \) is an i.i.d. process. We solve for the coefficients of \( \Psi_t^q(B) \) by computing \( E[X_iY_{i+p-i}] \), \( i \geq 0. \) Using model (13), we derive
\[
E[X_iY_{i+p-i}] = \sum_{j=0}^{p} \phi_{i+p-i,j} E[X_iX_{i+p-i-j}]
= \phi_{i+p-i,p-i} \sigma_X^2, \tag{37}\]
for \( i \geq 0. \) Using the backward model of (13), \( E[X_iY_{i+p-i}] \) can be computed alternatively as
\[
E[X_iY_{i+p-i}] = E[Y_{i+p-i}\left(\sum_{j=0}^{q} \psi_{i,j}Y_{i-j} + \tilde{N}_t\right)]
= \sum_{j=0}^{q} \psi_{i,j} E[Y_{i+p-i}Y_{i-j}]
= \sum_{j=0}^{q} \psi_{i,j} E[Y_{i+p-i}Y_{i-j}],
\]
where the last equality holds since \( |t + p - i - (j - t)| \leq p \) is equivalent to \( i - 2p \leq j \leq i. \) Thus, we obtain the following relationship for \( i \geq 0, \)
\[
\sum_{j=\max(0,i-2p)}^{i} \psi_{i,j} E[Y_{i+p-i}Y_{i-j}] = \phi_{i+p-i,p-i} \sigma_X^2.
\]
This can be written explicitly for \( i = 0 \) and \( i > 1 \) as follows,
\[
\psi_{i,0} E[Y_{i+p}Y_{i}] = \phi_{i+p,p} \sigma_X^2, \quad \text{for } i = 0
\]
and for \( i \geq 1, \)
\[
\psi_{i,j} E[Y_{i+p-i}Y_{i-j}] = \phi_{i+p-i,p-j} \sigma_X^2
- \sum_{j=\max(0,i-2p)}^{i-1} \psi_{i,j} E[Y_{i+p-i}Y_{i-j}]. \tag{38}\]

To further simplify the expression, we now show that \( E[Y_{i+p-i}Y_{i-j}] \neq 0. \) First observe that
\[
E[Y_{i}Y_{s}] = E\left[\sum_{j=0}^{p} \phi_{t,j}X_{t-j} + N_{t}\left(\sum_{k=0}^{p} \phi_{s,k}X_{s-k} + N_{s}\right)\right]
= \sigma_X^2 \sum_{j=0}^{p} \phi_{t,j} \phi_{s,s-t+j} + \sigma_N^2 I_{t=s},
\]
and \( E[Y_t Y_s] = 0 \) when \(|t - s| > p\). Thus we have, for \( i \geq 0 \),

\[
E[Y_{t+p-i} Y_{t-1}] = \sigma_X^2 \sum_{j=0}^{p} \phi_{t-i,j} \phi_{t+p-i,p+j}
\]

\[
= \sigma_X^2 \phi_{t+p-i,p} \phi_{t-i,0} \neq 0,
\]

which follows from the fact that \( \phi_{t+p-i,p+j} = 0 \) for \( j \geq 1 \).

As a result, we can divide \( E[Y_{t+p-i} Y_{t-1}] \) on both sides of (37) and obtain the solution

\[
\psi_{t,i} = -\frac{\sigma_X^2 \phi_{t-i,p}}{\sigma_X^2 \phi_{t+p-i,p} \phi_{t-i,0}} + \frac{\phi_{t+p-i,p} \phi_{t-i,0}}{\sigma_X^2 \phi_{t+p-i,p} \phi_{t-i,0}} - \sum_{j=\max(0,i-2p)}^{i-1} \psi_{t,j} \phi_{t+p-i,k} \phi_{t-j,-j-p} \phi_{t+p-i,p} \phi_{t-i,0}
\]

(38)

for \( i \geq 1 \). Similarly, we obtain \( \psi_{t,0} = 1/\phi_{t,0} \) for \( i = 0 \).

To obtain equation (14), we will need a technical lemma (postponed to be presented in Lemma 5 below), which shows that \( \Psi_t^*(B) \) is simply the inverse operator of \( \Psi_t^p(B) \) if \( \sigma_X^2 = 0 \). Thus the last two terms in (38) can be replaced by the coefficient of the inverse operator of \( \Phi_t^p(B) \) according to (15).

Therefore, the coefficients of \( \Psi_t^*(B) \) can be solved iteratively for all \( i \geq 0 \) using (14).

Now, we move on to the condition for the stationarity constraint. In Theorem 1 if \( \{N_t\} \) is an i.i.d. process, then (36) can be written as \( \text{Var}(W_t) = \sum_{j=1}^{\min(s,r)} \gamma_{t-j}\theta_{t-j,0}\gamma_{N(N)(0)} \), which in turn equals to \( \gamma_{N(N)(0)} \sum_{j=1}^{\min(s,r)} \eta_{t-j,0} \). Since the operator \( H_t^*(B) \) in the backward model (30) is uniquely determined by (14), we have \( \{W_t\} = \{N_t\} \) and \( \{\eta_{t,j}\} = \{\psi_{t,j}\} \). Thus, \( \text{Var}(N_t) = \sum_{j=1}^{\min(s,r)} \psi_{t,j} \psi_{t-j,0}\gamma_{N(N)(0)} \). The rest follows by invoking equation (15).

It remains to show the following technical lemma.

**Lemma 3:** Let \( \{X_t\} \) and \( \{N_t\} \) be i.i.d. processes. If \( \sigma_X^2 = 0 \) and \( \sigma_N^2 \neq 0 \), then \( \Psi_t^*(B) \) is the inverse operator of \( \Psi_t^p(B) \).

**Proof:** By replacing \( Y_t \) in the backward model (2) with the forward model (2), we obtain

\[
X_t = \Psi_t^*(B) \Phi_t^0(B) X_t + \Psi_t^p(B) N_t + \tilde{N}_t,
\]

(39)

where \( \{X_t\} \) and \( \{N_t\} \) are i.i.d. processes. By multiplying \( X_t \) and taking expectation to both sides of (39), we obtain

\[
E[X_t^2] = \psi_{t,0} \phi_{t,0} E[X_t^2] + E[X_t \tilde{N}_t]
\]

\[
\psi_{t,0} \phi_{t,0} E[X_t^2] + E[\tilde{N}_t^2],
\]

where the last equality is obtained using the backward model. By same argument leading up to (38), we have \( \psi_{t,0} \phi_{t,0} = 1 \), which implies \( E[\tilde{N}_t^2] = 0 \). It follows that \( E[N_t \tilde{N}_{t-k}] = 0 \) for any \( k \in \mathbb{Z} \) by the Cauchy-Schwarz inequality. Let \( \Theta_t^* \triangleq \Psi_t^*(B) \Phi_t^0(B) \). Similarly, by multiplying \( X_{t-k}, k \geq 1 \), and taking expectation to both sides of (39), we obtain

\[
0 = E[X_t X_{t-k}] = \theta_{t,k} E[X_{t-k}],
\]

which implies that \( \theta_{t,k} = 0 \) for all \( k \geq 1 \). Then \( \Theta_t^* = \Psi_t^*(B) \Phi_t^0(B) = 1 \). Therefore, \( \Psi_t^*(B) \) is the inverse operator of \( \Phi_t^0(B) \).

**APPENDIX C**

**Proof of Corollary 2**

**Proof 1:** We solve for the operator \( \Psi_t^*(B) \) by computing \( E[X_t Y_{t-j}] \) in two ways using the forward model (16) and the backward model (5), respectively. Using model (16) we have \( E[X_t Y_{t-j}] = \phi(t) \sigma_X^2 \mathbb{I}_{j=0} \), while model (3) leads to \( E[X_t Y_{t-j}] = \psi_{t,j} \phi_j^2 (t-j) \sigma_X^2 + \sigma_X^2 \). Thus we obtain \( \psi_{t,0} = \phi(t) / (\phi(t) + \sigma_N^2 / \sigma_X^2) \) and \( \psi_{t,j} = 0 \) for \( j \geq 1 \), which gives the coefficient of \( Y_t \) in (17).

Then, by substituting (16) into (17), we obtain

\[
\tilde{N}_t = \frac{\sigma_X^2 / \sigma_N^2 \phi(t)}{\phi(t) + \sigma_N^2 / \sigma_X^2} X_t - \frac{\phi(t)}{\phi(t) + \sigma_N^2 / \sigma_X^2} N_t,
\]

which shows that \( \{\tilde{N}_t\} \) is a sequence of independent random variables. The rest is followed by computing the variance of \( \{\tilde{N}_t\} \) and write \( \tilde{N}_t = \sqrt{\text{Var}(N_t)} W_t \), where \( \{W_t\} \) is an i.i.d. process with \( \sigma_W^2 = 1 \).

**APPENDIX D**

**Proof of Proposition 1**

By the monotone convergence theorem and the finiteness of \( \sum_{j=0}^{\infty} |\psi_{t,j}| \) and \( \sup_t E[|X_t|^2] \), we have

\[
E \left[ \sum_{j=0}^{\infty} |\psi_{t,j}| |X_{t-j}| \right] = \lim_{n \to \infty} E \left[ \sum_{j=0}^{n} |\psi_{t,j}| |X_{t-j}| \right]
\]

\[
\leq \lim_{n \to \infty} \left( \sum_{j=0}^{n} |\psi_{t,j}| \right) \sup_t E[|X_t|^2] < \infty,
\]

which shows that \( \sum_{j=0}^{\infty} |\psi_{t,j}| |X_{t-j}| \) is finite with probability one. If \( \sup_t E[|X_t|^2] \) is finite with probability one.

\[
E \left[ \sum_{m<j \leq n} |\psi_{t,j} X_{t-j}|^2 \right] = \sum_{m<j \leq n} \sum_{m<k \leq n} |\psi_{t,j} \tilde{\psi}_{t,k} E[X_{t-j} X_{t-k}]| \leq \left( \sum_{m<j \leq n} |\psi_{t,j}| \right) \sup_t E[|X_t|^2] \to 0,
\]

as \( n, m \to \infty \). Thus, by Cauchy criterion, the series \( \sum_{j=0}^{\infty} \psi_{t,j} X_{t-j} \) converges in mean square. Finally, let \( S \) denote the mean square limit and by Fatou’s lemma,

\[
E[|S - \Psi_t^*(B) X_t|^2] = E \left[ \liminf_{n \to \infty} \left| S - \sum_{j=0}^{n} \psi_{t,j} X_{t-j} \right|^2 \right]
\]

\[
\leq \liminf_{n \to \infty} E \left[ \left| S - \sum_{j=0}^{n} \psi_{t,j} X_{t-j} \right|^2 \right] = 0,
\]

which shows that the mean square limit \( S \) and \( \Psi_t^*(B) X_t \) are equal with probability one.
Proof of Proposition 2

Proof: The proof consists of two cases regarding $q$. For $T$ large enough, it suffices to consider two cases $0 < q < T$ and $q = \infty$. For $0 < q < T$, since the independence constraint of the backward model (3) implies $Y_t \perp N_t$ for each $t$, the coefficients of $\Psi_t^q(B)$ are determined by (20) as well. Therefore, we have $\Psi_t^q(B)Y_t - b_t^q + Y_t = 0$.

For $q = \infty$, first recall that a valid backward model (3) requires $\sum_{j=0}^{\infty} |\psi_{t,j}| < \infty$ for each $t \in \mathbb{Z}$. Using (3), we have

$$\text{Cov}(Y_t, X_t) = \text{Cov}(Y_t, \Psi_t^\infty(B)Y_t)$$

$$= \text{Cov}(Y_t, Y_t)\Psi_t + \sum_{j=T+1}^{\infty} \psi_{t,j}Y_{t-j})$$

$$= \text{Cov}(Y_t, Y_t)\Psi_t + o(1),$$

where we denote $\Psi_t = (\psi_{t,0}, \psi_{t,1}, \ldots, \psi_{t,T})^T$. The equality (40) uses the facts that $\sup_t E[|Y_t|^2] < \infty$ for the Gaussian process $\{Y_t\}$ and $\sum_{j=T+1}^{\infty} |\psi_{t,j}| = o_p(1)$ since $\Psi_t^\infty(B)Y_t$ converges absolutely with probability one by Proposition 1. Then, multiplying $\text{Cov}^{-1}(Y_t, Y_t)$ on both sides of (40) yields

$$\Psi_t = \text{Cov}^{-1}(Y_t, Y_t)\text{Cov}(Y_t, X_t) + o(1)$$

where $\inf_{T \geq 0} \lambda_{\text{min}}(\text{Cov}(Y_t, Y_t)) > 0$ guarantees that $\text{Cov}(Y_t, Y_t)$ is invertible as $T \to \infty$. Therefore,

$$\Psi_t^\infty(B)Y_t - b_t^\infty Y_t$$

$$= (\Psi_t^T - b_t^T)Y_t + \sum_{j=T+1}^{\infty} \psi_{t,j}Y_{t-j} = o_p(1).$$

Lemma 4: For a Gaussian process $\{Y_t\}$ and a lag operator $\Phi_p^\infty(B), p \in \mathbb{Z}_{\geq 0}$, we have that $\{Y_t\}$ is independent of $\Phi_p^\infty(B)Y_t$ only if the covariance function of $\{\Phi_p^\infty(B)Y_t\}$ is zero.

Proof: The claim is trivial when $\Phi_p^\infty(B) = 0$. Suppose that $\{Y_t\}$ and $\{\Phi_p^\infty(B)Y_t\}$ are independent, which implies that

$$\text{Cov}(Y_{t+1}, \Phi_p^\infty(B)Y_{t}) = \Phi_p^\infty(B)\gamma_tY_{t+1} = 0.$$

By applying $\Phi_p^\infty(B)$ to (44), we have that

$$\Phi_p^\infty(B)\Phi_p^\infty(B)\gamma_tY_{t+1} = 0,$$

where the left-hand side is simply the covariance function of $\{\Phi_p^\infty(B)Y_t\}$.

Lemma 5: For $A \in \mathbb{C}^{n \times n}$ and $x \in \mathbb{C}^{n \times 1}$, we have

1. $||Ax||_p \leq ||A||_p ||x||_p$.
2. $||A_1A_2 \ldots A_k||_p \leq ||A_1||_p ||A_2||_p \ldots ||A_k||_p$, where $A_1, A_2, \ldots, A_k \in \mathbb{C}^{n \times n}$.
3. For any two matrix norms, there exists a constant $0 < C_{\alpha,\beta} < \infty$ such that $||A||_\alpha \leq C_{\alpha,\beta}||A||_\beta$ for any matrix $A \in \mathbb{C}^{n \times n}$.
4. For any $\delta > 0$, there exists a matrix norm $||\cdot||_\delta$ such that $0 \leq ||A||_\delta \leq \rho(A) + \delta$.

The last two lemmas focus on a particular form of matrix called the companion matrix (51). For a product of companion matrices, the following lemma provides a condition for its spectral radius to be bounded by an exponentially decreasing sequence.

Lemma 6 (51): Let $A_i \in \mathbb{R}^{n \times n}, 1 \leq i \leq k$, be companion matrices of the form

$$A_i = \begin{bmatrix} 0 & \cdots & I_{n-1} \\ -a_{1,n} & \cdots & -a_{i-1,n} \\ \vdots & \ddots & \vdots \\ -a_{1,1} & \cdots & -a_{i,n} \\ 0 & \cdots & 0 \\ \end{bmatrix},$$

where $I_n$ denotes the $n \times n$ identity matrix. If $a_{i,0} \geq 1 > a_{i,1} > \cdots > a_{i,n-1} > a_{i,n} \geq 0$ for each $i$, then there exists

$$\varepsilon \geq \max_{1 \leq i \leq k, 1 \leq j \leq n} \frac{a_{i,j}}{a_{i,j-1}} < 1$$
such that \( \rho(A_k A_{k-1} \ldots A_1) \leq \epsilon^k < 1 \).

Finally, we establish the following lemma on a product of companion matrices inspired by \( [52] \).

**Lemma 7:** Let \( A_i \in \mathbb{R}^{n \times n}, 1 \leq i \leq n \), be companion matrices of the form in \( [45] \). If
\[
0 < \sum_{j=1}^{n} |a_{i,j}| < 1
\]
for \( 1 \leq i \leq n \), then there exists some \( 0 < \epsilon < 1 \) such that \( ||A_i A_{i-1} \ldots A_1||_\infty \leq \epsilon \). Moreover, for companion matrices \( A_i \) with \( 1 \leq i \leq Nn \) and \( N \geq 1 \), if \( (46) \) holds for \( 1 \leq i \leq Nn \), then
\[
||A_{Nn} A_{Nn-1} \ldots A_1||_\infty \leq \epsilon^N
\]
for some \( 0 < \epsilon < 1 \).

**Proof:** Let \( T_i \triangleq A_i A_{i-1} \ldots A_1, 1 \leq i \leq n \), and let \( t^{(i)}_j, 1 \leq j \leq n \), denote the \( j \)th row of \( T_i \). We first claim that
\[
||t^{(i)}_j||_1 = \begin{cases} 1, & 1 \leq j \leq n - i, \\ < 1, & n - i + 1 \leq j \leq n, \\
\end{cases}
\]
for \( 1 \leq i \leq n - 1 \), and
\[
||t^{(n)}_j||_1 < 1, \quad 1 \leq j \leq n,
\]
which implies that \( ||T_n||_\infty \leq \epsilon < 1 \) for some \( 0 < \epsilon < 1 \).

Now, we prove this claim by induction. For \( i = 1 \), the statement follows directly from the assumption that \( 0 < \sum_{j=1}^{n} |a_{i,j}| < 1 \). For any \( 1 \leq i \leq n - 1 \), if \( (47) \) holds, then following from the structure of \( A_i \) in \( [45] \), we obtain
\[
||t^{(i+1)}_j||_1 = \begin{cases} 1, & 1 \leq j \leq n - i - 1, \\ ||t^{(i)}_j||_1 < 1, & n - i \leq j \leq n - i - 1, \\
\end{cases}
\]
and
\[
||t^{(n)}_j||_1 \leq \sum_{k=1}^{n} |a_{i+1,k}| ||t^{(i)}_{n-k+1}||_1 \leq \sum_{k=1}^{n} |a_{i+1,k}| < 1, \quad (a)
\]
where (a) is due to \( ||t^{(i)}_{n-i+1}||_1 \leq 1 \) from \( (47) \) and \( (48) \). Thus, we have proved the first part of the lemma by induction.

For the second part, by Lemma \( [53] \) combined with the first part, we find that
\[
||A_{Nn} A_{Nn-1} \ldots A_1||_\infty \leq \prod_{k=1}^{N} ||A_{kn} A_{kn-1} \ldots A_{kn-n+1}||_\infty \leq \epsilon^N
\]
for some \( 0 < \epsilon < 1 \), as claimed.

**APPENDIX I**

**PROOF OF LEMMA 1**

According to \( [52] \) (equation (4.10))), \( \phi_{t,0} \neq 0 \) is a necessary condition for \( \Phi_t(B) \) to be invertible and the coefficients of the inverse operator \( \Theta_t(B) \) can be solved iteratively by
\[
\theta_{t, i} = \begin{cases} 1/\phi_{t,0}, & i = 0, \\ -(1/\phi_{t-i,0}) \sum_{j=1}^{i} \theta_{t,i-j} \phi_{t-i+j,j}, & 1 \leq i \leq p - 1, \\ -(1/\phi_{t-i,0}) \sum_{j=1}^{p} \theta_{t,i-j} \phi_{t-i+j,j}, & i \geq p. \end{cases}
\]

For a fixed \( t \), we take \( \{\theta_{t,i}, 0 \leq i \leq p - 1\} \) as the initial value, then \( (49) \) is a homogeneous linear difference equation, which can be represented in a multi-dimensional form
\[
x_{t,n} = \begin{cases} A_{t,n} x_{t,n-1}, & n \geq 1, \\ x_{t,0}, & n = 0, \end{cases}
\]
where \( x_{t,n} = [\theta_{t,n}, \ldots, \theta_{t,n+p-1}]^T \) and
\[
A_{t,n} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{p} -a_{t,n,p-i} & \cdots & -a_{t,n,1} \end{bmatrix}
\]
with \( a_{t,n,j} = (\phi_j(\phi_{t+n+p-j-k} - \phi_{t+n+p-j}))/\phi_j(\phi_{t+n+p-j-k}) \) for \( 1 \leq j \leq p \). By the Leibniz formula of determinant \( [50] \), we obtain
\[
||A_{t,n}|| = (-1)^{p+1} a_{t,n,p}.
\]

Given the initial value \( x_{t,0} \), the solution of equation \( (50) \) is given by
\[
x_{t,n} = A_{t,n} x_{t,n-1} \ldots A_{t,1} x_{t,0} \triangleq T_{t,n} x_{t,0},
\]
where \( x_{t,0} \) is constantly non-zero due to \( \phi_{t,0} \neq 0 \).

**Proof of Lemma 1**

(1) We start with the sufficient conditions. Recall the condition \( (9) |\phi_{t,0}| > \sum_{j=1}^{p} |\phi_{t+j,j}| > 0 \) in Lemma \( [1] \) Note that \( \phi_{t,0} \neq 0 \), for all \( t \), follows directly from this sufficient condition. We will show that if \( (9) \) holds, then an inverse operator \( \Theta_t(B) \) exists. It suffices to prove that the coefficients of \( \Theta_t(B) \) are absolutely summable. This is trivial when \( r \) is finite. The remainder of the proof is thus devoted to the case when \( r = \infty \).

First, since condition \( (9) \) implies that \( \sum_{j=1}^{p} |a_{t,n,j}| < 1 \), for matrix \( A_{t,n} \) in \( (51) \), we have \( ||A_{t,n}||_\infty \leq 1 \) for all \( n \geq 1 \). It follows that \( ||T_{t,n}||_\infty = ||A_{t,n} T_{t,n-1}||_\infty \leq ||A_{t,n}||_\infty ||T_{t,n-1}||_\infty \leq ||T_{t,n-1}||_\infty \), where the first inequality is from Lemma \( [51] \). We thus observe that \( ||T_{t,n}||_\infty \), for \( n \geq 1 \), is a non-increasing sequence in \( n \).

Note that the sequence of companion matrices \( A_{t,n} \) satisfies condition \( (46) \) in Lemma \( [2] \) thus there exists \( 0 < \epsilon < 1 \) such that \( ||T_{t,jp+k}||_\infty \leq ||T_{t,jp}||_\infty \leq \epsilon^j \), for any \( j \geq 0 \) and \( 1 \leq k \leq p \).

Now we show that the coefficients of \( \Theta_t(B) \) are absolutely summable. Note that
\[
\sum_{i=0}^{\infty} |\theta_{t,i}| \leq ||x_{t,0}||_1 + \frac{1}{p} \sum_{i=1}^{\infty} ||x_{t,i}||_1
\]
due to the additional non-negative terms. We can upper bound \( \frac{1}{p} \sum_{i=1}^{\infty} ||x_{t,i}||_1 \) as follows,
\[
\frac{1}{p} \sum_{i=1}^{\infty} ||x_{t,i}||_1 \leq \frac{||x_{t,0}||_1}{p} \sum_{i=1}^{\infty} ||T_{i,j}||_1
\]
\[
\leq \frac{||x_{t,0}||_1 C}{p} \sum_{j=0}^{\infty} ||T_{j,p+j}||_\infty
\]
\[
= \frac{||x_{t,0}||_1 C}{p} \sum_{j=0}^{\infty} ||T_{j,p+j}||_\infty
\]

\(
(52) \)

\( (53) \)
\[ \leq ||x_{t,0}||_1 C \sum_{j=0}^{\infty} \varepsilon^j < \infty, \]

where (52) and (53) use Lemma 5(1) and Lemma 5(3) respectively. Therefore we have \( \sum_{i=0}^{\infty} ||\theta_{t,i}|| < \infty \). This completes the proof of the sufficient condition (9) in Lemma [1].

Recall the second sufficient condition in (10) \( \phi_{t,0} > \phi_{t+1,1} > \ldots > \phi_{t+p,p} \geq 0 \). Note again that \( \phi_{t,0} \neq 0 \), for all \( t \), follows directly from this sufficient condition. Similar to the first sufficient condition, we focus on the case when \( r = \infty \), and show that if (10) holds, then the coefficients of \( \Theta_t^\infty(B) \) are absolutely summable. Since (10) implies that

\[ a_{t,n,0} = 1 > a_{t,n,1} > \ldots > a_{t,n,p-1} > a_{t,n,p} \geq 0, \]

it follows from Lemma 6 that there exists \( 0 < \varepsilon < 1 \) such that \( \rho(T_{t,n}) \leq \\varepsilon^n \). Then by Lemma 5(4) there exists a matrix norm \( ||\cdot||_{(n)} \) such that \( ||T_{t,n}||_{(n)} \leq \rho(T_{t,n}) + 2^{-n} \) for each \( n \geq 1 \). Hence there exists \( 0 < C_n < \infty \) and \( 0 < \varepsilon < 1 \) such that

\[ ||T_{t,n}||_1 \leq C_n ||T_{t,n}||_{(n)} \leq C_n \left( \rho(T_{t,n}) + 2^{-n} \right) \leq C_n \left( \varepsilon^n + 2^{-n} \right), \]

where the first inequality follows from Lemma 5(3). Using equation (52) again, we find that the coefficients of \( \Psi_t^\infty(B) \)

are absolute summable since

\[ \frac{||x_{t,0}||_1}{\rho} \sum_{i=1}^{\infty} ||T_{t,i}||_1 \leq \frac{||x_{t,0}||_1}{\rho} \sum_{i=1}^{\infty} C_i (\varepsilon^i + 2^{-i}) \leq \frac{(d)}{\rho} ||x_{t,0}||_1 C \sum_{i=1}^{\infty} (\varepsilon^i + 2^{-i}) < \infty, \]

where (d) is due to \( 0 < C \triangleq \max_{i \geq 1} C_i < \infty \). Putting together the pieces yields the two sufficient conditions.

(II) Now we move on to the necessary condition. Assume that \( \Theta_q^\infty(B) \) exists for finite \( q \), then \( \phi_{t,0} \neq 0 \) [32]. Recall \( x_{t,n} = [\theta_{t,n}, \ldots, \theta_{t,n+p-1}]^T \) and note that we have \( x_{t,q+1} = 0 \) due to the finiteness of \( q \). This leads to \( x_{t,q+1} = T_{t,q+1}x_{t,0} = 0 \), where the only solution of this homogeneous linear system is zero if \( T_{t,q+1} \) is nonsingular [50]. However, since \( x_{t,0} \) is non-zero, we must have that \( T_{t,q+1} \) is singular, i.e.,

\[ |T_{t,q+1}| = \prod_{i=1}^{q+1} |A_{t,i}| = 0. \]

This implies that \( \prod_{i=0}^{q} \phi_{t-i,p} = 0 \). Combined with \( \phi_{t,0} \neq 0 \), we have shown the necessary condition (11) in Lemma [1] as claimed.  

[ ]