Causal Inference from Slowly Varying Nonstationary Processes

Kang Du and Yu Xiang

Abstract—Causal inference from observational data following the restricted structural causal model (SCM) 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 [2]–[5], following the pioneering works on structural causal models (SCMs) by Pearl [6]. 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) [2], the non-linear additive noise models (ANMs) [3], [7], and the post-nonlinear causal model [4]. The structure identifiability can be proved either exactly [2] or *in generic cases* [3], [4], [7], 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 [8], the *time series models with independent noise* (TiMINo) considers 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 [9] without considering instantaneous effects, while LiNGAM-t [10] incorporates instantaneous effects and non-Gaussian noise. A bivariate *deterministic* model via a *linear time-invariant* filter is studied in [11]. The directed information

rate [12], [13] from information theory is defined for bivariate stationary processes (see [14] for its relationship with Granger causality).

There are a few works on causal inference through the lens of nonstationarity [15]–[17]. The *time-dependent causal model* [15] (referred to as TCM in this paper) and CD-NOD [17] 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 [16] study a linear model (where the coefficients follow the autoregressive models) with additive noise that are uncorrelated in time, and the estimation step relies on nonlinear state-space model estimation procedures. However, none of them is built on well-established frameworks for *slowly varying nonstationary processes* such as evolutionary spectra [18], Wigner-Ville spectral analysis [19], and locally stationary processes [20] among others, 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 causal structure identification in generic cases and can be reliably estimated leveraging the bivariate evolutionary spectra framework [21], [22].

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 [2] or nonlinearity [7] is required for identification. Third, we develop efficient estimation algorithms, leveraging a recent variant of the evolutionary spectra estimate [23], 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 these results to a network setting in Section IV and report our experimental results in Section V.

A. Notation

Let $\mathbb{Z}$, $\mathbb{Z}_{\geq 0}$, and $\mathbb{C}$ denote the integers, non-negative integers, and complex numbers, respectively. We use $\bar{\mathbb{Z}}_{\geq 0}$ 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 $\mu_{X,t} = \mathsf{E}[X_t]$ and auto-covariance function $\gamma_{XX}(r, s) = \mathrm{Cov}(X_r, X_s)$. 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 is referred to as *wide-sense stationary* processes. We use the capital Greek letter (e.g., Φ, Ψ, H) for polynomial function and the corresponding lower case (i.e., ϕ, ψ, η) 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 $\mathsf{B}^j X_t \triangleq X_{t-j}, j \in \bar{\mathbb{Z}}_{\geq 0}$. Let $\Phi_t^p(z) \triangleq \sum_{j=0}^p \phi_{t,j} z^j, z \in \mathbb{C}$, denote a time-dependent polynomial function of finite degree p. If $\Phi_t^p(z)$ is not constantly zero, we require $\phi_{t,p} \neq 0$ for some t. For infinite degree $p = \infty$, we define $\Phi_t^\infty(z) \triangleq \sum_{j=0}^\infty \phi_{t,j} z^j$, with $z \in \mathbb{C}$ such that $|z| \leq 1$, where the coefficients of $\Phi_t^\infty(z)$ are assumed to be *absolutely summable*, i.e., $\sum_{j=0}^\infty |\phi_{t,j}| < \infty$. Given a polynomial function $\Phi_t^p(z)$ of degree $p \in \bar{\mathbb{Z}}_{\geq 0}$, a *time-varying (lag-polynomial) operator* of order p is defined as

$$\Phi_t^p(\mathsf{B}) \triangleq \sum_{j=0}^p \phi_{t,j} \mathsf{B}^j. \tag{1}$$

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

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

$$Y_t = \Phi_t^p(\mathsf{B})X_t + N_t, \quad p \in \mathbb{Z}_{\geq 0}, \quad \{N_t\} \perp\!\!\!\perp \{X_t\}, \tag{2}$$

where the noise $\{N_t\}$ is a stationary process, and we assume that $\Phi_t^p(\mathsf{B})$ is invertible. The assumption that $\Phi_t^p(\mathsf{B})$ is invertible (see Lemma 1 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 [10]. Also note that (2) can be equivalently written as any invertible time-invariant operator

applied to its both sides (since the noise remains stationary). We will thus focus on the representation in (2) for simplicity.

Remark 1: We do not specify the generating process of $\{X_t\}$, which is in contrast to the bivariate version of SCMs in [8], [10], [15], [16] 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 IV, 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 $\Psi_t^q(\mathsf{B}), q \in \bar{\mathbb{Z}}_{\geq 0}$, and a stationary process $\{\widetilde{N}_t\}$ such that

$$X_t = \Psi_t^q(\mathsf{B})Y_t + \widetilde{N}_t, \quad \{\widetilde{N}_t\} \perp\!\!\!\perp \{Y_t\}. \tag{3}$$

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 (3). Note that a valid backward model requires the coefficients of $\Psi_t^q(\mathsf{B})$, i.e., $\{\psi_{t,i}\}$ to be absolutely summable.

B. Identifiability

Recall that for bivariate Gaussian processes, the backward model defined in (3) has to satisfy two constraints: *the independence constraint* ($\{\widetilde{N}_t\} \perp\!\!\!\perp \{Y_t\}$) and *the stationarity constraint* ($\{\widetilde{N}_t\}$ 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

$$Y_t = \Phi_t^p(\mathsf{B})X_t + N_t, \tag{4}$$

where we assume that $\Phi_t^p(\mathsf{B})$ is invertible. Then a backward model of (4) exists only if the following two conditions are satisfied.

(I) *Condition for the independence constraint.* The equation $\gamma_{XX}(t_1, t_2) = H_{t_2}^s(\mathsf{B})\alpha(t_1, t_2)$ with respect to $H_{t_2}^s(\mathsf{B})$, where

$$\begin{aligned}\alpha(t_1, t_2) = \Phi_{t_2}^p(\mathsf{B})\gamma_{XX}(t_1, t_2) \\ + (\Phi_{t_1}^p(\mathsf{B}))^{-1}\gamma_{NN}(t_2 - t_1),\end{aligned}$$

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

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

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

is *time-invariant*.

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 [3] 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 [21], [22], 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 $\{X_t\}$ and $\{N_t\}$ 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 $\mathsf{E}[X_t^2] = \sigma_X^2$ and $\mathsf{E}[N_t^2] = \sigma_N^2$, and $\phi_{t,p} \neq 0$ for all t, then the coefficients of the operator $\Psi_t^q(\mathsf{B})$ in (3) are determined by $\psi_{t,0} = 1/\phi_{t,0}$ and

$$\psi_{t,i} = \frac{-1}{\phi_{t-i,0}} \left(\sum_{j=1}^{\min(p,i)} \psi_{t,i-j}\phi_{t-i+j,j} + \frac{\sigma_N^2 \psi_{t,i-p}}{\sigma_X^2 \phi_{t+p-i,p}} \right) \quad (5)$$

for $i \geq 1$. A backward model (3) exists only if $\{\psi_{t,i}\}$ is absolutely summable and $\mathrm{Var}(\widetilde{N}_t) = \sum_{j=1}^{\min(s,r)} \psi_{t,j}\theta_{t-j,j}\gamma_{NN}(0)$ is time-invariant, where $\{\theta_{t,j}\}$ are the coefficients of the inverse operator of $\Phi_t^p(\mathsf{B})$ (see equation (6) below).

Remark 2: As the SNR σ_X^2/σ_N^2 goes to infinity, the coefficients $\{\psi_{t,i}\}$, converges to $\theta_{t,0} = 1/\phi_{t,0}$ and

$$\theta_{t,i} = -\frac{1}{\phi_{t-i,0}} \sum_{j=1}^{\min(p,i)} \theta_{t,i-j}\phi_{t-i+j,j}, \quad i \geq 1, \quad (6)$$

which are the coefficients of $\Theta_t^r(\mathsf{B}) \triangleq (\Phi_t^p(\mathsf{B}))^{-1}$ (see the derivation of (6) in [24, equation (4.10)]). Thus the invertibility of $\Phi_t^p(\mathsf{B})$ is a necessary condition for a backward model to exist when the SNR is sufficiently high. Since we assume that $\Phi_t^p(\mathsf{B})$ is invertible, we thus focus on the cases when the identifiability is more difficult to show.

In Corollary 1, we show that $\{\psi_{t,i}\}$ can be solved iteratively, and the variance of $\{\widetilde{N}_t\}$ is written as a combination of $\{\psi_{t,i}\}$ and $\{\theta_{t,i}\}$. In general, it could be hard to check whether $\{\psi_{t,i}\}$ is absolutely summable and whether $\{\widetilde{N}_t\}$ is stationary. To get a concrete sense of the identifiability result, we simplify the setting by letting $\Phi_t^p(\mathsf{B})$ 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 σ_X^2 and σ_N^2, respectively. Consider the following forward model with $\phi(t) \neq 0$,

$$Y_t = \phi(t)X_t + N_t, \quad \{N_t\} \perp\!\!\!\perp \{X_t\}. \quad (7)$$

Then there exists a model as follows,

$$X_t = \frac{\phi(t)}{\phi^2(t) + \sigma_N^2/\sigma_X^2}Y_t + \widetilde{N}_t, \quad \{\widetilde{N}_t\} \perp\!\!\!\perp \{Y_t\}, \quad (8)$$

where $\{\widetilde{N}_t\}$ is determined by $\widetilde{N}_t = \frac{\sigma_N}{\sqrt{\phi^2(t)+\sigma_N^2/\sigma_X^2}}W_t$, where $\{W_t\}$ is an i.i.d. process with $\sigma_W^2 = 1$.

Remark 3: Due to the stationarity constraint on $\{\widetilde{N}_t\}$, a backward model exists only if $|\phi(t)|$ is time-invariant. The noise $\{\widetilde{N}_t\}$ has the form of a stationary process multiplied by a nonnegative function, which belongs to a class of nonstationary processes call the uniformly modulated process (UMP) [18] (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

$$\widetilde{N}_t = (1 - \Psi_t^q(\mathsf{B})\Phi_t^p(\mathsf{B}))X_t - \Psi_t^q(\mathsf{B})N_t,$$

which is a sum of two independent processes. In generic cases, $\{\widetilde{N}_t\}$ 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.

C. 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 $\Phi_t^p(\mathsf{B})$ 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 $\Phi_t^p(\mathsf{B})$ to $\{X_t\}$, we obtain

$$\Phi_t^p(\mathsf{B})X_t = \sum_{j=0}^{p} \phi_{t,j}X_{t-j}. \quad (9)$$

Since we focus on Gaussian processes and operators with absolutely summable coefficients, we would like to have any series of the form in (9) to converge even when $p = \infty$. To 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 [25] (and we include it in Appendix D for completeness).

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

$$\Psi_t^\infty(\mathsf{B})X_t = \sum_{j=0}^{\infty} \psi_{t,j}X_{t-j} \quad (10)$$

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

We now discuss the relationship between different operators. First, we say two operators $\Phi_t^p(\mathsf{B})$ and $\Psi_t^q(\mathsf{B})$, with $p, q \in \bar{\mathbb{Z}}_{\geq 0}$, are equivalent if $\phi_{t,j} = \psi_{t,j}$ holds for all $j \geq 0$ and $t \in \mathbb{Z}$, and we write $\Phi_t^p(\mathsf{B}) = \Psi_t^q(\mathsf{B})$. Otherwise, we use $\neq$ 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 $\Phi_t^p(\mathsf{B}) = \Psi_t^q(\mathsf{B})$ if

$$\Phi_t^p(z) = \Psi_t^q(z) \quad (11)$$

holds for z in some open set $\mathbb{E} \subseteq C$ that contains 0 (see Appendix E 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.,

$$\Phi_t^p(\mathsf{B})\Psi_t^q(\mathsf{B}) \neq \Psi_t^q(\mathsf{B})\Phi_t^p(\mathsf{B}). \tag{12}$$

Remark 4: For two operators $\Phi_{t_1}^p(\mathsf{B})$ and $\Psi_{t_2}^q(\mathsf{B})$ such that t_1 and t_2 do not depend on each other, we have $\Phi_{t_1}^p(\mathsf{B})\Psi_{t_2}^q(\mathsf{B}) = \Psi_{t_1}^q(\mathsf{B})\Phi_{t_2}^p(\mathsf{B})$.

As a consequence of the non-commutative property in (12), $\Phi_t^p(\mathsf{B})$ may not be the inverse operator of $(\Phi_t^p(\mathsf{B}))^{-1}$ in general. It is known that, when $\Phi_t^p(\mathsf{B})$ is time-invariant, a necessary and sufficient condition for $\Phi^p(\mathsf{B})$ to be invertible is

$$\sum_{j=0}^{p} \phi_j z^j \neq 0, \tag{13}$$

for $|z| \leq 1$, which says that the roots of the polynomial in (13) are strictly outside the unit circle. A similar statement was proved in [25]. When $\Phi_t^p(\mathsf{B})$ is time-varying and $\phi_{t,p} \neq 0$ for all t, a necessary and sufficient condition for the invertibility of $\Phi_t^p(\mathsf{B})$ is provided in [26] 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 $\Phi_t^p(\mathsf{B})$ 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 $\Phi_t^p(\mathsf{B})$ to be invertible.*
The inverse operator $(\Phi_t^p(\mathsf{B}))^{-1}$ exists if either of the following conditions holds,

$$\text{(a)} \quad |\phi_{t,0}| > \sum_{j=1}^{p} |\phi_{t+j,j}| > 0, \tag{14}$$

$$\text{(b)} \quad \phi_{t,0} > \phi_{t+1,1} > \ldots > \phi_{t+p,p} \geq 0. \tag{15}$$

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

$$\prod_{i=0}^{q} \phi_{t-i,p} = 0. \tag{16}$$

Remark 5: If $\Phi_t^p(\mathsf{B})$ is time-invariant, then (16) reduces to $\phi_{t,p} = 0$ for all t, which contradicts the definition of $\Phi_t^p(\mathsf{B})$ since it requires that $\phi_{t,p} \neq 0$ for some t. Thus the inverse operator of $\Phi_t^p(\mathsf{B})$ cannot be of finite order in the time-invariant case.

The proof of Lemma 1 is provided in Appendix G. In Section III-D, we show that there is a close relationship between condition (14) 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 $\Phi_t^1(\mathsf{B}) \triangleq 1 + \phi_{t,1}\mathsf{B}$, where $\phi_{t,1} = 1$ when t is even, and $\phi_{t,1} = 0$ otherwise. One can check that condition (16) holds for all $q \geq 1$. Then using (6), it is straightforward to find that

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

Example 2. Consider the first-order operator $\Psi_t^1(\mathsf{B}) \triangleq 1 + \psi_{t,1}\mathsf{B}$, with $0 < |\psi_{t,1}| < 1$. One can check that condition (16) does not hold for any $q \geq 1$ while condition (14) holds immediately. Then by (6), we obtain

$$(1 + \psi_{t,1}\mathsf{B})^{-1} = 1 + \sum_{j=1}^{\infty} \left((-1)^j \prod_{k=1}^{j} \psi_{t-k+1,1} \right) \mathsf{B}^j. \tag{17}$$

Example 3. Consider the operator $\Psi_t^1(\mathsf{B})$ 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 (16) does not hold for any $q \geq 1$. It then follows from (14) in Lemma 1 that $\Psi_t^1(\mathsf{B})$ has an inverse operator of the form (17). This operator was employed in [21] as the transfer function for an open loop system.

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 $x \to y$ if the following model holds,

$$Y_t = \sum_{u=0}^{\infty} d_t(u)X_{t-u} + N_t, \quad \{X_t\} \perp\!\!\!\perp \{N_t\}, \tag{18}$$

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, $X_t = \sum_{u=0}^{\infty} \tilde{d}_t(u)Y_{t-u} + \tilde{N}_t, \{Y_t\} \perp\!\!\!\perp \{\tilde{N}_t\}$, where $\{\tilde{N}_t\}$ is a stationary process, then we say the causal direction is $y \to x$. The assumptions on model (18) 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 $H_0 : x \to y$, and the test for $y \to x$ can be done in the same manner. Let $p_I^{x \to y}$ denotes the p-value from the independence test and $q_S^{x \to y} = 1$ if the residual is stationary and 0 if nonstationary. Similarly, we obtain $p_I^{y \to x}$ and $q_S^{y \to x}$ from the test for $y \to x$. We accept or reject H_0 by checking the following conditions. For a prefixed α, we accept H_0 if $p_I^{x \to y} \geq \alpha$ and $p_I^{y \to x} < \alpha$. (Similarly, we reject H_0 if $p_I^{x \to y} < \alpha$ and $p_I^{y \to x} \geq \alpha$.) If $p_I^{x \to y} \geq \alpha$ and $p_I^{y \to x} \geq \alpha$, then we rely on the stationarity test: We accept H_0 if $q_S^{x \to y} = 1$ and $q_S^{y \to x} = 0$ (or reject H_0 if $q_S^{x \to y} = 0$ and $q_S^{y \to x} = 1$). 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 (18) considered by our estimation procedure is $\lfloor N_F/2 \rfloor$ (see Section V for details). The order p can be selected using AIC [27] or BIC [28]. For a similar independence test task, previous works [8], [15] have used a kernel-based independence test developed for i.i.d. data [29] (referred to as HSIC), which may suffer from high false positive rates in certain cases [30]. The estimation of the filter and the stationarity test are based on the evolutionary spectra framework [31] by incorporating the multitaper method as in the univariate case [23], [32] (see more details below).

B. Univariate nonstationary processes

To set the stage, we start with a brief review of the evolutionary spectra framework [18]. Consider a class of nonstationary processes $\{X_t\}$, with $\mathsf{E}[X_t] = 0$ and $\mathsf{E}[|X_t|^2] < \infty$ for $t \in \mathbb{Z}$, such that

$$X_t = \int_{-\pi}^{\pi} \phi_t(w) dZ(w), t \in \mathbb{Z}, \tag{19}$$

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 $\mathsf{E}[|dZ(w)|^2] = d\mu(w)$. 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 (19) 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 [18] 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 [33], is characterized [23]. Interesting methodologies on neural processes can be found in [34].

It is hard to characterize *characteristic widths* [18], which quantifies the length of a "stable" segment, exactly for semi-stationary processes [35]. 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)* [18], is of the following form:

$$X_t = c(t)Y_t, \tag{20}$$

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 $\mathsf{E}|dZ(w)|^2 = dF_Y(w)$. The process introduced in (20) is an oscillatory process since $\mathcal{F}_Y = \{c(t)e^{iwt}\}$ 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$(x \to y)$
    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$(y \to x)$

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 $\mathsf{E}|dZ_x(w)|^2 = d\mu_x(w)$ and $\{Z_y(w)\}$ with $\mathsf{E}|dZ_y(w)|^2 = d\mu_y(w)$ are two orthogonal increment processes, and $\mathsf{E}[dZ_x(w)dZ_y^*(w)] \triangleq d\mu_{xy}(w)$. The evolutionary cross-spectrum [21] 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

$$f_{t,xy}(w) = A_{t,x}(w) A_{t,y}^*(w) \frac{d\mu_{xy}(w)}{dw}.$$

For $\{Y_t\} \triangleq \{X_t\}$, the cross-spectral density of $\{X_t, Y_t\}$ reduces to the auto-spectral density of $\{X_t\}$. Note that $f_{t,xy}(w)$ is in general a complex function. In this work, we adopt the multitaper method approach [23] for the estimation of the auto-spectral densities $f_{t,xx}(w)$ and $f_{t,yy}(w)$ and the cross-spectral density $f_{t,xy}(w)$.

D. Estimation of the filter

Following the model assumptions in [21], we assume $\{X_t\}$ and $\{Y_t\}$ are semi-stationary processes. The filter $d_t(u)$ is assumed to satisfy the *slowing-varying condition* [22] described as follows. Let $D_t(w)$ denote the Fourier transform of $d_t(u)$ with respect to u. For each w, consider $D_t(w)$ 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 θ both with period 2π. It suffices to define the slowly varying condition in the period $-\pi < w, \theta \leq \pi$. By leveraging the powerful estimation procedure in [21], we estimate $d_t(u)$ by

$$\hat{d}_t(u) = \mathscr{F}_w^{-1}\{\hat{D}_t(w)\} = \mathscr{F}_w^{-1}\left\{\frac{\hat{f}_{t,yx}(w)}{\hat{f}_{t,xx}(w)}\right\},$$

where $\mathscr{F}_w^{-1}$ denotes the inverse Fourier transform.

It is thus 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 4. 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[1]. 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 (14) is satisfied. The functions $D_t(w)$ and $L_w(\theta)$ in the period $-\pi < w, \theta \leq \pi$ are given as follows,

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

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

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

Example 5. Consider the filter $d_t(u)$ with $d_t(0) = 1$ and $\sum_{-\infty}^{\infty} |d_t(u)| < \infty$ for $1 \leq u \leq p$, for which we can assume that condition (14) or condition (15) holds. Then, we have

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

for $-\pi < w, \theta \leq \pi$, where

$$\left|\sum_{u=1}^{p} \sum_{t=-\infty}^{\infty} d_t(u) e^{-iuw} e^{-it\theta}\right| \leq \sum_{u=1}^{p} \sum_{t=-\infty}^{\infty} |d_t(u)| < \infty.$$

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 replace 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

---

[1]The unit impulse function is referred to as either the Dirac delta function $\delta(t)$ (with $\delta(0) = \infty$) for $t \in \mathbb{R}$ or the unit sample function $\delta(n)$ (with $\delta(0) = 1$) for $n \in \mathbb{Z}$.

$a(t)d_t(u) = |A(0)|d_t(u) + (a(t) - |A(0)|)d_t(u)$, where the first term is dominating, the invertibility condition could hold for $a(t)d_t(u)$.

E. Stationarity test

In order to verify the stationarity of the residual processes $\hat{N}_t \triangleq Y_t - \sum_{u=0}^{p} \hat{d}_t(u) X_{t-u}$ as in Algorithm 1, we make use of an improved version of the original Priestley and Subba Rao stationary test (PSR test) [36] by incorporating the multitaper method [23], [32] to obtain $\{\hat{f}_{t_i}^K(w_j), 1 \leq i \leq I, 1 \leq j \leq J\}$ 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 $\hat{\bar{f}}_t^K(w)$ denote the multitaper estimate with K tapers and N as the length of the sample records. For $i \in \{1, \ldots, I\}$, with $I = \lfloor T/N \rfloor$, and $j \in \{1, \ldots, J\}$, with $J = \lfloor (N+1)/(K+1) \rfloor$, let $W_{ij} \triangleq \log \hat{f}_{t_i}^K(w_j) + \psi(k) + \log K$, where $\psi(\cdot)$ is the digamma function. The stationary test is based on applying the two-way analysis of variance (ANOVA) test to $\{W_{ij}\}$. Let $W_{..} = (1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J} W_{ij}, W_{i\cdot} = (1/J)\sum_{j=1}^{J} W_{ij}$, and $W_{\cdot j} = (1/I)\sum_{i=1}^{I} W_{ij}$. The following statistics are needed: between time variance $S_T = J\sum_{i=1}^{I}(W_{i\cdot} - W_{..})^2$; between frequencies variance $S_F = I\sum_{j=1}^{J}(W_{\cdot j} - W_{..})^2$; interaction and residual variance $S_{I+R} = \sum_{i=1}^{I}\sum_{j=1}^{J}(W_{ij} - W_{i\cdot} - W_{\cdot j} + W_{..})^2$. The algorithm is described in Algorithm 2, where testing $S_{I+R}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ is essentially a UMP test (see details from [36]).

Algorithm 2 PSR stationarity test [36]

Input: $\{\hat{f}_{t_i}^K(w_j), 1 \leq i \leq I, 1 \leq j \leq J\}$
Output: accept or reject H_0
Compute $\{W_{ij}, 1 \leq i \leq I, 1 \leq j \leq J\}$
Test $S_{I+R}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$
if significant **then**
  reject H_0
else
  Test $S_T/\sigma^2 \sim \chi^2_{(I-1)}$
  if significant **then**
    reject H_0
  else
    accept H_0

IV. EXTENSION TO THE NETWORK SETTING

Our bivariate model shows how nonstationarity can be used for identifying the causal relation of a pair of processes. A more general setting is to identify the causal relations of a set of processes that corresponds to a DAG. In this section, we continue to exploit nonstationarity for identifying the underlying DAG of a set of Gaussian processes.

For a DAG $\mathcal{G}$ with nodes $\boldsymbol{V} = \{1, \ldots, N\}$, we use $\boldsymbol{PA}(j)$ and $\boldsymbol{ND}(j)$ to denote the set of parents and set of non-descendants of a node $j \in \boldsymbol{V}$, respectively. The set of non-descendants $\boldsymbol{ND}(j)$ is the set of all nodes in $\boldsymbol{V}$ such that there is no path from j to any $k \in \boldsymbol{ND}(j)$.

A. Model identifiability

Consider a N dimensional Gaussian process $\{\boldsymbol{X}_t\} = \{\{X_t^1\}, \ldots, \{X_t^N\}\}$ generated by the following SCM corresponding to a DAG $\mathcal{G}$ with nodes $\boldsymbol{V} = \{1, \ldots, N\}$,

$$X_t^j = \sum_{k \in \boldsymbol{PA}(j)} \Phi_t^{k \to j}(\mathsf{B}) X_t^k + N_t^j, \quad j \in \boldsymbol{V}, \quad (21)$$

where the noise processes $\{N_t^j\}, j \in \boldsymbol{V}$, are jointly independent and stationary. The maximal order of the operators $\Phi_t^{k \to j}(\mathsf{B})$'s is p.

Remark 6: Note that when $N = 2$, model (21) reduces to a bivariate model with a stationary process as the cause. While in our bivariate model (2), the generating process of the cause $\{X_t\}$ is not specified. Thus model (21) is the extension of a special case of our bivariate model (2) to the network setting.

As a consequence of the time-varying operators in (21), $\{\boldsymbol{X}_t\}$ is a set of stationary/nonstationary processes. Then a natural question is: What kind of nonstationarity is needed for identifying the DAG $\mathcal{G}$? First, let us start with the following example to show that the time-varying operators may not lead to nonstationarity.

Example 6. Let $Y_t = \Phi_t^2(\mathsf{B}) X_t$, where $\{X_t\}$ is i.i.d. and $\Phi_t^2(\mathsf{B})$ is defined by $\phi_{t,0} = (-1)^t$ and $\phi_{t,1} = (-1)^{t-1}$. Then $\{Y_t\}$ is stationary since $\mathsf{E}[Y_t] = (-1)^t \mu_X + (-1)^{t-1} \mu_X = 0$, $\mathrm{Cov}(Y_t, Y_s) = 2\sigma_X^2$ for $t = s$, $\mathrm{Cov}(Y_t, Y_s) = \sigma_X^2$ for $|t-s| = 1$, and $\mathrm{Cov}(Y_t, Y_s) = 0$ for $|t - s| \geq 2$.

This example can be easily generalized by applying any time-invariant operator to Y_t. Thus, simply using time-varying operators in (21) may not lead to nonstationarity in some non-generic cases. In order for the complete graph $\mathcal{G}$ to be identifiable, we need the following assumption.

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

In general, Assumption 1 is satisfied when all operators in (21) are time-varying and chosen generically. In the i.i.d. setting, the restricted ANMs [7] assume that the model $X_j = f_j(X_{\boldsymbol{PA}(j)}) + N_j$ belongs to a bivariate identifiable class if one conditions on $X_{\boldsymbol{PA}(j) \setminus k} = \boldsymbol{x}$ for each $k \in \boldsymbol{PA}(j)$. In particular, for a model with Gaussian noise, the function f_j needs to remain nonlinear when $X_{\boldsymbol{PA}(j) \setminus k} = \boldsymbol{x}$ are conditioned on, which is similar to how we require the nonstationarity to exist when $\{X_t^k = \boldsymbol{x}_{\boldsymbol{S}}\}, k \in \boldsymbol{S}$, are conditioned on.

Let $\boldsymbol{R} \subseteq \boldsymbol{V}$ denote the set of root nodes in $\mathcal{G}$ (i.e., all nodes j's such that $\boldsymbol{PA}(j) = \varnothing$). The identifiability of the graph $\mathcal{G}$ is built on the following lemma and the definition of *causal ordering*.

Lemma 2: A process $\{X_t^j\}, j \in \boldsymbol{V}$, is stationary if and only if $j \in \boldsymbol{R}$.

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

Definition 1 (Causal ordering): A *causal ordering* of the nodes $\boldsymbol{V}$ of a DAG $\mathcal{G}$ is an ordering of $\boldsymbol{V}$ such that there is no path from a later node to any earlier node.

From the definition of causal ordering, the parents of each node in $\boldsymbol{V}$ are contained in the previous nodes, which motivates the proof the following theorem.

Theorem 2: The graph $\mathcal{G}$ entailed in (21) is identifiable.

Proof: First, we classify the nodes $\boldsymbol{V}$ to K classes $\{\boldsymbol{V}^1, \ldots, \boldsymbol{V}^K\}$ as follows. Since the set of root nodes $\boldsymbol{R}$ is identifiable by Lemma 2, let $\boldsymbol{V}^1 = \boldsymbol{R}$. For $k \geq 2$, by conditioning on the processes $\{X_t^j\}, j \in \boldsymbol{V}^{i-1}, i \leq k$, to be zeros, we define $\boldsymbol{V}^k$ as the nodes in $\boldsymbol{V} \setminus \cup_{i=1}^{k-1} \boldsymbol{V}^i$ such that the corresponding processes are stationary. The iteration stops if $\cup_{i=1}^k \boldsymbol{V}^i = \boldsymbol{V}$. The iteration will stop within $K \leq N$ steps due to the existence of a (unknown) causal ordering.

By Assumption 1, the conditioning step implies that $\boldsymbol{PA}(j) \subseteq \cup_{i=1}^{k-1} \boldsymbol{V}^i$ for each $j \in \boldsymbol{V}^k$ and $2 \leq k \leq K$, which means that the parents of each node are in previous classes. Again, Assumption 1 implies that $\boldsymbol{PA}(j)$ of $j \in \boldsymbol{V}^k$ is the smallest set $\boldsymbol{S} \subseteq \cup_{i=1}^{k-1} \boldsymbol{V}^i$ such that $\{X_t^j\}$ is stationary when the processes that correspond to $\boldsymbol{S}$ are conditioned on to be zeros. Since the parents of each node $j \in \boldsymbol{V}$ are identified, the graph $\mathcal{G}$ is identifiable. ∎

B. Model estimation

In Section III-D, we described an estimation procedure of the time-varying filter for the bivariate model (18), while the estimation of time-varying filters for general multivariate models remains an open problem. Our causal inference procedure for the network setting is motivated by the following observation. By replacing each X_t^k in (21) with the corresponding structural equation iteratively, we obtain an equivalent representation of model (21),

$$X_t^j = \sum_{k \in \boldsymbol{AN}(j)} \Psi_t^{k \to j}(\mathsf{B}) N_t^k + N_t^j, \quad j \in \boldsymbol{V}, \quad (22)$$

where $\boldsymbol{AN}(j)$ denotes the set of ancestors of the node j (i.e., all nodes k's such that there exists a path from k to j) and each operator $\Psi_t^{k \to j}(\mathsf{B})$ is given by

$$\Psi_t^{k \to j}(\mathsf{B}) = \sum_{(k, v_1, v_2, \ldots, j)} \Phi_t^{k \to v_1}(\mathsf{B}) \Phi_t^{v_1 \to v_2}(\mathsf{B}) \ldots \Phi_t^{v_d \to j}(\mathsf{B}),$$

where $(k, v_1, v_2, \ldots, j)$ denotes any path of any length $d + 1$ from k to j. Note that the operator $\Psi_t^{k \to j}(\mathsf{B})$ in (22) and the operator $\Phi_t^{k \to j}(\mathsf{B})$ in (21) are equal for each j and $k \in \boldsymbol{PA}(j)$. In our algorithm for the network setting, we use $d_t^{k \to j}(u)$ to denote $\Psi_t^{k \to j}(\mathsf{B})$ 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 (22) 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 will see later that our algorithm naturally provides estimates of the residuals.

Based on model (22), 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, $\boldsymbol{PA}(j)$ is

the smallest set $Q \subseteq \boldsymbol{AN}(j)$ such that $\{X_t^j\}$ is stationary when $\{X_t^k\}, k \in Q$, are conditioned on to be zeros. But such conditioning is hard to evaluate in practice. To introduce our procedure for identifying $\boldsymbol{PA}(j)$ (i.e., Procedure 2), we need the following assumption, which is again generally satisfied, based on which we show the correctness of Procedure 2 in the proposition below.

Assumption 2: For any $Q \subseteq \boldsymbol{AN}(j)$ such that $Q \neq \boldsymbol{PA}(j)$, the equation

$$W_t^j = X_t^j - \sum_{k \in Q} \Psi_t^{k \to j}(\mathsf{B})X_t^k \tag{23}$$

determines a nonstationary process $\{W_t^j\}$.

Proposition 2: For any $Q \subseteq \boldsymbol{AN}(j)$, the process $\{W_t^j\}$ determined by (23) is stationary if and only if $Q = \boldsymbol{PA}(j)$

It is straightforward to see that we obtain $W_t^j = N_t^j$ when $Q = \boldsymbol{PA}(j)$ in (23), using $\Psi_t^{k \to j}(\mathsf{B}) = \Phi_t^{k \to j}(\mathsf{B})$ and (21). 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 $\{x_t^i\}$, $i \in \boldsymbol{V}$, $t = 1, \ldots, T$
Output: adjacency matrix A, estimated residuals $\{\hat{N}_t^i\}$
Initialization: $S = \boldsymbol{V}$, $A = \boldsymbol{0}_{N \times N}$, $\{\hat{N}_t^i\} = \{x_t^i\}$
while $S \neq \varnothing$ **do**
  $i^* = \text{MinStationary}(\{\hat{N}_t^i, i \in S\})$
  if $i^* = \varnothing$ **then** break
  $S \leftarrow S \setminus i^*$
  for $j \in S$ **do**
    Initialization: $c_t^j = \boldsymbol{0}_{T \times 1}$, $\boldsymbol{AN}_j = \varnothing$
    for $k \in S^c$ **do**
      Estimate the filter $d_t^{k \to j}(u)$
      if $\{x_t^j\} \not\perp\!\!\!\perp \{\hat{N}_t^k\}$ **then**
        $\boldsymbol{AN}_j \leftarrow \boldsymbol{AN}_j \cup k$
        $c_t^j \leftarrow c_t^j + \sum_u \hat{d}_t^{k \to j}(u)\hat{N}_{t-u}^k$
    $n_t^j = x_t^j - c_t^j$
  $j^* = \text{MinStationary}(\{n_t^j, j \in S\})$
  if $j^* \neq \varnothing$ **then**
    $\{\hat{N}_t^{j^*}\} = \{n_t^{j^*}\}$
    $\boldsymbol{PA}_{j^*} = \text{SelectParents}(\boldsymbol{AN}_{j^*}, \{x_t^i\}, \{\hat{d}_t^{k \to j^*}\})$
    $A(i, j^*) = 1$, $\forall i \in \boldsymbol{PA}_{j^*}$

Our algorithm follows the main idea of Theorem 2. In each iteration of the while loop, the task is to 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 for the estimation of the filters. There are three places in the algorithm where we need to select the time series that minimizes some 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 [30], which could be computationally demanding. An efficient approximation is to test whether $(1/T)|\mu| = |\sum_t d_t^{k \to j}(u)| < a$ for $u = 1, \ldots, q$. In practice, when our model assumptions are violated, one can test the joint independence of the estimated residuals $\{\hat{N}_t^i\}, i \in \boldsymbol{V}$, at the end of the algorithm. In Section V, this step is omitted since our algorithm is applied to the data generated by model (21).

Procedure 1 MinStationary

Input: N time series $\{x_t^j\}$, $j \in J$
Output: $j^* \in J$
$U = \{j \in J : \{x_t^j\} \text{ is a UMP}\}$
if $U \neq \varnothing$ **then**
  $j^* = \arg\min_{j \in U} S_T(\{x_t^j\})$
else $j^* = \varnothing$

Procedure 2 SelectParents

Input: $\boldsymbol{AN}_j, \{x_t^i\}, \{\hat{d}_t^{k \to j}(u)\}$
Output: $\boldsymbol{PA}_j$
for each $Q \subseteq \boldsymbol{AN}_j$ **do**
  $\hat{W}_t(Q) = x_t^j - \sum_{m \in Q} \sum_u \hat{d}_t^{m \to j}(u)x_{t-u}^k$
$\boldsymbol{PA}_j = \text{MinStationary}(\{\hat{W}_t(Q)\})$

Remark 7: From our experiments on synthetic data (i.e., Experiment 5 in Section V), the selected parents $\boldsymbol{PA}_j$ in Procedure 2 may be empty in certain cases, due to the estimation procedure. In such cases, one could replace $\boldsymbol{PA}_j$ with $\boldsymbol{AN}_j$ in Procedure 2, resulting in additional edges in the inferred graph (which is a subgraph of the transitive closure of $\mathcal{G}$ [37]). 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 [30]. The significance level is denoted by α 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 $\lfloor N_F/2 \rfloor$, 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 [8], TCM [15], LiNGAM-t [10] and Granger causality [9].

Fig. 1: Synthetic data A1. The red lines denote $x \to y$, and the blue lines denote $y \to x$. For the clarity of the figure, we present the results for $N_F = 64, 128$, and 256.

A. Synthetic Data

Experiment 1: First-order Models. We consider first-order models from [21],

$$Y_t = X_t + a(t)X_{t-1} + N_t, \quad 0 \le t \le T-1, \quad (24)$$

where $\{X_t\}$ is a UMP defined by $X_t = b(t)Z_t$, with $b(t)$ being a Gaussian kernel $\mathcal{N}(\mu_b, \sigma_b^2)$. We choose $\mu_b = 0.5T$ and $\sigma_b = 0.2T$, with the same ratios to T as in [21]. The process $\{Z_t\}$ is defined by a second-order AR model, $Z_t = 0.8Z_{t-1} - 0.4Z_{t-2} + \varepsilon_t$, in which $\{\varepsilon_t\}$ is a white Gaussian noise with $\varepsilon_t \sim \mathcal{N}(0, 100^2)$. The stationary noise process[2] $\{N_t\}$ is defined by $N_t = 0.8N_{t-1} - 0.16N_{t-2} + e_t$, where $\{e_t\}$ is a white Gaussian noise with $e_t \sim \mathcal{N}(0, \sigma_N^2)$.

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 ($L \ge 100$) 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 σ_N to control the SNR level. Let $a(t) = 0.5\cos(t/200)$. For each σ_N in $\{5, 10, 15, 20, 25, 40, 55, 70, 85, 100\}$, we test 100 models. For $\sigma_N = 25$, one can tell from Fig. 2a that the residuals of $x \to y$ is more likely to be nonstationary than the residuals of $y \to x$. 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 σ_N is larger). Note that even for a fixed σ_N, the SNR changes over time with $\sigma_X(t)$ (Fig. 2a). So 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.

---

[2]The only difference between the model in (24) and that in [21] is that the latter considers $\{N_t\}$ to be a UMP process.

(a)   (b)

Fig. 2: Synthetic data A2. (a-1) The standard deviation of $\{X_t\}$. (a-2) The estimated residuals of $x \to y$ when $\sigma_N = 25$. (a-3) The estimated residuals of $y \to x$ when $\sigma_N = 25$. (b) Estimated $a(t)$'s for different σ_N. The red lines denote the ground truth, and the blue lines denote the estimate.

Experiment 2: High-order Models. In this experiment, we focus on high-order models with smooth filters. We generalize model (24) to incorporate higher orders p which is generate from $\text{Unif}\{1, \ldots, 5\}$. Specifically, we have

$$Y_t = c(t)\sum_{k=0}^{p} a_k(t)X_{t-k} + N_t, \quad 0 \le t \le T-1, \quad (25)$$

where $c(t) = 1 + A\cos(t/L)$, $A \sim \text{Unif}[0.05, 0.2]$, $L \sim \text{Unif}[400, 800]$, $a_0(t) = 1$, $a_k(t) = g_k(f(P(t/T)) + S(t))$, $k \ge 1$, in which $S(t) = \text{sinc}((t - a)/b)$, $\text{sinc}(t) \triangleq \sin(\pi t)/(\pi t)$, $a \sim \text{Unif}[300, 1500]$, $b \sim \text{Unif}[400, 800]$, and $P(t)$ is a polynomial function of degree $q \sim \text{Unif}\{1, \ldots, 6\}$ with roots sampled from $\text{Unif}[-1, 1]$. The functions f and g_k's are scaling functions defined as $f(P(t)) \triangleq P(t)/\max_t(|P(t)|)$ and $g_k(a_k(t)) \triangleq c_k a_k(t)/\max_t(|a_k(t)|)$, 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, $W_t = \Phi^2(\mathsf{B})W_t + \Theta^2(\mathsf{B})e_t$, where $\Phi^2(\mathsf{B}) = (d_1 + d_2)\mathsf{B} - d_1 d_2 \mathsf{B}^2$, $\Theta^2(\mathsf{B}) = 1 + (d_3 + d_4)\mathsf{B} + d_3 d_4 \mathsf{B}^2$, 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 [25]. 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 [8] and TCM [15] 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.

Table I: Experiment 2 results (%)

α for the stationarity test	0.01	0.05
$x \to y$	84.1	82.8
$y \to x$	0.6	0.9
both not independent	11.7	10.8
both stationary	1.0	0.4
both nonstationary	2.6	5.1

Table III: Experiment 4 results (%)

b	$x \to y$	$y \to x$
0	33	4
0.125	89	2
0.25	86	2
0.375	86	2
0.5	87	3

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 $g(t) = \exp((t - T/2)^2/(2\sigma_g^2))$, $\sigma_g \sim \text{Unif}[0.4T, 0.8T]$. $\{W_t\}$ 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) $W_t = 15V_t$, with $\{V_t\}$ being a sequence of i.i.d. variables following the student's t-distribution with degrees of freedom 5. We use of a variant of Algorithm 1 that replaces the stationarity test by a UMP test. We sample 1000 models from (25) for each case of the UMP noise. Our method works well for the three cases (see Table II). 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.

Table II: Experiment 3 results (%)

Gaussian	Ours	LiNGAM-t	Granger
$x \to y$	84.1	64.4	39.2
$y \to x$	0.7	35.6	2.5
undecided	15.2	0	58.3
uniform			
$x \to y$	84.2	66.1	29.9
$y \to x$	0.7	33.9	4.4
undecided	15.1	0	65.7
student's t			
$x \to y$	80.8	70.7	30.9
$y \to x$	0.3	29.3	3.5
undecided	18.9	0	65.6

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 (24) with $a(t) = 0.5 \text{tri}(t/200; b), 0 \le b < 1$, which is a triangle wave function. The function $\text{tri}(t; b)$ in one of its period $[0, 2\pi]$ is defined as

$$\text{tri}(t; b) = \begin{cases} 1 - \frac{t}{2\pi}, & 0 \le t \le 2\pi, b = 0 \\ \frac{t}{2\pi b}, & 0 \le t \le 2\pi b, b \ne 0 \\ \frac{t}{2\pi b - 2\pi} + \frac{1}{1-b}, & 2\pi b < t \le 2\pi, b \ne 0 \end{cases},$$

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.

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 (22), let $\{N_t^j\}$ be an i.i.d. Gaussian process with zero mean and variance $\sigma_j^2 \sim \text{Unif}[5, 10]$. Each time-varying filter is defined in the same way as the bivariate model (25), 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 3 with $\alpha = 0.01$ and $\alpha = 0.05$ for the UMP test, respectively. To approximately test the independence between $\{x_t^j\}$ and $\{\hat{N}_t^k\}$ in Algorithm 3, we test whether $|\mu| = (1/T)|\sum_t d_t^{k \to j}(u)| < a$, $a = 0.15$, 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 edges 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 percent 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 V). Our method infers more cases of wrong edges as a gets smaller, thus we suggest using $a \ge 0.1$.

Table IV: Experiment 5 results (%)

	Ours		Granger	LiNGAM-t
	0.01	0.05		
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 a (%)

a	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 [15] remains undecided.

Experiment 6: Input Gas Rate vs. Output CO₂. The Gas Furnace data set [38] has two variables: the input gas rate (x) and the output CO_2 (y). Our method with HSIC yields the correct direction with p-values $p_{x \to y} = 0.0895$ and $p_{y \to x} = 5.511 \cdot 10^{-11}$. TiMINo and Granger causality correctly infer the direction $x \to y$ [8], whereas LiNGAM-t gives the wrong direction.

Experiment 7: Duration vs. Time Interval. The Old Faithful data set [39], [40] 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 [8], 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 $p_{x \to y} = 0.0777$ and $p_{y \to x} = 9.364 \cdot 10^{-10}$. TiMINo and LiNGAM-t infer the correct direction [8]. 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 the respiratory sinus arrhythmia (RSA) [41]. This phenomenon is widely observed, especially among young and healthy individuals. We use the data from [42] to verify the causal relationship between breathing pattern and heart rate. Specifically, this data set contains two variables: chest volume (x) and heart rate (y). Since the mechanism of respiratory sinus arrhythmia is understood to be the synchronization of heart rate to breathing rhythm [44], we consider x as a cause for y. 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 B1), our method yields p-values $p_{x \to y} = 0.0610$ and $p_{y \to x} = 0.004$. 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 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. Acknowledgement

Appendix A
Proof of Theorem 1

Proof: Suppose there exists a backward model of the form

$$X_t = \Psi_t^q(\mathsf{B})Y_t + \widetilde{N}_t, \quad \{\widetilde{N}_t\} \perp\!\!\!\perp \{Y_t\}, \quad q \in \bar{\mathbb{Z}}_{\geq 0}, \tag{26}$$

where $\{\widetilde{N}_t\}$ is stationary.

(I) First, we show the condition for the independence constraint. Let $\mathcal{O}$ denote the class of operators such that

$\{Y_t\} \perp\!\!\!\perp \{X_t - H_t^s(\mathsf{B})Y_t\}$ for any $H_t^s(\mathsf{B}) \in \mathcal{O}$, $s \in \bar{\mathbb{Z}}_{\geq 0}$. Note that $\mathcal{O}$ is nonempty since $\Psi_t^q(\mathsf{B}) \in \mathcal{O}$. For any $H_t^s(\mathsf{B}) \in \mathcal{O}$, we obtain the model

$$X_t = H_t^s(\mathsf{B})Y_t + W_t, \tag{27}$$

where $\{W_t\}$ is defined by $W_t = X_t - H_t^s(\mathsf{B})Y_t$. We replace Y_t in (27) by that in (4), and have

$$W_t = (1 - H_t^s(\mathsf{B})\Phi_t^p(\mathsf{B}))X_t - H_t^s(\mathsf{B})N_t. \tag{28}$$

Now, we prove that the Gaussian process $\{H_t^s(\mathsf{B})Y_t\}$ has the same distribution for any $H_t^s(\mathsf{B}) \in \mathcal{O}$, which is equivalent to prove that $\{H_t^s(\mathsf{B})Y_t\}$ and $\{\Psi_t^q(\mathsf{B})Y_t\}$ have the same covariance function. According to the independence constraint, we know that both $\{W_t\}$ and $\{\widetilde{N}_t\}$ are independent of $\{Y_t\}$, which implies that $\{W_t - \widetilde{N}_t\} \perp\!\!\!\perp \{Y_t\}$. Then, subtracting (26) from (27) yields $W_t - \widetilde{N}_t = (\Psi_t^q(\mathsf{B}) - H_t^s(\mathsf{B}))Y_t$. It follows from Lemma 4 that the covariance function of $\{(\Psi_t^q(\mathsf{B}) - H_t^s(\mathsf{B}))Y_t\}$ is zero. Thus, we have shown that $\{H_t^s(\mathsf{B})Y_t\}$ has the same distribution for any $H_t^s(\mathsf{B}) \in \mathcal{O}$.

Moreover, we provide an explicit characterization of the operators in $\mathcal{O}$ as follows. Since two Gaussian processes are independent if and only if their cross-covariance function equals to zero, from (4) and (28), we have

$$\begin{aligned}
&\mathrm{Cov}(Y_{t_1}, W_{t_2}) \\
&= \mathrm{Cov}(\Phi_{t_1}^p(\mathsf{B})X_{t_1} + N_{t_1}, \\
&\quad (1 - H_{t_2}^s(\mathsf{B})\Phi_{t_2}^p(\mathsf{B}))X_{t_2} - H_{t_2}^s(\mathsf{B})N_{t_2}) \\
&= \Phi_{t_1}^p(\mathsf{B})(1 - H_{t_2}^s(\mathsf{B})\Phi_{t_2}^p(\mathsf{B}))\gamma_{XX}(t_1, t_2) \\
&\quad - H_{t_2}^s(\mathsf{B})\gamma_{NN}(t_2 - t_1) = 0. \quad (29)
\end{aligned}$$

Since we assume $\Phi_t^p(\mathsf{B})$ is invertible, (29) can be written as

$$\begin{aligned}
&(1 - H_{t_2}^s(\mathsf{B})\Phi_{t_2}^p(\mathsf{B}))\gamma_{XX}(t_1, t_2) \\
&\qquad = H_{t_2}^s(\mathsf{B})(\Phi_{t_1}^p(\mathsf{B}))^{-1}\gamma_{NN}(t_2 - t_1), \quad (30)
\end{aligned}$$

which can be further simplified as

$$\begin{aligned}
\gamma_{XX}(t_1, t_2) = &H_{t_2}^s(\mathsf{B})(\Phi_{t_2}^p(\mathsf{B})\gamma_{XX}(t_1, t_2) \\
&+ (\Phi_{t_1}^p(\mathsf{B}))^{-1}\gamma_{NN}(t_2 - t_1)). \quad (31)
\end{aligned}$$

Therefore we have shown that (31) determines a class of operators in $\mathcal{O}$, simply because that (31) 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 $H_t^s(\mathsf{B}) = \Psi_t^q(\mathsf{B})$, we obtain $\{W_t\} = \{\widetilde{N}_t\}$ in (27). Thus there exists $H_t^s(\mathsf{B}) \in \mathcal{O}$ such that $\{W_t\} = \{X_t - H_t^s(\mathsf{B})Y_t\}$ is stationary. The stationarity of $\{W_t\}$ implies that its covariance function is a function of $t_2 - t_1$. From (28) and (30), we have

$$\begin{aligned}
&\mathrm{Cov}(W_{t_1}, W_{t_2}) \\
&= (1 - H_{t_1}^s(\mathsf{B})\Phi_{t_1}^p(\mathsf{B}))(1 - H_{t_2}^s(\mathsf{B})\Phi_{t_2}^p(\mathsf{B}))\gamma_{XX}(t_1, t_2) \\
&\qquad\qquad\qquad + H_{t_1}^s(\mathsf{B})H_{t_2}^s(\mathsf{B})\gamma_{NN}(t_2 - t_1) \\
&= H_{t_2}^s(\mathsf{B})(\Phi_{t_1}^p(\mathsf{B}))^{-1}\gamma_{NN}(t_2 - t_1). \quad (32)
\end{aligned}$$

Let $t_1 = t_2 = t$ in (32) and $\Theta_t^r(\mathsf{B}) \triangleq (\Phi_t^p(\mathsf{B}))^{-1}$, we obtain the variance of $\{W_t\}$ as

$$\mathrm{Var}(W_t) = \sum_{j=1}^{s} \sum_{k=1}^{r} \eta_{t,j} \theta_{t-j,k} \gamma_{NN}(k - j), \tag{33}$$

which is time-invariant. ∎

Proof: Note that $\mathsf{E}[X_t X_s] = \sigma_X^2 \mathbb{1}_{t=s}$ since $\{X_t\}$ is an i.i.d. process. We solve for the coefficients of $\Psi_t^q(\mathsf{B})$ by computing $\mathsf{E}[X_t Y_{t+p-i}]$, $i \geq 0$. Using model (4), we derive

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

for $i \geq 0$. Using the backward model of (4), $\mathsf{E}[X_t Y_{t+p-i}]$ can be computed alternatively as

$$\mathsf{E}[X_t Y_{t+p-i}] = \mathsf{E}\left[Y_{t+p-i}\left(\sum_{j=0}^{q} \psi_{t,j} Y_{t-j} + \widetilde{N}_t\right)\right]$$
$$= \sum_{j=0}^{q} \psi_{t,j} \mathsf{E}[Y_{t+p-i} Y_{t-j}]$$
$$= \sum_{j=\max(0,i-2p)}^{i} \psi_{t,j} \mathsf{E}[Y_{t+p-i} Y_{t-j}],$$

where the last equality holds since $|t+p-i-(t-j)| \leq p$ is equivalent to $i - 2p \leq j \leq i$. Thus, we obtain the following relationship for $i \geq 0$,

$$\sum_{j=\max(0,i-2p)}^{i} \psi_{t,j} \mathsf{E}[Y_{t+p-i} Y_{t-j}] = \phi_{t+p-i,p-i} \sigma_X^2.$$

This can be written explicitly for $i = 0$ and $i > 1$ as follows,

$$\psi_{t,0} \mathsf{E}[Y_{t+p} Y_t] = \phi_{t+p,p} \sigma_X^2, \quad \text{for } i = 0$$

and for $i \geq 1$,

$$\psi_{t,i} \mathsf{E}[Y_{t+p-i} Y_{t-i}] = \phi_{t+p-i,p-i} \sigma_X^2$$
$$- \sum_{j=\max(0,i-2p)}^{i-1} \psi_{t,j} \mathsf{E}[Y_{t+p-i} Y_{t-j}]. \quad (34)$$

To further simplify the expression, we now show that $\mathsf{E}[Y_{t+p-i} Y_{t-i}] \neq 0$. First observe that

$$\mathsf{E}[Y_t Y_s] = \mathsf{E}\left[\left(\sum_{j=0}^{p} \phi_{t,j} X_{t-j} + N_t\right)\left(\sum_{k=0}^{p} \phi_{s,k} X_{s-k} + N_s\right)\right]$$
$$= \sigma_X^2 \sum_{j=0}^{p} \phi_{t,j} \phi_{s,s-t+j} + \sigma_N^2 \mathbb{1}_{t=s},$$

and $\mathsf{E}[Y_t Y_s] = 0$ when $|t-s| > p$. Thus we have, for $i \geq 0$,

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

which follows from the fact that $\phi_{t+p-i,p+j} = 0$ for $j \geq 1$. As a result, we can divide $\mathsf{E}[Y_{t+p-i} Y_{t-i}]$ on both sides of (34) and obtain the solution

$$\psi_{t,i} = -\frac{\sigma_N^2 \phi_{t,i-p}}{\sigma_X^2 \phi_{t+p-i,p} \phi_{t-i,0}} + \frac{\phi_{t+p-i,p-i}}{\phi_{t+p-i,p} \phi_{t-i,0}}$$

$$- \sum_{j=\max(0,i-2p)}^{i-1} \frac{\psi_{t,j} \sum_{k=0}^{p} \phi_{t+p-i,k} \phi_{t-j,i-j+p}}{\phi_{t+p-i,p} \phi_{t-i,0}} \quad (35)$$

for $i \geq 1$. Similarly, we obtain $\psi_{t,0} = 1/\phi_{t,0}$ for $i = 0$. To obtain equation (5), we will need a technical lemma (postponed to be presented in Lemma 3 below), which shows that $\Psi_t^q(\mathsf{B})$ is simply the inverse operator of $\Phi_t^p(\mathsf{B})$ if $\sigma_N^2 = 0$. Thus the last two terms in (35) can be replaced by the coefficent of the inverse operator of $\Phi_t^p(\mathsf{B})$ according to (6). Therefore, the coefficients of $\Psi_t^q(\mathsf{B})$ can be solved iteratively for all $i \geq 0$ using (5).

Now, we move on to the condition for the stationarity constraint. In Theorem 1, if $\{N_t\}$ is an i.i.d. process, then (33) can be written as $\mathrm{Var}(W_t) = \sum_{j=1}^{\min(s,r)} \eta_{t,j} \theta_{t-j,j} \gamma_{NN}(0)$, which in turn equals to $\gamma_{NN}(0) \sum_{j=1}^{\min(q,r)} \eta_{t,j} \theta_{t-j,j}$. Since the operator $H_t^s(\mathsf{B})$ in the backward model (27) is uniquely determined by (5), we have $\{W_t\} = \{\widetilde{N}_t\}$ and $\{\eta_{t,j}\} = \{\psi_{t,j}\}$. Thus, $\mathrm{Var}(\widetilde{N}_t) = \sum_{j=1}^{\min(s,r)} \psi_{t,j} \theta_{t-j,j} \gamma_{NN}(0)$. The rest follows by invoking equation (6). ∎

It remains to show the following technical lemma.

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

Proof: By replacing Y_t in the backward model (3) with the forward model (2), we obtain

$$X_t = \Psi_t^q(\mathsf{B})\Phi_t^p(\mathsf{B})X_t + \Psi_t^q(\mathsf{B})N_t + \widetilde{N}_t, \quad (36)$$

where $\{X_t\}$ and $\{N_t\}$ are i.i.d. processes. By multiplying X_t and taking expectation to both sides of (36), we obtain

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

where the last equality is obtained using the backward model. By same argument leading up to (35), we have $\psi_{t,0} \phi_{t,0} = 1$, which implies $\mathsf{E}[\widetilde{N}_t^2] = 0$. It follows that $\mathsf{E}[\widetilde{N}_t \widetilde{N}_{t-k}] = 0$ for any $k \in \mathbb{Z}$ by the Cauchy-Schwarz inequality. Let $\Theta_t^r \triangleq \Psi_t^q(\mathsf{B})\Phi_t^p(\mathsf{B})$. Similarly, by multiplying $X_{t-k}, k \geq 1$, and taking expectation to both sides of (36), we obtain

$$0 = \mathsf{E}[X_t X_{t-k}] = \theta_{t,k} \mathsf{E}[X_{t-k}^2],$$

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

Proof 1: We solve for the operator $\Psi_t^q(\mathsf{B})$ by computing $\mathsf{E}[X_t Y_{t-j}]$ in two ways using the forward model (7) and the backward model (3), respectively. Using model (7) we have $\mathsf{E}[X_t Y_{t-j}] = \phi(t)\sigma_X^2 \mathbb{1}_{j=0}$, while model (3) leads to $\mathsf{E}[X_t Y_{t-j}] = \psi_{t,j}\left(\phi^2(t-j)\sigma_X^2 + \sigma_N^2\right)$. Thus we obtain $\psi_{t,0} = \phi(t)/(\phi^2(t) + \sigma_N^2/\sigma_X^2)$ and $\psi_{t,j} = 0$ for $j \geq 1$, which gives the coefficient of Y_t in (8).

Then, by substituting (7) into (8), we obtain

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

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

APPENDIX D
PROOF OF PROPOSITION 1

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

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

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

which shows that $\sum_{j=0}^{\infty} |\psi_{t,j}||X_{t-j}|$ is finite with probability one. If $\sup_t \mathsf{E}[|X_t|^2] < \infty$ and $n > m > 0$, then

$$\mathsf{E}\left[\left|\sum_{m<j\leq n} \psi_{t,j}X_{t-j}\right|^2\right]$$

$$= \sum_{m<j\leq n}\sum_{m<k\leq n} \psi_{t,j}\bar{\psi}_{t,k}\,\mathsf{E}[X_{t-j}\bar{X}_{t-k}]$$

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

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,

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

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

which shows that the mean square limit S and $\Psi_t^{\infty}(\mathsf{B})X_t$ are equal with probability one. ∎

APPENDIX E
EQUIVALENT DEFINITIONS

To show the equivalence of the two definitions, it suffices to prove the following direction since the other direction is trivial.

Proposition 3: For two operators $\Phi_t^p(\mathsf{B})$ and $\Psi_t^q(\mathsf{B})$, with $p, q \in \bar{\mathbb{Z}}_{\geq 0}$, if

$$\Phi_t^p(z) = \Psi_t^q(z) \tag{37}$$

holds for z in some open set $\mathbb{E} \subseteq C$ that contains 0, then $\Phi_t^p(\mathsf{B}) = \Psi_t^q(\mathsf{B})$.

Proof: We prove that $\phi_{t,j} = \psi_{t,j}$ for all $j \geq 0$ by induction. Let $z = 0$ in (37), we obtain $\phi_{t,0} = \psi_{t,0}$ and $\sum_{k=1}^{\infty} \phi_{t,k}z^k = \sum_{k=1}^{\infty} \psi_{t,k}z^k$. Assume that $\phi_{t,k} = \psi_{t,k}$ for $k \leq j$. Then, for any $z \in \mathbb{E}$, we have

$$\sum_{k=j+1}^{\infty}(\phi_{t,k} - \psi_{t,k})z^k$$
$$= z^{j+1}\sum_{k=0}^{\infty}(\phi_{t,k+j+1} - \psi_{t,k+j+1})z^k = 0.$$

It follows that, we have $\sum_{k=0}^{\infty}(\phi_{t,k+j+1} - \psi_{t,k+j+1})z^k = 0$ for any $z \in \mathbb{E} \setminus 0$. Finally, taking $\lim_{z\to 0}$ on both sides of the last equality yields $\phi_{t,j+1} = \psi_{t,j+1}$. Therefore, we have proved that $\phi_{t,j} = \psi_{t,j}$ for all $j \geq 0$ as claimed. ∎

APPENDIX F
TECHNICAL LEMMAS

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

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

$$\text{Cov}(Y_{t_1}, \Phi_{t_2}^p(\mathsf{B})Y_{t_2}) = \Phi_{t_2}^p(\mathsf{B})\gamma_{YY}(t_1, t_2) = 0. \tag{38}$$

By applying $\Phi_{t_1}^p(\mathsf{B})$ to (38), we have that

$$\Phi_{t_1}^p(\mathsf{B})\Phi_{t_2}^p(\mathsf{B})\gamma_{YY}(t_1, t_2) = 0,$$

where the left-hand side is simply the covariance function of $\{\Phi_t^p(\mathsf{B})Y_t\}$. ∎

The following technical lemma contains a list of basic properties of matrix norms (see proofs in [45]).

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) $||A_1 A_2 \ldots A_k||_p \leq ||A_1||_p||A_2||_p \ldots ||A_k||_p$, 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 $||\cdot||_*$ such that $0 \leq ||A||_* - \rho(A) \leq \delta$.

The last two lemmas focus on a particular form of matrix called the *companion matrix* [46]. 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 ([46]): Let $A_i \in \mathbb{R}^{n\times n}, 1 \leq i \leq k$, be companion matrices of the form

$$A_i = \begin{bmatrix} 0 & & & \\ \vdots & & I_{n-1} & \\ 0 & & & \\ \hline -a_{i,n} & -a_{i,n-1} & \ldots & -a_{i,1} \end{bmatrix}, \tag{39}$$

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

$$\varepsilon = \max_{1\leq i\leq k, 1\leq j\leq n} \frac{a_{i,j}}{a_{i,j-1}} < 1$$

such that $\rho(A_k A_{k-1} \ldots A_1) \leq \varepsilon^k < 1$.

Finally, we establish the following lemma on a product of companion matrices inspired by [47].

Lemma 7: Let $A_i \in \mathbb{R}^{n\times n}, 1 \leq i \leq n$, be companion matrices of the form in (39). If

$$0 < \sum_{j=1}^{n} |a_{i,j}| < 1 \tag{40}$$

for $1 \le i \le n$, then there exists some $0 < \varepsilon < 1$ such that $||A_n A_{n-1} \ldots A_1||_\infty \le \varepsilon$. Moreover, for companion matrices A_i with $1 \le i \le Nn$ and $N \ge 1$, if (40) hold for $1 \le i \le Nn$, then

$$||A_{Nn}A_{Nn-1}\ldots A_1||_\infty \le \varepsilon^N$$

for some $0 < \varepsilon < 1$.

Proof: Let $T_i \triangleq A_i A_{i-1} \ldots A_1$, $1 \le i \le n$, and let $t_j^{(i)}, 1 \le j \le n$, denote the jth row of T_i. We first claim that

$$||t_j^{(i)}||_1 \begin{cases} = 1, & 1 \le j \le n-i, \\ < 1, & n-i+1 \le j \le n, \end{cases} \quad (41)$$

for $1 \le i \le n-1$, and

$$||t_j^{(n)}||_1 < 1, \quad 1 \le j \le n, \quad (42)$$

which implies that $||T_n||_\infty \le \varepsilon < 1$ for some $0 < \varepsilon < 1$.

Now, we prove this claim by induction. For $i = 1$, the statement follows directly from the assumption that $0 < \sum_{j=1}^n |a_{1,j}| < 1$. For any $1 \le i \le n-1$, if (41) holds, then following from the structure of A_i in (39), we obtain

$$||t_j^{(i+1)}||_1 = \begin{cases} 1, & 1 \le j \le n-i-1, \\ ||t_{j+1}^{(i)}||_1 < 1, & n-i \le j \le n-1, \end{cases}$$

and

$$||t_n^{(i+1)}||_1 \le \sum_{k=1}^n |a_{i+1,k}|||t_{n-k+1}^{(i)}||_1 \overset{(a)}{\le} \sum_{k=1}^n |a_{i+1,k}| < 1,$$

where (a) is due to $||t_{n-k+1}^{(i)}||_1 \le 1$ from (41) and (42). Thus, we have proved the first part of the lemma by induction.

For the second part, by Lemma 5.(3) combined with the first part, we find that

$$||A_{Nn}A_{Nn-1}\ldots A_1||_\infty$$
$$\le \prod_{k=1}^N ||A_{kn}A_{kn-1}\ldots A_{kn-n+1}||_\infty \le \varepsilon^N$$

for some $0 < \varepsilon < 1$, as claimed. ∎

APPENDIX G
PROOF OF LEMMA 1

According to [24, equation (4.10)]), $\phi_{t,0} \ne 0$ is a necessary condition for $\Phi_t^p(\mathsf{B})$ to be invertible and the coefficients of the inverse operator $\Theta_t^r(\mathsf{B})$ can be solved iteratively by

$$\theta_{t,i} = \begin{cases} 1/\phi_{t,0}, & i = 0, \\ -(1/\phi_{t-i,0}) \cdot \sum_{j=1}^i \theta_{t,i-j}\phi_{t-i+j,j}, & 1 \le i \le p-1, \\ -(1/\phi_{t-i,0}) \cdot \sum_{j=1}^p \theta_{t,i-j}\phi_{t-i+j,j}, & i \ge p. \end{cases} \quad (43)$$

For a fixed t, we take $\{\theta_{t,i}, 0 \le i \le p-1\}$ as the initial value, then (43) is a homogeneous linear difference equation, which can be represented in a multi-dimensional form

$$\boldsymbol{x}_{t,n} = \begin{cases} A_{t,n}\boldsymbol{x}_{t,n-1}, & n \ge 1, \\ \boldsymbol{x}_{t,0}, & n = 0, \end{cases} \quad (44)$$

where $\boldsymbol{x}_{t,n} = [\theta_{t,n}, \ldots, \theta_{t,n+p-1}]^T$ and

$$A_{t,n} = \begin{bmatrix} 0 & & \\ \vdots & & I_{p-1} \\ 0 & & \\ \hline -a_{t,n,p} & -a_{t,n,p-1} & \cdots & -a_{t,n,1} \end{bmatrix} \quad (45)$$

with $a_{t,n,j} \triangleq (\phi_{t-(n+p-1)+j,j})/(\phi_{t-(n+p-1),0})$ for $1 \le j \le p$. By the Leibniz formula of determinant [45], we obtain

$$|A_{t,n}| = (-1)^{p+1}a_{t,n,p}.$$

Given the initial value $\boldsymbol{x}_{t,0}$, the solution of equation (44) is given by

$$\boldsymbol{x}_{t,n} = A_{t,n}A_{t,n-1}\ldots A_{t,1}\boldsymbol{x}_{t,0} \triangleq T_{t,n}\boldsymbol{x}_{t,0},$$

where $\boldsymbol{x}_{t,0}$ is constantly non-zero due to $\phi_{t,0} \ne 0$.

Proof of Lemma 1: (I) We start with the sufficient conditions. Recall the condition (14) $|\phi_{t,0}| > \sum_{j=1}^p |\phi_{t+j,j}| > 0$ in Lemma 1. Note that $\phi_{t,0} \ne 0$, for all t, follows directly from this sufficient condition. We will show that if (14) holds, then an inverse operator $\Theta_t^r(\mathsf{B})$ exists. It suffices to prove that the coefficients of $\Theta_t^r(\mathsf{B})$ 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 (14) implies that $\sum_{j=1}^p |a_{t,n,j}| < 1$, for matrix $A_{t,n}$ in (45), we have $||A_{t,n}||_\infty \le 1$ for all $n \ge 1$. It follows that $||T_{t,n}||_\infty = ||A_{t,n}T_{t,n-1}||_\infty \le ||A_{t,n}||_\infty||T_{t,n-1}||_\infty \le ||T_{t,n-1}||_\infty$, where the first inequality is from Lemma 5.(1). We thus observe that $||T_{t,n}||_\infty$, for $n \ge 1$, is a non-increasing sequence in n.

Note that the sequence of companion matrices $A_{t,n}$ satisfies condition (40) in Lemma 7, thus there exists $0 < \varepsilon < 1$ such that $||T_{t,jp+k}||_\infty \le ||T_{t,jp}||_\infty \le \varepsilon^j$, for any $j \ge 0$ and $1 \le k \le p$.

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

$$\frac{1}{p}\sum_{i=1}^\infty ||\boldsymbol{x}_{t,i}||_1 \le \frac{||\boldsymbol{x}_{t,0}||_1}{p}\sum_{i=1}^\infty ||T_{t,i}||_1 \quad (46)$$

$$\le \frac{||\boldsymbol{x}_{t,0}||_1 C}{p}\sum_{i=1}^\infty ||T_{t,i}||_\infty \quad (47)$$

$$= \frac{||\boldsymbol{x}_{t,0}||_1 C}{p}\sum_{j=0}^\infty\sum_{k=1}^p ||T_{t,jp+k}||_\infty$$

$$\le ||\boldsymbol{x}_{t,0}||_1 C\sum_{j=0}^\infty \varepsilon^j < \infty,$$

where (46) and (47) 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 (14) in Lemma 1.

Recall the second sufficient condition in (15) $\phi_{t,0} > \phi_{t+1,1} > \ldots > \phi_{t+p,p} \ge 0$. Note again that $\phi_{t,0} \ne 0$, 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 (15) holds, then the coefficients of $\Theta_t^\infty(\mathsf{B})$ are absolutely summable. Since (15) 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

$$\begin{aligned} ||T_{t,n}||_1 &\leq C_n ||T_{t,n}||_{(n)} \\ &\leq C_n \left(\rho(T_{t,n}) + 2^{-n} \right) \leq C_n \left(\varepsilon^n + 2^{-n} \right), \end{aligned}$$

where the first inequality follows from Lemma 5.(3). Using equation (46) again, we find that the coefficients of $\Psi_t^\infty(\mathsf{B})$ are absolute summable since

$$\begin{aligned} \frac{||\boldsymbol{x}_{t,0}||_1}{p} \sum_{i=1}^\infty ||T_{t,i}||_1 &\leq \frac{||\boldsymbol{x}_{t,0}||_1}{p} \sum_{i=1}^\infty C_i(\varepsilon^i + 2^{-i}) \\ &\overset{(d)}{\leq} \frac{||\boldsymbol{x}_{t,0}||_1 C}{p} \sum_{i=1}^\infty (\varepsilon^i + 2^{-i}) < \infty, \end{aligned}$$

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 $\Theta_t^q(\mathsf{B})$ exists for finite q, then $\phi_{t,0} \neq 0$ [24]. Recall $\boldsymbol{x}_{t,n} = [\theta_{t,n}, \ldots, \theta_{t,n+p-1}]^T$ and note that we have $\boldsymbol{x}_{t,q+1} = 0$ due to the finiteness of q. This leads to $\boldsymbol{x}_{t,q+1} = T_{t,q+1}\boldsymbol{x}_{t,0} = 0$, where the only solution of this homogeneous linear system is zero if $T_{t,q+1}$ is nonsingular [45]. However, since $\boldsymbol{x}_{t,0}$ 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 (16) in Lemma 1, as claimed. ∎

REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

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

[6] J. Pearl, "Models, reasoning and inference," *Cambridge, UK: Cambridge University Press*, 2000.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] J. Massey et al., "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 1990, pp. 303–305.

[13] G. Kramer, *Directed information for channels with feedback*. Citeseer, 1998.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] M. B. Priestley, "Evolutionary spectra and non-stationary processes," *Journal of the Royal Statistical Society. B*, vol. 28, no. 1, pp. 228–240, 1966.

[19] W. Martin and P. Flandrin, "Wigner-Ville spectral analysis of nonstationary processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1461–1470, 1985.

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

[21] M. B. Priestley and H. Tong, "On the analysis of bivariate non-stationary processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 153–166, 1973.

[22] 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.

[23] 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.

[24] N. Abdrabbo 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.

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

[26] M. Hallin, "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.

[27] 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.

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

[29] 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.

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

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

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

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

[34] 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.

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

[36] 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.

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

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

[39] 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.

[40] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[41] 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.

[42] 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.

[43] Y. Ichimaru and G. Moody, "Development of the polysomnographic database on cd-rom," *Psychiatry and clinical neurosciences*, vol. 53, no. 2, pp. 175–177, 1999.

[44] 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.

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

[46] 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.

[47] 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.