Assumption-Weak Discovery of Forecasting Signals

Michael Wieck-Sosa* Michel F. C. Haddad† Aaditya Ramdas‡

September 27, 2024

Abstract

Most time series encountered in practice are not stationary, linear, or Gaussian. However, these strong assumptions are often made to simplify the problem. Motivated by the goal of discovering new forecasting signals in an "assumption-weak" framework, we develop a new test for strong Granger causality that does not require any of these unrealistic assumptions. Our framework is designed for the practical setting in which just one realization of a possibly high-dimensional nonstationary nonlinear time series is observed. Our simple testing procedure relies on *adaptive forecasting*, which makes it naturally appealing to time series researchers and forecasting practitioners alike. We will discuss different adaptive prediction techniques for nonstationary time series that can be used with widely-available, light-weight learning algorithms to construct very powerful tests. We allow the forecasting signals to be informative about the forecasting target during some periods, but uninformative during others. For situations with this *time-varying structure*, we show how our framework can be used with multiple testing procedures to discover short time-windows during which the forecasting signal is informative. We discuss how to gain power by forecasting *groups of time series*, such as energy prices in nearby geographic regions, the demand for similar consumer products, or the number of viral infections in neighboring municipalities. In our companion paper [WHR24], we use our framework to discover new epidemic forecasting signals from city-level infectivity data.

1 Introduction

In this work, we introduce an assumption-weak testing framework for conditional independence (and strong Granger causality in particular, see [FM82; SF22]). Our test only requires one realization of the time series, which can be high-dimensional, nonstationary, and nonlinear. Our test statistic can be viewed as an extension of the *generalized covariance measure* (GCM) from Shah and Peters [SP20] to the nonstationary time series setting. As we will show, moving to this complicated setting introduces many nontrivial complexities.

We show that our test has uniformly asymptotic Type-I error control provided that the adaptive predictors satisfy modest rate requirements. We utilize a bootstrap procedure based on a distribution-uniform extension of the strong Gaussian approximation from Mies and Steland [MS22]. Along the way, we develop a distribution-uniform theoretical framework for conditional covariance processes of high-dimensional nonstationary nonlinear time series, which can be of independent interest. Our framework makes very few assumptions about observed processes. Consequently, it can handle very general forms of non-stationarity and temporal dependence. Instead, our assumptions focus on the quality of the adaptive forecasts and the error processes, which will be introduced in the next section.

Recently, Shah and Peters [SP20] showed that conditional independence testing is fundamentally impossible without making further assumptions, even in the idealized iid setting. Consequently, we *must* restrict the null hypothesis in some way. This result is widely known in statistics, but it is still unfamiliar to many in the time series research community. To diffuse the awareness about this hardness result, we will clarify that it also implies that conditional independence testing between general discrete-time stochastic processes is very hard to test for. Throughout the paper, we give many examples in the context of forecasting to motivate the more technical discussions. In practice, our

*Department of Statistics and Data Science, Carnegie Mellon University. Email: mwiecksosa@cmu.edu.
†Department of Business Analytics and Applied Economics, Queen Mary University of London.
‡Department of Statistics and Data Science, Machine Learning Department, Carnegie Mellon University.

simple testing procedure leverages *adaptive forecasting* to construct powerful tests. We will discuss several adaptive forecasting techniques that can be used with widely-available, light-weight learning algorithms. Arguably, this aspect of our test makes it appealing to time series researchers and practitioners because they are already familiar the practice of forecasting.

Our contributions and paper outline. The rest of the paper is structured as follows. In Section 2, we introduce a practical CI test for time series which can exhibit very complex forms of nonstationarity. In Section H, we discuss the simulation results. In Section 3, we discuss promising avenues for future work. We leave several discussions for the Appendix. In Section A, we discuss simultaneous testing so that we can reliably discover short time windows during which conditional independence holds for particular signals and forecasting horizons. In Section D, we discuss a practical CI test for a well-studied special class of nonstationary processes called locally stationary time series. In Section E, we introduce distribution-uniform extensions of many results from Mies and Steland [MS22], such as the strong Gaussian approximation for high-dimensional nonstationary nonlinear time series. As we will make clear, this distribution-uniform theory is necessary for proving that our CI tests have uniformly asymptotic Type-I error control.

1.1 Background

Many fundamental concepts in statistics can be expressed in terms of conditional independence, such as sufficiency, ancillarity, and the Markov condition [Daw79]. Moreover, conditional independence plays a key role in causal inference [Pea09], causal discovery [SGS00], and graphical modeling [KF09]. See [SP20; Zha+12; NBW21; PJS17; Run+23a; Zha+24] for more background. A related problem to ours is (unconditional) independence testing for nonstationary time series [Liu+23; Bru22; Bee21]. Concepts from graphical modeling that were originally developed for the iid setting have been extended to the time series setting [Dah00; Eic12]. Recently, Basu and Rao [BR23] introduced a graphical modeling framework for nonstationary time series.

There is a large literature on testing for conditional independence (CI) relationships between stationary time series, often in the context of forecasting. For example, there are tests based on characteristic functions [SW07] and copulas [BRT12]. Notably, the conditional mutual information-based test from Runge [Run18b] was used in Runge et al. [Run+19a]. See [SP11; Con12; AM12; Cha82; QKC15; FM82] for more background on conditional independence, transfer entropy, and Granger causality.

Classical Granger causality methods assume linear stationary dynamics with Gaussian errors. It is well-known that these assumptions are unlikely to hold for time series encountered in complicated settings, such as epidemiology and finance. Consequently, these classical methods will often draw erroneous conclusions. There have been significant developments to nonlinear Granger causality in recent years, such as neural Granger causality [Tan+21]. However, this approach still requires training a *computationally-expensive* neural network model and assuming that the time series evolves according to nonlinear *stationary* dynamics. In many situations, this is not problematic. In comparison, our perspective is based on adaptively forecasting *nonstationary* time series using *light-weight* learning algorithms.

CI testing is used in *constraint-based* and *hybrid* causal discovery algorithms for time series [Run+19c]. Recently, there has also been work on constraint-based causal discovery algorithms for nonstationary time series. See Huang et al. [Hua+20] and the related discussion in Dong et al. [Don+23] in the context of finance. We note that the underlying CI test that the causal discovery algorithm is based on must be well-suited to the data. If the underlying CI test is not appropriate for the data (e.g. if the test does not account for temporal dependence and/or nonstationarity), then the causal discovery algorithm will draw erroneous conclusions about the causal structure.

The existing literature on CI testing for nonstationary time series is very limited. In fact, the literature is practically non-existent once we restrict ourselves to the practical setting in which only one realization of the process is observed. We discuss the two tests (of which we are aware) that can be used with one realization of a nonstationary time series. First, Malinsky and Spirtes [MS19] focus on a type of nonstationarity in which the processes exhibit "stochastic trends" so that the first difference of the time series is stable. The consistency of their procedure requires assuming that the data were generated by a vector autoregressive linear model with iid Gaussian errors. Second, Flaxman, Neill, and Smola [FNS15] introduce a CI testing framework based on pre-whitening that can be used with nonstationary time series.

Throughout the paper, we will give many examples of how our CI test can be used for discovering forecasting signals. In the forecasting context, we use the terms conditional independence and strong Granger causality interchangeably. However, we also discuss conditional independence hypotheses that are more general than strong Granger causality with applications to causal discovery in mind.

2 The Generalized Covariance Measure for Nonstationary Time Series

2.1 Setting and notation

In this section, we employ the usual long-run asymptotics where the sample size n approaching infinity means that we observe the process for a longer amount of time. The processes described here are allowed to have very complicated nonstationarity that can be both abrupt and smooth. Again, we emphasize that we are only considering the practical setting in which we observe one realization of the process.

We work in a triangular array framework for high-dimensional nonstationary time series. We allow the dimensions of each of the processes to grow with n and the collection of distributions of the processes to change with n. Thus, we write the dependence of the processes, distributions, and dimensions on n explicitly going forward. Let $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t\in[n]}$, $[n] = \{1, \dots, n\}$, be the observed sequence. Let $d_X = d_{X,n}$, $d_Y = d_{Y,n}$, $d_Z = d_{Z,n}$ denote the dimensions of the processes X, Y, Z, respectively. Denote dimension $i \in [d_X]$ of $X_{t,n}$ by $X_{t,n,i}$, dimension $j \in [d_Y]$ of $Y_{t,n}$ by $Y_{t,n,j}$, and dimension $k \in [d_Z]$ of $Z_{t,n}$ by $Z_{t,n,k}$.

Next, we introduce notation for the time-offsets of $X_{t,n}$, $Y_{t,n}$, and $Z_{t,n}$ because we are often interested in time-lagged CI relationships. Denote by the time-offset $a \in A_i$ of $X_{t,n,i}$ by $X_{t,n,i,a} = X_{t+a,n,i}$, the time-offset $b \in B_j$ of $Y_{t,n,j}$ by $Y_{t,n,j,b} = Y_{t+b,n,j}$, and the time-offset $c \in C_k$ of $Z_{t,n,k}$ by $Z_{t,n,k,c} = Z_{t+c,n,k}$. Negative time-offsets are called *lags* of the process, and positive time-offsets are called *leads* of the process. Here, $A_i, B_j \subset \{-n+1, \dots, n-1\}$ are the set of time-offsets of $X_{t,n,i}$ and $Y_{t,n,j}$ under consideration. Similarly, $C_k \subset \{-n+1, \dots, -1, 0\}$ is the set of time-offsets of $Z_{t,n,k}$ under consideration, which we restrict to be non-positive so that the time-offset is known at time t. Denote the set of all time-offsets of $X_{t,n}$ by $A = \bigcup_{i=1}^{d_X} A_i$, all time-offsets of $Y_{t,n}$ by $B = \bigcup_{j=1}^{d_Y} B_j$, and all time-lags of $Z_{t,n}$ by $C = \bigcup_{k=1}^{d_Z} C_k$. Denote $a_{\max} = \max(A)$, $b_{\max} = \max(B)$, and $c_{\max} = \max(C)$. Denote the vectors with all dimensions and time-offsets by

$$\boldsymbol{X}_{t,n} = (X_{t,n,i,a})_{i\in[d_X], a\in A_i}, \quad \boldsymbol{Y}_{t,n} = (Y_{t,n,j,b})_{j\in[d_Y], b\in B_j}, \quad \boldsymbol{Z}_{t,n} = (Z_{t,n,k,c})_{k\in[d_Z], c\in C_k}.$$

Denote the dimensions of $\boldsymbol{X}_{t,n}$, $\boldsymbol{Y}_{t,n}$, $\boldsymbol{Z}_{t,n}$ by $\boldsymbol{d_X} = \sum_{i=1}^{d_X} |A_i|$, $\boldsymbol{d_Y} = \sum_{j=1}^{d_Y} |B_j|$, $\boldsymbol{d_Z} = \sum_{k=1}^{d_Z} |C_k|$, respectively. Also, denote the processes by

$$\boldsymbol{X}_n = (\boldsymbol{X}_{t,n})_{t\in\mathcal{T}_n}, \quad \boldsymbol{Y}_n = (\boldsymbol{Y}_{t,n})_{t\in\mathcal{T}_n}, \quad \boldsymbol{Z}_n = (\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}.$$

We allow the number of time-offsets and the number of dimensions to grow with n. To simplify the notation, let

$$\mathcal{M}_n = \{(i, j, a, b) : i \in [d_X], j \in [d_Y], a \in A_i, b \in B_j\}$$

be an index set of the dimensions and time-offsets. Going forward, we will refer to the tuple $(i, j, a, b) \in \mathcal{M}_n$ by $m \in \mathcal{M}_n$. The index set $\mathcal{M}_n$ depends on the sample size n through the dimensions and time-offsets, so its cardinality $M_n = |\mathcal{M}_n|$ may grow with n.

Since we are interested in time-lagged conditional independence relationships, it is often useful to refer to the subset of original times $\mathcal{T}_n \subseteq \{1, \dots, n\}$ in which *all* time-offsets of each dimension of $X_{t,n}$, $Y_{t,n}$, and $Z_{t,n}$ are actually observed, where

$$\mathcal{T}_n = \{1 + \lambda_n - \min(\{0\} \cup A \cup B \cup C), n - \max(\{0\} \cup A \cup B \cup C)\},$$

where $\lambda_n \geq 0$ is an optional late-starting parameter that will be used in Subsection B.2.

Note that if no negative time-offsets (i.e. lags) are used then $\min(\{0\} \cup A \cup B \cup C) = 0$, and if no positive time-offsets (i.e. leads) are used then $\max(\{0\} \cup A \cup B \cup C) = 0$. We will write $t \in \mathcal{T}_n$ instead of $t \in [n]$ to emphasize that we are only using the subset of times in which all time-offsets are

observed. Going forward, we will denote the first time of $\mathcal{T}_n$ by $\mathbb{T}_n^- = \min(\mathcal{T}_n)$ and the last time of $\mathcal{T}_n$ by $\mathbb{T}_n^+ = \max(\mathcal{T}_n)$

We require that the largest (in magnitude) time-offset grows at a slower rate than n such that as $n \to \infty$ we have $\min(\{0\} \cup A \cup B \cup C)/n \to 0$ and $\max(\{0\} \cup A \cup B \cup C)/n \to 0$ so that $T_n \to \infty$ arbitrarily slowly. In our theoretical analyses, we consider the asymptotics as $n \to \infty$ as usual. In the "best case" scenario in terms of regularity conditions for the temporal dependence and nonstationarity, the fastest that the number of dimensions and time-offsets M_n can grow with T_n is $M_n = O(T_n^{\frac{1}{4}-\delta})$ for some $\delta > 0$. See Subsection E.2 for a more detailed discussion about how quickly M_n can grow with T_n.

Now that the necessary notation has been introduced, we introduce the causal representation of the processes. For each $n \in \mathbb{N}$, we assume that each dimension of the observed sequence $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t\in[n]}$ can be represented as a nonlinear function of iid random elements. What follows is most similar to the definition of high-dimensional nonstationary time series from Mies and Steland [MS22], which builds on the work of Zhou and Wu [ZW09] and Wu [Wu05]. This representation for time series has a long history going back to at least Rosenblatt [Ros61] and Wiener [Wie66].

Assumption 1 (Causal representation of the observed processes)**.** *Let $\mathcal{H}_t^X = (\eta_t^X, \eta_{t-1}^X, \ldots)$, $\mathcal{H}_t^Y = (\eta_t^Y, \eta_{t-1}^Y, \ldots)$, $\mathcal{H}_t^Z = (\eta_t^Z, \eta_{t-1}^Z, \ldots)$ where $(\eta_t^X)_{t\in\mathbb{Z}}$, $(\eta_t^Y)_{t\in\mathbb{Z}}$, $(\eta_t^Z)_{t\in\mathbb{Z}}$ are sequences of iid random elements. Assume that for each time $t \in \mathcal{T}_n$ we can represent each dimension of the observed sequence as the output of an evolving nonlinear system that was given a sequence of iid inputs:*

$$X_{t,n,i} = G_{t,n,i}^X(\mathcal{H}_t^X),\ Y_{t,n,j} = G_{t,n,j}^Y(\mathcal{H}_t^Y),\ Z_{t,n,k} = G_{t,n,k}^Z(\mathcal{H}_t^Z).$$

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, we assume that $G_{t,n,i}^X(\cdot)$, $G_{t,n,j}^Y(\cdot)$, and $G_{t,n,k}^Z(\cdot)$ are each measurable functions from $\mathbb{R}^\infty$ to $\mathbb{R}$ (where we endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections) such that $G_{t,n,i}^X(\mathcal{H}_s^X)$, $G_{t,n,j}^Y(\mathcal{H}_s^Y)$, $G_{t,n,k}^Z(\mathcal{H}_s^Z)$ are each well-defined random variables for each $s \in \mathbb{Z}$ and $(G_{t,n,i}^X(\mathcal{H}_s^X))_{s\in\mathbb{Z}}$, $(G_{t,n,j}^Y(\mathcal{H}_s^Y))_{s\in\mathbb{Z}}$, $(G_{t,n,k}^Z(\mathcal{H}_s^Z))_{s\in\mathbb{Z}}$ are each stationary ergodic time series.

In the rest of this section, we will introduce several more causal representations. Let us state some properties that all causal representations will have to avoid repeating the same ideas many times. These causal representations will all be measurable functions on $\mathbb{R}^\infty$, and we will always endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections. As stated in Assumption 1, the causal mechanism at each time $t \in \mathcal{T}_n$ with the input sequence up to some time $s \in \mathbb{Z}$ is a well-defined r.v., and the process induced by considering the sequence of inputs up to each time $s \in \mathbb{Z}$ is a stationary ergodic time series.

In view of Assumption 1, we have the following causal representations for the observed process including all dimensions as

$$X_{t,n} = G_{t,n}^X(\mathcal{H}_t^X) = (G_{t,n,i}^X(\mathcal{H}_t^X))_{i\in[d_X]},$$

$$Y_{t,n} = G_{t,n}^Y(\mathcal{H}_t^Y) = (G_{t,n,j}^Y(\mathcal{H}_t^Y))_{j\in[d_Y]},$$

$$Z_{t,n} = G_{t,n}^Z(\mathcal{H}_t^Z) = (G_{t,n,k}^Z(\mathcal{H}_t^Z))_{k\in[d_Z]}.$$

Also, we have a causal representation for dimensions $i \in [d_X]$, $j \in [d_Y]$, $k \in [d_Z]$ with time-offsets $a \in A_i$, $b \in B_j$, $c \in C_k$

$$X_{t,n,i,a} = G_{t,n,i,a}^X\left(\mathcal{H}_{t,a}^X\right) = G_{t+a,n,i}^X\left(\mathcal{H}_{t+a}^X\right),$$

$$Y_{t,n,j,b} = G_{t,n,j,b}^Y\left(\mathcal{H}_{t,b}^Y\right) = G_{t+b,n,j}^Y\left(\mathcal{H}_{t+b}^Y\right),$$

$$Z_{t,n,k,c} = G_{t,n,k,c}^Z\left(\mathcal{H}_{t,c}^Z\right) = G_{t+c,n,k}^Z\left(\mathcal{H}_{t+c}^Z\right)$$

where $\mathcal{H}_{t,a}^X = (\eta_{t+a}^X, \eta_{t-1+a}^X, \ldots)$, $\mathcal{H}_{t,b}^Y = (\eta_{t+b}^Y, \eta_{t-1+b}^Y, \ldots)$, and $\mathcal{H}_{t,c}^Z = (\eta_{t+c}^Z, \eta_{t-1+c}^Z, \ldots)$. We can then write the causal representation of the vectors with all dimensions and time-offsets as

$$\boldsymbol{X}_{t,n} = \boldsymbol{G}_{t,n}^{\boldsymbol{X}}(\mathcal{H}_t^{\boldsymbol{X}}) = (G_{t,n,i,a}^X(\mathcal{H}_{t,a}^X))_{i\in[d_X],a\in A_i},$$

$$\boldsymbol{Y}_{t,n} = \boldsymbol{G}_{t,n}^{\boldsymbol{Y}}(\mathcal{H}_t^{\boldsymbol{Y}}) = (G_{t,n,j,b}^Y(\mathcal{H}_{t,b}^Y))_{j\in[d_Y],b\in B_j},$$

$$\boldsymbol{Z}_{t,n} = \boldsymbol{G}_{t,n}^{\boldsymbol{Z}}(\mathcal{H}_t^{\boldsymbol{Z}}) = (G_{t,n,k,c}^Z(\mathcal{H}_{t,c}^Z))_{k\in[d_Z],c\in C_k},$$

where $\mathcal{H}_t^{\boldsymbol{X}} = (\eta_t^{\boldsymbol{X}}, \eta_{t-1}^{\boldsymbol{X}}, \ldots)$, $\mathcal{H}_t^{\boldsymbol{Y}} = (\eta_t^{\boldsymbol{Y}}, \eta_{t-1}^{\boldsymbol{Y}}, \ldots)$, $\mathcal{H}_t^{\boldsymbol{Z}} = (\eta_t^{\boldsymbol{Z}}, \eta_{t-1}^{\boldsymbol{Z}}, \ldots)$, and $\eta_t^{\boldsymbol{X}} = \eta_{t+a_{\max}}^{X}$, $\eta_t^{\boldsymbol{Y}} = \eta_{t+b_{\max}}^{Y}$, $\eta_t^{\boldsymbol{Z}} = \eta_{t+c_{\max}}^{Z}$.

Let Ω be a sample space, $\mathcal{B}$ the Borel sigma-algebra, and $(\Omega, \mathcal{B})$ a measurable space. For fixed $n \in \mathbb{N}$, let $(\Omega, \mathcal{B})$ be equipped with a family of probability measures $(\mathbb{P}_P)_{P \in \mathcal{P}_n}$ so that the distribution of the stochastic system

$$(G_{t,n}^X(\mathcal{H}_s^X), G_{t,n}^Y(\mathcal{H}_s^Y), G_{t,n}^Z(\mathcal{H}_s^Z))_{t \in [n], s \in \mathbb{Z}}$$

under $\mathbb{P}_P$ is $P \in \mathcal{P}_n$ where $\mathcal{P}_n$ is a collection of such distributions. The family of probability measures $(\mathbb{P}_P)_{P \in \mathcal{P}_n}$ is defined with respect to the same measurable space $(\Omega, \mathcal{B})$, but need not have the same dominating measure. Denote the family of probability spaces by $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P \in \mathcal{P}_n}$ and a sequence of such families of probability spaces by $((\Omega, \mathcal{B}, \mathbb{P}_P)_{P \in \mathcal{P}_n})_{n \in \mathbb{N}}$. For a given sample size $n \in \mathbb{N}$ and distribution $P \in \mathcal{P}_n$, let $\mathbb{E}_P(\cdot)$ denote the expectation of a random variable with distribution determined by P, and let $\mathbb{P}_P(E)$ denote the probability of an event $E \in \mathcal{B}$.

Lastly, we use the notation $o_{\mathcal{P}}(\cdot)$ and $O_{\mathcal{P}}(\cdot)$ in the same way that Shah and Peters [SP20] do, so we replicate their notation here. Let $(V_{P,n})_{n \in \mathbb{N}, P \in \mathcal{P}_n}$ be a family of sequences of random variables with distributions determined by $P \in \mathcal{P}_n$ for some collection of distributions $\mathcal{P}_n$ that will be made clear from the context. We write $V_{P,n} = o_{\mathcal{P}}(1)$ to mean that for all $\epsilon > 0$, we have

$$\sup_{P \in \mathcal{P}_n} \mathbb{P}_P(|V_{P,n}| > \epsilon) \to 0.$$

Also, by $V_{P,n} = O_{\mathcal{P}}(1)$ we mean for all $\epsilon > 0$, there exists a constant $K > 0$ such that

$$\sup_{n \in \mathbb{N}} \sup_{P \in \mathcal{P}_n} \mathbb{P}_P(|V_{P,n}| > K) < \epsilon.$$

Let $(W_{P,n})_{n \in \mathbb{N}, P \in \mathcal{P}_n}$ be another family of sequences of random variables. By $V_{P,n} = o_{\mathcal{P}}(W_{P,n})$ we mean $V_{P,n} = W_{P,n} R_{P,n}$ and $R_{P,n} = o_{\mathcal{P}}(1)$, and by $V_{P,n} = O_{\mathcal{P}}(W_{P,n})$ we mean $V_{P,n} = W_{P,n} R_{P,n}$ and $R_{P,n} = O_{\mathcal{P}}(1)$.

2.2 Conditional independence for nonstationary time series

In this section, we show how to construct a test for

$$H_{0,n}^{\mathrm{CI}} : X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } m \in \mathcal{M}_n, \tag{1}$$

that is asymptotically valid as $n \to \infty$, uniformly over a large collection of distributions for which this global null hypothesis holds. Let $\mathcal{P}_{0,n}^{\mathrm{CI}}$ be a collection of distributions such that $H_{0,n}^{\mathrm{CI}}$ is true, and denote a sequence of such collections by $(\mathcal{P}_{0,n}^{\mathrm{CI}})_{n \in \mathbb{N}}$. In the univariate case where $d_X = 1$ and $d_Y = 1$, simply ignore the dimension/time-offset indices $m = (i, j, a, b) \in \mathcal{M}_n$.

In the familiar context of forecasting, the global null hypothesis (1) can be viewed as a null hypothesis of (strong) nonlinear Granger noncausality at all times. In this setting, $\boldsymbol{Z}_{t,n}$ would be the vector of covariates known at time t, $X_{t,n,i,a}$ would be an auxiliary time series known at time t, and $Y_{t,n,j,b}$ would be the forecasting target. Note that a must be a non-positive integer indicating the current value or a-th lag, and b must be a non-negative integer indicating the current value for nowcasting or the b-th forecast horizon.

There are four different alternative hypotheses $H_{1,n}^{\mathrm{CI}}$ that can be used. We can always use

$$X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for at least one } t \in \mathcal{T}_n, \text{ for at least one } m \in \mathcal{M}_n. \tag{2}$$

However, it may not necessarily be useful to know that conditional independence fails to hold at *some* point in time t for *some* index m. Fortunately, we can often conclude more than this. If domain knowledge suggests that it is reasonable to restrict the collection of distributions $\mathcal{P}_n$ to be those in which either conditional independence *or* conditional dependence holds for *all* times, then we can replace the alternative hypothesis (2) with

$$X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for at least one } m \in \mathcal{M}_n. \tag{3}$$

In this case, each time series is allowed to be nonstationary in non-specific ways, but the conditional independence relationships for each of the dimension/time-offset indices $m = (i, j, a, b) \in \mathcal{M}_n$ must

be *time-invariant*. Additionally, if the indices m can be viewed as a "group of time series" that all share the same time-invariant conditional independence structure, then we can replace the alternative hypothesis (2) with

$$X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } m \in \mathcal{M}_n. \tag{4}$$

Again, each time series is allowed to be nonstationary in non-specific ways. However, in this case the conditional independence relationship must be both time-invariant and index-invariant. In this "group of time series" setting, we can often gain power by increasing the number of time series in the group (e.g. nodes in a sensor network). Lastly, if the indices $m \in \mathcal{M}_n$ can be treated as a group of time series with a *time-varying* conditional independence structure, then we can replace the alternative hypothesis (2) with

$$X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for at least one } t \in \mathcal{T}_n, \text{ for all } m \in \mathcal{M}_n. \tag{5}$$

Overall, our testing framework offers a large toolkit that can be used for variety of settings when we only have one observation of a high-dimensional nonstationary time series.

In Section A, we discuss simultaneous testing procedures so that we can determine *when* certain conditional independence relationships hold or not. More specifically, we can recover time-windows in which conditional independence holds for particular dimension/time-offset indices $m \in \mathcal{M}_n$ all while controlling the familywise error rate. This multiple testing perspective can be very useful because in many complex settings it may be more realistic to assume that conditional independence relationships are stable over shorter time segments, but not necessarily over the entire duration of the time series. Under a different set of assumptions, it is also possible to use the locally stationary time series framework [Dah97; Dah12; DRW19] to infer conditional independence at specific points in time. We devote Section D entirely to this setting.

We formulate the null hypothesis of conditional independence (1) in this way because it makes explicit how conditional dependencies could arise among the different leads and lags of each dimension at different points in time. In other words, our formulation emphasizes the dynamic "flow of information" between two time series conditional on the information contained in a third time series. To put this statement into context, let us consider different ways that we could have formulated the null hypothesis of conditional independence.

As discussed in Hochsprung et al. [Hoc+23], for each $n \in \mathbb{N}$, the null hypothesis that

$$\boldsymbol{X}_{t,n} \perp\!\!\!\perp \boldsymbol{Y}_{t,n} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \tag{6}$$

implies our formulation (1). Hence, to test for formulation (6), one can test for the hypothesis (1) as we do and then reject the other formulation (6) if and only if we reject (1). By the same arguments of Lemma 1 from Hochsprung et al. [Hoc+23], this induced test for (6) will have uniformly asymptotic level if the original test for the hypothesis (1) does.

2.3 Basic setup and main ideas

For a fixed sample size $n \in \mathbb{N}$, distribution $P \in \mathcal{P}_n$, time $t \in \mathcal{T}_n$ and dimension/time-offset index tuple $m = (i, j, a, b) \in \mathcal{M}_n$, we can always decompose

$$X_{t,n,i,a} = f_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}) + \varepsilon_{P,t,n,i,a}, \; Y_{t,n,j,b} = g_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}) + \xi_{P,t,n,j,b},$$

where $f_{P,t,n,i,a}(\boldsymbol{z}) = \mathbb{E}_P(X_{t,n,i,a}|\boldsymbol{Z}_{t,n} = \boldsymbol{z})$ and $g_{P,t,n,j,b}(\boldsymbol{z}) = \mathbb{E}_P(Y_{t,n,j,b}|\boldsymbol{Z}_{t,n} = \boldsymbol{z})$ are the time-varying regression functions. Denote the corresponding product of errors at time t by

$$R_{P,t,n,m} = \varepsilon_{P,t,n,i,a}\xi_{P,t,n,j,b}.$$

Next, let $\hat{f}_{t,n,i,a}$ and $\hat{g}_{t,n,j,b}$ be estimates of $f_{P,t,n,i,a}$ and $g_{P,t,n,j,b}$ created by time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively. Let

$$\hat{\varepsilon}_{t,n,i,a} = X_{t,n,i,a} - \hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}),$$
$$\hat{\xi}_{t,n,j,b} = Y_{t,n,j,b} - \hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}),$$

be the corresponding residuals, and denote the product of these residuals at time t by

$$\hat{R}_{t,n,m} = \hat{\varepsilon}_{t,n,i,a}\hat{\xi}_{t,n,j,b}.$$

The covariate process $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ is a nonstationary time series and each of the dimensions can depend on one another. We can, for instance, include any lags of any of the dimensions of the original time series $Z_{t,n}$ since these lags are known at time t. The error processes $(\varepsilon_{P,t,n,i,a})_{t\in\mathcal{T}_n}$, $(\xi_{P,t,n,j,b})_{t\in\mathcal{T}_n}$ can also be nonstationary time series that depend on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$, $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$, respectively. In Section B, we discuss the technical concepts that are necessary for understanding the theoretical justifications of our test.

Our testing framework can be considered an extension of the *generalized covariance measure* (GCM) CI test from Shah and Peters [SP20] to the nonstationary time series setting in which the conditional independence relationships can change over time. We first briefly summarize the univariate version of the original GCM test. Let X, Y be two random variables and let Z be a random vector. Assume the joint distribution of (X, Y, Z) is absolutely continuous with respect to the Lebesgue measure. The GCM test is based on the "weak" conditional independence criterion of Daudin [Dau80], which states that if $X \perp\!\!\!\perp Y \mid Z$ then $\mathbb{E}_P[\phi(X,Z)\varphi(Y,Z)] = 0$ for all functions $\phi \in L^2_{X,Z}$ and $\varphi \in L^2_{Y,Z}$ such that $\mathbb{E}_P[\phi(X,Z) \mid Z] = 0$ and $\mathbb{E}_P[\varphi(Y,Z) \mid Z] = 0$. Thus, under the null hypothesis of conditional independence, the expectation of the products of errors $\mathbb{E}_P(\varepsilon\xi)$ from the regressions $X = \phi(Z) + \varepsilon$ and $Y = \varphi(Z) + \xi$, or equivalently the expected conditional covariance $\mathbb{E}_P[\mathrm{Cov}_P(X, Y|Z)]$, is equal to zero. The GCM test statistic is based on the normalized sum of the products of residuals from the regressions of X on Z and Y on Z.

Let us translate the "weak" conditional independence criterion of Daudin [Dau80] into the nonstationary time series setting. Assume that for each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$ the joint distribution of $(X_{t,n,i,a}, Y_{t,n,j,b}, \boldsymbol{Z}_{t,n})$ is absolutely continuous with respect to the Lebesgue measure. For some $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, if $X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n}$ then $\mathbb{E}_P[\phi(X_{t,n,i,a}, \boldsymbol{Z}_{t,n})\varphi(Y_{t,n,j,b}, \boldsymbol{Z}_{t,n})] = 0$ for all functions $\phi \in L^2_{X_{t,n,i,a},\boldsymbol{Z}_{t,n}}$ and $\varphi \in L^2_{Y_{t,n,j,b},\boldsymbol{Z}_{t,n}}$ such that $\mathbb{E}_P[\phi(X_{t,n,i,a}, \boldsymbol{Z}_{t,n}) \mid \boldsymbol{Z}_{t,n}] = 0$ and $\mathbb{E}_P[\varphi(Y_{t,n,j,b}, \boldsymbol{Z}_{t,n}) \mid \boldsymbol{Z}_{t,n}] = 0$ and hence the (time-varying) expected conditional covariance at time $t \in \mathcal{T}_n$

$$\rho_{P,t,n,m} = \mathbb{E}_P[\mathrm{Cov}_P(X_{t,n,i,a}, Y_{t,n,j,b}|\boldsymbol{Z}_{t,n})]$$

is equal to zero. Hence, the process of error products from the time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ has mean zero. The exact notation will be introduced in Subsection 2.3.

Whereas other tests for conditional independence for stochastic processes treat the processes as entire objects, our testing framework is based on the perspective of detecting *conditional flows of information* between the processes via their time-varying expected conditional covariances. In particular, under the global null hypothesis of conditional independence (1), all of the expected conditional covariances $\rho_{P,t,n,m}$ are equal to zero. Hence, we aim to detect conditional dependencies by determining whether the time-varying expected conditional covariances $\rho_{P,t,n,m}$ deviate from zero *at any point in time along the path* for any index $m \in \mathcal{M}_n$.

Note that while our test is based on the expected conditional covariance functional, our framework can be easily adapted to be used with any functional that is equal to zero under the null of conditional independence. Zhang and Janson [ZJ20] discuss some of the shortcomings of the expected conditional covariance function, in particular that it lacks sensitivity to nonlinear relationships and interactions. We leave further explorations and comparisons with other functionals for future work.

Our test statistic is based on the products of residuals from the time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively. We reject the null hypothesis of conditional independence (1) if the magnitude of the partial sum process of the products of residuals ever becomes "too large" *at some point in time* over the entire time period. The limiting distribution of our test statistic is therefore based on the *strong Gaussian approximation* for high-dimensional nonstationary time from Mies and Steland [MS22], as opposed to a central limit theorem.

Analogously to the GCM test from Shah and Peters [SP20], our test only has power against alternatives in which the time-varying expected conditional covariances are non-zero for at least *some points in time*. In other words, if the time-varying expected conditional covariances are *always* zero we cannot hope to detect whether conditional dependence holds at some or all times. Note that for nonstationary time series the time-varying expected conditional covariances can be zero at some times

and non-zero at other times even if the corresponding conditional dependence relationships hold at all times. Hence, our test is useful for both alternative hypotheses (2) and (3).

2.4 A practical test

The key to our bootstrap-based testing procedure is a consistent estimator for the local long-run covariance matrices of the products of errors based on the products of residuals. We use the same cumulative covariance estimator from Mies and Steland [MS22]. However, the theoretical results in Mies and Steland [MS22] for covariance estimation are for using the estimator with the original time series. Hence, we extend the covariance estimation results so that it can be used with products of residuals. In Section E, we slightly extend many of the theoretical results from Mies and Steland [MS22] so that they hold distribution-uniformly.

Let $\hat{\boldsymbol{R}}_{t,n} = (\hat{R}_{t,n,m})_{m\in\mathcal{M}_n}$ be the high-dimensional vector process containing all indices of the products of residuals. We use the following estimator to estimate the *cumulative* covariance process of the product of errors $\boldsymbol{R}_{t,n} = (R_{P,t,n,m})_{m\in\mathcal{M}_n}$,

$$\hat{Q}^{\boldsymbol{R}}_{k,n} = \sum_{r=L_n+\mathbb{T}_n^- -1}^{k} \frac{1}{L_n} \left(\sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{R}}_{s,n} \right) \left(\sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{R}}_{s,n} \right)^\top,$$

where $L_n \in \mathbb{N}$ is a lag window size parameter where $L_n \asymp T_n^\zeta$ for some $\zeta \in (0, \frac{1}{2})$. For expository purposes, we delay the very technical discussions about this estimator and the cumulative covariance process until Subsections B.1 and E.3. This way, we can present the simple algorithm for our practical test and give high-level intuition for why it works. Going forward, we will denote and $\hat{Q}^{\boldsymbol{R}}_n = (\hat{Q}^{\boldsymbol{R}}_{t,n})_{t=L_n+\mathbb{T}_n^- -1}^{\mathbb{T}_n^+}$. Similarly, define the rolling-window estimator of the instantaneous covariance by $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n} = \hat{Q}^{\boldsymbol{R}}_{t,n} - \hat{Q}^{\boldsymbol{R}}_{t-1,n}$ which will play a key role in the algorithm.

Define $\hat{\boldsymbol{R}}_n = (\hat{\boldsymbol{R}}_n)_{t=L_n+\mathbb{T}_n^- -1}^{\mathbb{T}_n^+}$. We will show that test statistics

$$S_{n,p}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- +L_n-1,\ldots,\mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t\le s} \hat{\boldsymbol{R}}_{t,n} \right\|_p$$

based on the ℓ_p-norm for any $p \ge 2$ will have uniformly asymptotic level under the assumptions from Subsection B.3. For instance, we can use

$$S_{n,\infty}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- +L_n-1,\ldots,\mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t\le s} \hat{\boldsymbol{R}}_{t,n} \right\|_\infty,$$

or

$$S_{n,2}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- +L_n-1,\ldots,\mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t\le s} \hat{\boldsymbol{R}}_{t,n} \right\|_2.$$

It is well known that max-norm test statistics have good power against sparse alternatives while ℓ_2-norm test statistics have good power against dense alternatives. Since in practice we might not know which test statistic to use, we suggest using a Bonferroni combination test, as in Zhang and Shao [ZS24] and Gao, Wang, and Shao [GWS23]. This way we can achieve good power against both dense and sparse alternatives. That is, we reject the null with the Bonferroni combination test at significance level α if the p-value of either the test using $S_{n,\infty}(\hat{\boldsymbol{R}}_n)$ or $S_{n,2}(\hat{\boldsymbol{R}}_n)$ drops below $\alpha/2$. We leave the theoretical results for the asymptotic independence of $S_{n,\infty}(\hat{\boldsymbol{R}}_n)$ and $S_{n,2}(\hat{\boldsymbol{R}}_n)$ as well as the simulation results for the Bonferroni combination test for future work.

Next, we will introduce additional assumptions and details before stating the main theorem. Define the offsets $\nu_n \to 0$ and $\tau_n \to 0$

$$\nu_n \gg \log(T_n) M_n \left[\left(\frac{M_n}{T_n} \right)^{2\xi(\bar{q}^R, \bar{\beta}^R)} + \tau_n^{-2} \left(\varphi_{n,1} + \varphi_{n,2} \right) \right], \tag{7}$$

where

$$\varphi_{n,1} = T_n^{-\frac{1}{2}}(\bar{\Gamma}_n^R)^{\frac{1}{2}} L_n^{\frac{1}{4}} + T_n^{-\frac{1}{4}} M_n^{\frac{1}{4}} L_n^{\frac{1}{4}} + L_n^{-\frac{1}{2}} + L_n^{1-\frac{\bar{\beta}^R}{2}} + T_n^{-1},$$

comes from the covariance estimation error,

and

$$\varphi_{n,2} = \tau_n^{\frac{1}{2}} M_n^{-\frac{1}{4}} + \tau_n M_n^{-\frac{1}{2}},$$

comes from the prediction errors since we use the products of residuals.

The constants $\bar{\beta}^R > 2$, $\bar{q}^R > 4$, are related to regularity conditions controlling the temporal dependence and nonstationarity uniformly over the collection of distributions $\mathcal{P}_n$ in Subsection B.3 and $\xi(\bar{q}^R, \bar{\beta}^R)$ is a rate defined in Subsection E.2. In practice, we discuss reasonable choices for the offsets in Section H based on the simulation results.

The following *rate double robustness* property allows for our conditional independence test to be tolerant of one of the predictors having relatively large prediction errors.

Assumption 2 (Doubly robust rate requirements for predictors)**.** *For $n \in \mathbb{N}$ and some collection of distributions $\mathcal{P}_n$, assume that the predictors satisfy*

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left|\left| \hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right|\right|_2^2 \right) \sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left|\left| \hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right|\right|_2^2 \right) \right) = o(T_n \tau_n^2),$$

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left|\left| \hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right|\right|_2^2 \right) \right) = o(T_n M_n^{-1} \tau_n^2),$$

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left|\left| \hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right|\right|_2^2 \right) \right) = o(T_n M_n^{-1} \tau_n^2).$$

This double robustness property can be especially important in the context of forecasting. Suppose one of the predictors, say $\hat{g}_{t,n,j,b}$, is forecasting the target $Y_{t,n,j,b}$ (i.e. the time-offset $b > 0$), and the other predictor $\hat{f}_{t,n,i,a}$ is nowcasting a new auxiliary signal $X_{t,n,i,a}$ (i.e. the time-offset $a = 0$). We would like to test whether there is any additional information about $Y_{t,n,j,b}$ contained in $X_{t,n,i,a}$ that is not already in the existing forecasting signals $\boldsymbol{Z}_{t,n}$. The nowcasting model $\hat{f}_{t,n,i,a}$ will have (relatively small) prediction errors that will "make up" for the (relatively large, if b is large) prediction errors of the forecasting model $\hat{g}_{t,n,j,b}$.

We will calculate the cumulative covariance process $\hat{Q}_n^{\boldsymbol{R}}$ based on the process of products of residuals. In practice, we can approximate the $(1-\alpha)$ quantile via Monte Carlo. We reject the null hypothesis of conditional independence at level α if $S_n(\hat{\boldsymbol{R}}_n) > a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}}) + \tau_n$ for some offsets $\nu_n \to 0$ and $\tau_n \to 0$ defined in Subsection B.2. The following result shows that if the time-varying regression estimators satisfy certain requirements, then $a_\alpha(\hat{Q}_n^{\boldsymbol{R}})$ closely approximates the $(1-\alpha)$ quantile of the test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ and thus we can use it to calibrate a test. For expository purposes, we delay discussion of the Assumptions 4, 5, 6, 7, 8 controlling the temporal dependence and nonstationarity until Section B.

Theorem 2.1. *Suppose that Assumptions 1, 2, 3, 4, 5, 6, 7, 8, all hold for the collection of distributions $\mathcal{P}^*_{0,n} \subset \mathcal{P}^{CI}_{0,n}$ and the predictors. If the offsets $\tau_n \to 0$ and $\nu_n \to 0$ are chosen such that condition (7) holds, we have that*

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(S_{n,p}(\hat{\boldsymbol{R}}_n) > a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}}) + \tau_n \right) \leq \alpha.$$

We can summarize the main steps of our bootstrap-based test with the following procedure:

1. For each $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$ calculate the time-t residuals $\hat{\varepsilon}_{t,n,i,a}$ and $\hat{\xi}_{t,n,j,b}$ using adaptive regression procedures based on $(X_{s,n,i,a}, \boldsymbol{Z}_{s,n})_{s \leq t}$ and $(Y_{s,n,j,b}, \boldsymbol{Z}_{s,n})_{s \leq t}$, respectively.

2. For each $t = L_n + \mathbb{T}_n^- - 1, \ldots, \mathbb{T}_n^+$, calculate

$$\hat{\boldsymbol{\Sigma}}_{t,n} = \frac{1}{L_n} \left(\sum_{s=t-L_n+1}^{t} \hat{\boldsymbol{R}}_{s,n} \right) \left(\sum_{s=t-L_n+1}^{t} \hat{\boldsymbol{R}}_{s,n} \right)^\top.$$

3. For each $t = L_n + \mathbb{T}_n^- - 1, \ldots, \mathbb{T}_n^+$, simulate *independent* Gaussian random vectors $\tilde{\boldsymbol{R}}_{t,n} \sim \mathcal{N}(0, \hat{\boldsymbol{\Sigma}}_{t,n})$ and let $\tilde{\boldsymbol{R}}_n$ denote the entire simulated process. For some $p \geq 2$, calculate the test statistic $S_{n,p}(\tilde{\boldsymbol{R}}_n)$ using the simulated process.

4. Repeat the previous step many times and compute the offset-corrected $1-\alpha$ quantile $a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}})+\tau_n$.

5. For some $p \geq 2$, calculate the test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ using the observed residuals.

6. Reject the null hypothesis at level α if $S_{n,p}(\hat{\boldsymbol{R}}_n)$ exceeds $a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}}) + \tau_n$.

2.5 A note on adaptive prediction for nonstationarity time series

Let us briefly reflect on the rate requirements for the adaptive predictors from Assumption 2 needed for Theorem 2.1. The sum of the squared prediction errors must grow sublinearly, which means that we must be able to slowly *improve* our forecasts as we observe the nonstationary process for a longer amount of time. It is clearly not possible to meet these rate requirements for processes with *arbitrary* nonstationarity. Hence, we must introduce *some* restrictions on the class of nonstationary processes that we consider.

The literature on statistical learning for nonstationary time series has much to say about this. Over the last decade, this literature has been able to move beyond the restrictive assumptions of stationarity and mixing [Yu94; WLT20; KV02; ALW13] (or asymptotic stationarity [AD12]) by describing nonstationarity in terms of discrepancy measures [KM14; KM15; KM17; MK20; HY19; Han21b]. Most of this work has been done in the context of forecasting. Notably, Hanneke and Yang [HY19] shows that for bounded VC subgraph classes, if the sum of changes in the marginal distributions for contiguous time points grows sublinearly in the number of samples, then the cumulative excess risk will grow sublinearly in the number of predictions.

Recently, the notion of *learnability* for target functions under general stochastic processes was introduced in Dawid and Tewari [DT20]. Additionally, Hanneke [Han21a] studied the question of whether there exist learning rules that achieve a small long-run loss for any target function (almost surely, as the duration of the process $n \to \infty$) given only that such *learning is possible* for the stochastic process. Hanneke [Han21a] establishes that there do exist such learning rules in the *self-adaptive* learning setting, where the predictor can be updated after each prediction. Dawid and Tewari [DT20] and Hanneke [Han21a] focus on the problem of estimating a (time-invariant) target function using non-iid observations. On the other hand, we are concerned with adaptive predictions for nonstationary time series, which is more similar to the motivations of the aforementioned works [KM14; KM15; KM17; MK20; HY19; Han21b].

There do not exist statistical guarantees for adaptive prediction algorithms in a theoretical framework similar to that of Mies and Steland [MS22]. We see this as an opportunity for new avenues of research in statistical learning for time series with complicated forms of nonstationarity. One research direction will involve identifying a more basic set of conditions for the time-varying regression functions such that the required rates can be achieved. After developing some additional technical tools, it may also be possible to develop theoretical guarantees for an adaptive learning algorithm for nonstationary processes within this framework. This line of theoretical inquiry has connections to the problems of domain adaptation and covariate shift [Han21a; Hua+06; Cor+08; Ben+10; HY19; HK24; RCS21] and may inspire practical algorithms for forecasting nonstationary time series (e.g. by systematically identifying similar regimes across time). In Section D, we discuss examples of time-varying regression estimators for the more well-studied framework of locally stationary time series, such as the L_2 boosting algorithm of Yousuf and Ng [YN21].

In practice, users of our CI test are tasked with coming up with estimates of the time-varying regression functions from an observed finite-length nonstationary time series. Users can employ *any* time-varying regression procedure for nonstationary time series. Given that learning algorithms for nonstationary processes are not widely available, we discuss some practical approaches which can leverage popular learning algorithms.

Consider the popular ad-hoc approach of estimating regression models on a transformed time series that is "more stationary" than the original time series. Possible transformations include differencing, detrending, or logarithmic differencing. After estimating this model, its predictions can be transformed

back to the scale of the original time series although this is not necessary. Note that this approach requires changing the hypothesis of conditional independence so that it reflects the transformed time series (e.g. using log returns instead of prices, if a logarithmic difference is used). See the example in Malinsky and Spirtes [MS19] about how the causal structure of a differenced time series can differ from the original time series. In many cases, the transformed time series will still be nonstationary despite the practitioner's best efforts.

Hence, to come up with estimates of the time-varying regression functions, we suggest estimating separate nonlinear regression models on *moving time-windows* of the (possibly transformed) time series. For decades, rolling-window estimation of relatively simple regression models been the approach favored by many forecasting practitioners. If the time-windows are moderately sized, it may be reasonable to fit nonlinear regression models such as random forests [Bre01], XGBoost [CG16], and LightGBM [Ke+17]. Notably, forecasting approaches based on LightGBM have performed well in the M5 forecasting competition [MSA22]. Building on this idea, it is also possible to use sample weights that decay exponentially in time (or down-weighted using one-way kernel smoothing) so that more recent time points have a higher weight.

In many cases, we will already have an forecasting model $\hat{g}_{t,n,j,b}$ for $Y_{t,n,j,b}$ that was computationally expensive to train. In this case, we can simply use the historical residuals (often already stored in a database) that were formed from the real-time forecasts of $Y_{t,n,i,a}$. When faced with a large database of alternative forecasting signals, it is often prudent to opt for lightweight regression methods to construct the adaptive nowcasting model $\hat{f}_{t,n,i,a}$, such as rolling-window approaches or nonstationary online learning methods.

In some sense, the framework of *locally stationary time series* [Dah97; Dah12; DRW19] formalizes the idea of fitting "local" regressions based on rolling time-windows or decaying sample weights based on one-way kernel-weighted time-averages. Instead of the usual time series long-run asymptotics, the locally stationary framework utilizes infill asymptotics so that time is rescaled to the unit interval. That is, $n \to \infty$ no longer means that we get more information about the future. Instead, $n \to \infty$ means that we get more observations about each *local* structure of the nonstationary time series. This can be done by using time-varying nonlinear regression estimators for locally stationary time series, such as Yousuf and Ng [YN21]. This theoretical framework requires different arguments, notation, and assumptions, so we dedicate Section D entirely to the locally stationary time series setting.

To be clear, we prove the theoretical guarantees for the *same testing procedure* using *two different asymptotic frameworks* for nonstationary time series. The framework introduced in this section uses long-run asymptotics and allows for more general forms of nonstationarity, but it is less well-studied than the locally stationary time series framework. On the other hand, the locally stationary time series framework has a long history but requires that the time series changes smoothly and "slowly" over time.

Throughout this discussion, we have completely avoided the assumption of stationarity. However, it is worth considering how things would simplify if one is willing to assume stationarity. The essential difference is that one must use a suitable (i.e. for stationary time series) Gaussian approximation, regression estimator, and covariance matrix estimator. More specifically, if one assumes that the regression functions are time-invariant, then standard regression estimators can be used. The statistical guarantees of many learning algorithms have been studied the stationary time series setting [LKS05; SA09; SHS09; DN20]. Further, if one assumes that the expected conditional covariances are stationary, then it suffices to use a test statistic based on the ℓ_p norm of the full sum of the residual products

$$S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n) = \left|\left| \frac{1}{\sqrt{T_n}} \sum_t \hat{\boldsymbol{R}}_{t,n} \right|\right|_p$$

for some $p \geq 2$. The covariance matrix of the error products can be estimated from the residual products by using standard covariance estimators for stationary time series, see Wu and Xiao [WX12] and Wu [Wu11]. One could even allow *some* aspects of the process to be to be nonstationary. For example, we can allow for the product of errors to be mean-stationary but covariance-nonstationary. In this case, we can use the same estimator $\hat{Q}^{\boldsymbol{R}}_{k,n}$ of the cumulative covariance defined above. However, we would only need an estimate $\hat{Q}^{\boldsymbol{R}}_{\mathbb{T}^+_n,n}$ of the full cumulative covariance up to the last time point $\mathbb{T}^+_n$. The $1-\alpha$ quantile of $S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n)$ can be approximated via Monte Carlo and the theoretical justifications can be based on the same strong Gaussian approximation for nonstationary time series. Under the

assumption of time-invariant conditional independence relationships, one could use the alternative hypothesis $X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n}$ either (1) for all $t \in \mathcal{T}_n$, for at least one $m \in \mathcal{M}_n$, or (2) for all $t \in \mathcal{T}_n$, for all $m \in \mathcal{M}_n$ in the "groups of time series" setting.

3 Discussion and Future Work

In this paper, we introduced a CI test and proved its theoretical justification in two different theoretical frameworks for nonstationary time series: (1) a new framework in Section 2, and (2) the well-studied special case of locally stationary time series in Section D. In practice, we suggest using rolling time-window approaches with widely-available learning algorithms (e.g. random forest, XGBoost, or LightGBM). We can then use the residuals from these regressions as the basis of a powerful CI test by using the bootstrap-based testing procedure from Subsection 2.4. We discussed how our testing framework can (1) be used to discover new forecasting signals, (2) serve as the basis of a causal discovery algorithm for nonstationary processes to infer causal graphs, and (3) gain power by leveraging the high-dimensionality of a "group of time series" such as a sensor network. In Section A, we show how our framework can be used with simultaneous testing procedures to discover short time-windows during which conditional independence holds or not for particular dimensions and time-offsets.

Our motivation for developing this CI test was to improve the practice of discovering new forecasting signals In particular, our original goal was to create a Granger causality testing framework that is appropriate for the complexities of real-world time series. Specifically, we wanted to avoid assuming that the time series is stationary, linear, and Gaussian. In our companion paper [WHR24], we apply our testing framework to determine whether viral infectivity data contains any auxiliary information about future case growth rates of COVID-19 data.

There are many promising avenues for future work. First, we will deal with the problem of *conditional mean independence testing*. As discussed in Lundborg et al. [Lun+22], the null hypothesis of conditional mean independence is untestable (without making further assumptions) because the smaller conditional independence null is untestable [SP20]. Consequently, we must restrict the null hypothesis. In future work, we will develop a conditional mean independence testing framework for high-dimensional nonstationary nonlinear time series.

Second, we will develop distribution-uniform extensions for many of the results about locally stationary time series introduced in Dahlhaus, Richter, and Wu [DRW19]. We plan to apply these new distribution-uniform results to more develop additional conditional independence independence tests for the locally stationary time series setting as discussed in Subsections D.8 and D.9 . In particular, these results will allow us to (1) test for conditional independence at particular rescaled times by utilizing kernel smoothing, and (2) infer expected conditional covariances curves which can be used for simultaneous testing of conditional independence.

Third, we plan to develop a unified framework for causal inference and causal discovery for high-dimensional nonstationary time series based on all of this work. The previously discussed framework can be adapted so that it can be used for inferring other functionals of interest in the emerging field of causal inference for time series [Sag+20; RGR22; Run+23b; Run+19b; Run18a; RS21; RS19; Bon+21]. However, this line of inquiry will require additional Donsker-type assumptions or novel developments in sample-splitting methodology.

References

[AD12] Alekh Agarwal and John C. Duchi. "The generalization ability of online algorithms for dependent data". In: *IEEE Transactions on Information Theory* 59.1 (2012), pp. 573–587.

[ALW13] Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. "Prediction of time series by statistical learning: general losses and fast rates". In: *Dependence Modeling* (2013).

[AM12] Pierre-Olivier Amblard and Olivier J. J. Michel. "The Relation between Granger Causality and Directed Information Theory: A Review". In: *Entropy* (2012).

[AB12] D. W. K. Andrews and P. J. Barwick. "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure". In: *Econometrica* (2012).

[AG09] D. W. K. Andrews and P. Guggenberger. "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities". In: *Econometrica* (2009).

[AS10] D. W. K. Andrews and G. Soares. "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection". In: *Econometrica* (2010).

[BW23] Lujia Bai and Weichi Wu. "Time-varying correlation network analysis of non-stationary multivariate time series with complex trends". 2023.

[BR23] Sumanta Basu and Suhasini Subba Rao. "Graphical models for nonstationary time series". In: *The Annals of Statistics* (2023).

[Bee21] Carina Beering. "A functional central limit theorem and its bootstrap analogue for locally stationary processes with application to independence testing". 2021.

[Ben+10] Shai Ben-David et al. "A theory of learning from different domains". In: *Machine learning* (2010).

[Bic+93] Peter J. Bickel et al. "Efficient and adaptive estimation for semiparametric models". In: *Annals of Statistics* (1993).

[Bon+21] Matteo Bonvini et al. "Causal inference in the time of Covid-19". 2021.

[BRT12] Taoufik Bouezmarni, Jeroen VK Rombouts, and Abderrahim Taamouti. "Nonparametric copula-based test for conditional independence with applications to Granger causality". In: *Journal of Business and Economic Statistics* (2012).

[Bre01] Leo Breiman. "Random forests". In: *Machine Learning* (2001).

[Bru22] Guy-Niklas Brunotte. "A test of independence under local stationarity based on the local characteristic function". 2022.

[Cha82] Gary Chamberlain. "The general equivalence of Granger and Sims causality". In: *Econometrica* (1982).

[CG16] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016).

[CCK13] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors". In: *The Annals of Statistics* 41.6 (2013), pp. 2786–2819.

[CPH22] Alexander Mangulad Christgau, Lasse Petersen, and Niels Richard Hansen. "Nonparametric Conditional Local Independence Testing". In: *arXiv preprint arXiv:2203.13559* (2022).

[Con12] Nonparametric Copula-Based Test for Conditional Independence with Applications to Granger Causality. "Taoufik Bouezmarni and Jeroen V.K. Rombouts and Abderrahim Taamouti". In: *Journal of Business and Economic Statistics* (2012).

[Cor+08] Corinna Cortes et al. "Sample selection bias correction theory". In: *International conference on algorithmic learning theory* (2008).

[Dah97] Rainer Dahlhaus. “Fitting time series models to nonstationary processes”. In: *The Annals of Statistics* (1997).

[Dah00] Rainer Dahlhaus. “Graphical interaction models for multivariate time series”. In: *Metrika* (2000).

[Dah12] Rainer Dahlhaus. “Locally stationary processes”. In: *Handbook of statistics* (2012).

[DRW19] Rainer Dahlhaus, Stefan Richter, and Wei Biao Wu. “Towards a general theory for nonlinear locally stationary processes”. In: *Bernoulli* (2019).

[Dau80] J. J. Daudin. “Partial association measures and an application to qualitative regression”. In: *Biometrika* 67.3 (1980), pp. 581–590.

[DN20] Richard A. Davis and Mikkel S. Nielsen. “Modeling of time series using random forests: Theoretical developments”. In: *Electronic Journal of Statistics* (2020).

[Daw79] Philip A. Dawid. “Conditional independence in statistical theory”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979), pp. 1–15.

[DT20] Philip A. Dawid and Ambuj Tewari. “On learnability under general stochastic processes”. 2020.

[Del20] Anne van Delft. “A note on quadratic forms of stationary functional time series under mild conditions”. In: *Stochastic Processes and their Applications* (2020).

[DD24] Anne van Delft and Holger Dette. “A general framework to quantify deviations from structural assumptions in the analysis of nonstationary function-valued processes”. In: *The Annals of Statistics* (2024).

[Don+23] Xinshuai Dong et al. “On the Three Demons in Causality in Finance: Time Resolution, Nonstationarity, and Latent Factors”. 2023.

[Eic12] Michael Eichler. “Graphical modelling of multivariate time series”. In: *Probability Theory Related Fields* (2012).

[FNS15] Seth R. Flaxman, Daniel B. Neill, and Alexander J. Smola. “Gaussian processes for independence tests with non-iid data in causal inference”. In: *ACM Transactions on Intelligent Systems and Technology* 7.2 (2015), pp. 1–23.

[FM82] J. P. Florens and M. Mouchart. “A note on noncausality”. In: *Econometrica* (1982).

[GWS23] Hanjia Gao, Runmin Wang, and Xiaofeng Shao. “Dimension-agnostic Change Point Detection”. 2023.

[GF12] Jelle J. Goeman and Livio Finos. “The inheritance procedure: multiple testing of tree-structured hypotheses”. In: *Statistical applications in genetics and molecular biology* (2012).

[GS10] Jelle J. Goeman and Aldo Solari. “The sequential rejection principle of familywise error control”. In: *The Annals of Statistics* (2010).

[Han21a] Steve Hanneke. “Learning whenever learning is possible: Universal learning under general stochastic processes”. In: *Journal of Machine Learning Research* (2021).

[Han21b] Steve Hanneke. “Learning whenever learning is possible: Universal learning under general stochastic processes”. In: *The Journal of Machine Learning Research* 22.1 (2021), pp. 5751–5866.

[HK24] Steve Hanneke and Samory Kpotufe. “A More Unified Theory of Transfer Learning”. 2024.

[HY19] Steve Hanneke and Liu Yang. “Statistical learning under nonstationary mixing processes”. In: *The 22nd International Conference on Artificial Intelligence and Statistics* 22.1 (2019), pp. 5751–5866.

[Hoc+23] Tom Hochsprung et al. “Increasing effect sizes of pairwise conditional independence tests between random vectors”. In: *Uncertainty in Artificial Intelligence* (2023).

[Hua+20] Biwei Huang et al. “Causal discovery from heterogeneous/nonstationary data”. In: *Journal of Machine Learning Research* (2020).

[Hua+06] Jiayuan Huang et al. "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems* (2006).

[IM04] G. W. Imbens and C. F. Manski. "Confidence Intervals for Partially Identified Parameters". In: *Econometrica* (2004).

[Kal97] Olav Kallenberg. *Foundations of modern probability*. Springer, 1997. URL: https://link.springer.com/book/10.1007/978-3-030-61871-1.

[KV02] Rajeeva L. Karandikar and Mathukumalli Vidyasagar. "Rates of uniform convergence of empirical means with mixing processes". In: *Statistics and probability letters* 58.3 (2002), pp. 297–307.

[Kas18] Maximilian Kasy. "Uniformity and the delta method". In: *Journal of Econometric Methods* (2018).

[Ke+17] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems* (2017).

[Ken22] Edward Kennedy. "Semiparametric doubly robust targeted double machine learning: a review". 2022.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[KKK22] Arun K. Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. "Post-selection inference". In: *Annual Review of Statistics and Its Application* (2022).

[KBW23] Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. "Median regularity and honest inference". In: *Biometrika* (2023).

[KKS23] Daisuke Kurisu, Kengo Kato, and Xiaofeng Shao. "Gaussian approximation and spatially dependent wild bootstrap for high-dimensional spatial data". In: *Journal of the American Statistical Association* (2023).

[KM17] Vitaly Kuznetsov and Mehryar Mohri. "Generalization bounds for non-stationary mixing processes". In: *Machine Learning* 106.1 (2017), pp. 93–117.

[KM14] Vitaly Kuznetsov and Mehryar Mohri. "Generalization bounds for time series prediction with non-stationary processes". In: *Algorithmic Learning Theory* (2014).

[KM15] Vitaly Kuznetsov and Mehryar Mohri. "Learning theory and algorithms for forecasting non-stationary time series". In: *Advances in neural information processing systems* 28 (2015).

[LLZ22] Jia Li, Zhipeng Liao, and Wenyu Zhou. "A general test for functional inequalities". 2022.

[Li89] Ker-Chau Li. "Honest confidence regions for nonparametric regression". In: *The Annals of Statistics* (1989).

[Li+11] Lingling Li et al. "Higher order inference on a treatment effect under low regularity conditions". In: *Statistics probability letters* (2011).

[Liu+23] Zhaolu Liu et al. "Kernel-based Joint Independence Tests for Multivariate Stationary and Nonstationary Time-Series". In: *arXiv* (2023).

[LKS05] Aurelie C. Lozano, Sanjeev Kulkarni, and Robert E. Schapire. "Convergence and consistency of regularized boosting algorithms with stationary b-mixing observations". In: *Advances in Neural Information Processing Systems* (2005).

[Lun+22] Anton Rask Lundborg et al. "The Projected Covariance Measure for assumption-lean variable significance testing". 2022.

[LSP22] Anton Rask Lundborg, Rajen D. Shah, and Jonas Peters. "Conditional independence testing in Hilbert spaces with applications to functional data analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.5 (2022), pp. 1821–1850.

[LW23] Tianpai Luo and Weichi Wu. "Simultaneous inference for monotone and smoothly time varying functions under complex temporal dynamics". 2023.

[MSA22] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. "M5 accuracy competition: Results, findings, and conclusions". In: *International Journal of Forecasting* (2022).

[MS19] Daniel Malinsky and Peter Spirtes. "Learning the structure of a nonstationary vector autoregression". In: *The 22nd International Conference on Artificial Intelligence and Statistics* (2019).

[Mei08] Nicolai Meinshausen. "Hierarchical testing of variable importance". In: *Biometrika* (2008).

[MS22] Fabian Mies and Ansgar Steland. "Sequential Gaussian approximation for nonstationary time series in high dimensions". In: *arXiv preprint arXiv:2203.03237* (2022).

[MK20] Mehryar Mohri and Vitaly Kuznetsov. "Discrepancy-based theory and algorithms for forecasting non-stationary time series". In: *Annals of Mathematics and Artificial Intelligence* 88.4 (2020), pp. 367–399.

[MH21] Pablo Montero-Manso and Rob J. Hyndman. "Principles and algorithms for forecasting groups of time series: Locality and globality". In: *International Journal of Forecasting* (2021).

[NBW21] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. "Minimax Optimal Conditional Independence Testing". In: *The Annals of Statistics* (2021).

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

[QKC15] Christopher J. Quinn, Negar Kiyavash, and Todd P. Coleman. "Directed Information Graphs". In: *IEEE Transactions on Information Theory* (2015).

[RS19] Ashesh Rambachan and Neil Shephard. "A nonparametric dynamic causal model for macroeconometrics". 2019.

[RS21] Ashesh Rambachan and Neil Shephard. "When do common time series estimands have nonparametric causal meaning". 2021.

[Reb80] Rolando Rebolledo. "Central limit theorems for local martingales". In: *Probability Theory and Related Fields* (1980).

[RCS21] Henry WJ Reeve, Timothy I. Cannings, and Richard J. Samworth. "Adaptive transfer learning". In: *The Annals of Statistics* (2021).

[RGR22] Nicolas-Domenic Reiter, Andreas Gerhardus, and Jakob Runge. "Causal inference for temporal patterns". 2022.

[RWG19] Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. "Bootstrapping and sample splitting for high-dimensional, assumption-lean inference". In: *The Annals of Statistics* (2019).

[Rob+08] James Robins et al. "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman* (2008).

[Rob+09] James Robins et al. "Semiparametric minimax rates". In: *Electronic journal of statistics* (2009).

[Rob+17] James M. Robins et al. "Minimax estimation of a functional on a structured high-dimensional model". In: *Annals of Statistics* (2017).

[Rob18] Whitney K. Neweyand James R. Robins. "Cross-fitting and fast remainder rates for semiparametric estimation". In: *Annals of Statistics* (2018).

[RS08] J. P. Romano and A. M. Shaikh. "Inference for Identifiable Parameters in Partially Identified Econometric Models". In: *Journal of Statistical Planning and Inference* (2008).

[RSW14] J. P. Romano, A. M. Shaikh, and M. Wolf. "A Practical Two-step Method for Testing Moment Inequalities". In: *Econometrica* (2014).

[RW05] Joseph P. Romano and Michael Wolf. "Exact and approximate stepdown methods for multiple hypothesis testing". In: *Journal of the American Statistical Association* (2005).

[Ros61] Murray Rosenblatt. “Independence and dependence”. In: *Proc. 4th Berkeley sympos. math. statist. and prob* (1961).

[Run18a] Jakob Runge. “Causal network reconstruction from time series: From theoretical assumptions to practical estimation”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* (2018).

[Run18b] Jakob Runge. “Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information”. In: *International Conference on Artificial Intelligence and Statistics* (2018).

[Run+23a] Jakob Runge et al. “Causal inference for time series”. In: *Nature Reviews Earth and Environment* (2023).

[Run+23b] Jakob Runge et al. “Causal inference for time series”. In: *Nature Reviews Earth Environment* (2023).

[Run+19a] Jakob Runge et al. “Detecting and quantifying causal associations in large nonlinear time series datasets”. In: *Science advances* (2019).

[Run+19b] Jakob Runge et al. “Detecting and quantifying causal associations in large nonlinear time series datasets”. In: *Science Advances* (2019).

[Run+19c] Jakob Runge et al. “Inferring causation from time series in Earth system sciences”. In: *Nature communications* 10.1 (2019).

[Sag+20] Elena Saggioro et al. “Reconstructing regime-dependent causal relationships from observational time series”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* (2020).

[SP11] Sohan Seth and Jose C. Principe. “Assessing Granger non-causality using nonparametric measure of conditional independence”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2011).

[SP20] Rajen D. Shah and Jonas Peters. “The hardness of conditional independence testing and the generalised covariance measure”. In: *Annals of Statistics* 48.3 (2020), pp. 1514–1538.

[SF22] Ali Shojaie and Emily B. Fox. “Granger causality: A review and recent advances”. In: *Annual Review of Statistics and Its Application* (2022).

[SGS00] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[SA09] Ingo Steinwart and Marian Anghel. “Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise”. In: *Annals of Statistics* (2009).

[SHS09] Ingo Steinwart, Don Hush, and Clint Scovel. “Learning from dependent observations”. In: *Journal of Multivariate Analysis* (2009).

[SW07] Liangjun Su and Halbert White. “A consistent characteristic function-based test for conditional independence”. In: *Journal of Econometrics* (2007).

[Tan+21] Alex Tank et al. “Neural Granger Causality”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[Tib+18] Ryan J Tibshirani et al. “Uniform asymptotic inference and the bootstrap after model selection”. In: *The Annals of Statistics* (2018).

[Vog12] Michael Vogt. “Nonparametric regression for locally stationary time series”. In: *Annals of Statistics* (2012), pp. 2601–2633.

[WNR23a] Jonas Wahl, Urmi Ninad, and Jakob Runge. “Foundations of Causal Discovery on Groups of Variables”. 2023.

[WNR23b] Jonas Wahl, Urmi Ninad, and Jakob Runge. “Vector causal inference between two groups of variables”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2023).

[WR23] Ian Waudby-Smith and Aaditya Ramdas. “Distribution-uniform anytime-valid inference”. 2023.

[Wau+21] Ian Waudby-Smith et al. “Time-uniform central limit theory and asymptotic confidence sequences”. In: *Annals of Statistics* (2021).

[WHR24] Michael Wieck-Sosa, Michel F. C. Haddad, and Aaditya Ramdas. “Epidemic Forecasting with City-Level Viral Infectivity Data”. 2024.

[Wie66] Norbert Wiener. *Nonlinear problems in random theory*. MIT Press, 1966.

[WLT20] Kam Chung Wong, Zifan Li, and Ambuj Tewari. “Lasso guarantees for beta-mixing heavy-tailed time series”. In: *Annals of Statistics* 48.2 (2020), pp. 1124–1142.

[Wu11] Wei Biao Wu. “Asymptotic theory for stationary processes”. In: *Statistics and its Interface* (2011).

[Wu05] Wei Biao Wu. “Nonlinear system theory: Another look at dependence”. In: *Proceedings of the National Academy of Sciences* 102.40 (2005), pp. 14150–14154.

[WX12] Wei Biao Wu and Han Xiao. “Covariance matrix estimation in time series”. In: *Handbook of Statistics* 30 (2012), pp. 187–209.

[YN21] Kashif Yousuf and Serena Ng. “Boosting high dimensional predictive regressions with time varying parameters”. In: *Journal of Econometrics* (2021).

[Yu94] Bin Yu. “Rates of convergence for empirical processes of stationary mixing sequences”. In: *The Annals of Probability* (1994), pp. 94–116.

[Zha+12] Kun Zhang et al. “Kernel-based conditional independence test and application in causal discovery”. In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* (2012).

[ZJ20] Lu Zhang and Lucas Janson. “Floodgate: inference for model-free variable importance”. 2020.

[ZW15] Ting Zhang and Wei Biao Wu. “Time-varying nonlinear regression models: nonparametric estimation and model selection”. In: *Annals of Statistics* (2015), pp. 741–768.

[ZS24] Yi Zhang and Xiaofeng Shao. “Another look at bandwidth-free inference: a sample splitting approach”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024).

[Zha+24] Yi Zhang et al. “Doubly Robust Conditional Independence Testing with Generative Neural Networks”. 2024.

[ZW09] Zhou Zhou and Wei Biao Wu. “Local linear quantile estimation for nonstationary time series”. In: *The Annals of Statistics* (2009).

A Simultaneous Testing

We will describe two approaches for simultaneous testing for conditional independence using the framework from Section 2. First, we introduce a stepdown procedure based on Romano and Wolf [RW05]. Second, we describe an inheritance procedure based on Goeman and Finos [GF12].

Let $W_n \in \mathbb{N}$ be the number of time-windows, and let $\mathcal{W}_n = \{1, \dots, W_n\}$ be indices for all of the time-window. For some $w \in \mathcal{W}_n$, let $\mathcal{T}_n^w \subset \mathcal{T}_n$ denote the time-window. The time-windows consist of possibly overlapping sequences consecutive of positive integers such that $\mathcal{T}_n = \bigcup_{w \in \mathcal{W}_n} \mathcal{T}_n^w$. Denote the window size by the cardinality $T_n^w = |\mathcal{T}_n^w|$ for some $w \in \mathcal{W}_n$. Mainly for the sake of convenience, we will require that the time-windows are the same length, i.e. $T_n^{w_1} = T_n^{w_2}$ for all $w_1, w_2 \in \mathcal{W}_n$. For $n \in \mathbb{N}$, we discuss how to simultaneously test for the hypotheses

$$H_{0,n,m}^{\mathrm{CI},w} : X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n^w$$

for *each* pair of dimension/time-offset indices and time-window numbers $(m, w) \in \mathcal{M}_n \times \mathcal{W}_n$.

We will require that the window sizes T_n^w grow with the sample size n at the same rate as T_n so that the long-run asymptotic arguments used in Section 2 directly apply. For instance, if we choose two time windows corresponding to the first and second halves of the time series, then $\mathcal{T}_n^1 = \{1, \dots, \lfloor T_n/2 \rfloor\}$ and $\mathcal{T}_n^2 = \{\lfloor T_n/2 \rfloor + 1, \dots, T_n\}$. There is a natural duality with this idea and the infill asymptotic framework of locally stationary time series described in Section D where time is rescaled to the unit interval. That is, the asymptotic justifications for the simultaneous testing procedure explained in this section only make sense when viewing the time-windows as fractions of the total sample size (i.e. so that we get more observations within each time-window). Note that these asymptotic considerations are typically not as relevant in practice when we are simply given a dataset with a finite number of observations.

There are two different alternative hypotheses $H_{1,n,m}^{\mathrm{CI},w}$ that can be used. We can always use

$$H_{1,n,m}^{\mathrm{CI},w} : X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for at least one } t \in \mathcal{T}_n^w$$

Similar to the discussion in Subsection 2.2, if domain knowledge suggests that the conditional independence relationships are stable within time-windows of length W_n, then we can replace the alternative hypothesis with

$$H_{1,n,m}^{\mathrm{CI},w} : X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n^w.$$

Several remarks are in order. First, note that the predictions and subsequent residuals are formed in the same way as for the test from Subsection 2.4 as explained in Subsection B.2. Thus, to be abundantly clear, choosing more windows does not affect the amount of data we have to estimate the time-varying regression functions. Second, the number of windows can grow with the sample size $n \in \mathbb{N}$. Third, note that there is a balance between choosing the time-windows $\mathcal{T}_n^w$ to be large enough so that each test has enough power to detect conditional dependence, but small enough so that each time-window is of scientific interest.

For each pair $(m, w) \in \mathcal{M}_n \times \mathcal{W}_n$, denote

$$\mathcal{P}_{0,n,m}^{\mathrm{CI},w} = \bigcap_{t \in \mathcal{T}_n^w} \mathcal{P}_{0,n,m,t}^{\mathrm{CI}}$$

where $\mathcal{P}_{0,n,m,t}^{\mathrm{CI}} \subset \mathcal{P}_n$ is a collection of distributions such that $X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n}$. That is, each $\mathcal{P}_{0,n,m}^{\mathrm{CI},w}$ is a collection of distributions for which conditional independence holds for all times $t \in \mathcal{T}_n^w$ in window $w \in \mathcal{W}_n$ for index $m \in \mathcal{M}_n$.

Let $\mathcal{M}_n^* \times \mathcal{W}_n^* \subset \mathcal{M}_n \times \mathcal{W}_n$ be a subset of pairs of indices and time windows. Consider the intersection of null hypotheses of the form

$$\mathcal{P}_{0,n}^{\mathrm{CI}}(\mathcal{M}_n^* \times \mathcal{W}_n^*) = \bigcap_{(m,w) \in \mathcal{M}_n^* \times \mathcal{W}_n^*} \mathcal{P}_{0,n,m}^{\mathrm{CI},w}.$$

In words, $\mathcal{P}_{0,n}^{\mathrm{CI}}(\mathcal{M}_n^* \times \mathcal{W}_n^*)$ is a collection of distributions such that conditional independence always holds during *certain* time windows for *specific* indices.

Inspired by Chernozhukov, Chetverikov, and Kato [CCK13] and Kurisu, Kato, and Shao [KKS23], we combine our bootstrap procedure with the stepdown procedure for strong FWER control from

Romano and Wolf [RW05]. The main idea is that we treat each time-window as a separate index, which is equivalent to considering a time-lead equal to the start of each window as another index. Hence, instead of considering the max over the times $\mathcal{T}_n$ as in Subsection 2.4, we take the max over the times $1, \ldots, T_n^w$ corresponding to the window size.

For each pair $(m, w) \in \mathcal{M}_n \times \mathcal{W}_n$ define

$$S_{n,m,w}(\hat{\boldsymbol{R}}_n) = \max_{s=1,\ldots,T_n^w} \left| \frac{1}{\sqrt{T_n}} \sum_{t \leq s} \hat{R}_{t,n,m} \right|.$$

At the first step $\ell = 1$, set $\mathcal{M}_n^{(\ell)} = \mathcal{M}_n$ and $\mathcal{W}_n^{(\ell)} = \mathcal{W}_n$. At each step ℓ, we reject any of the hypotheses $H_{0,n,m}^{\mathrm{CI},w}$ in which the corresponding $S_{n,m,w}(\hat{\boldsymbol{R}}_n)$ is greater than the $(1-\alpha)$ quantile of $S_{n,m,w,\infty}^{(\ell)}(\hat{\boldsymbol{R}}_n)$ where

$$S_{n,m,w,\infty}^{(\ell)}(\hat{\boldsymbol{R}}_n) = \max_{(m,w) \in \mathcal{M}_n^{(\ell)} \times \mathcal{W}_n^{(\ell)}} \max_{s=1,\ldots,T_n^w} \left| \frac{1}{\sqrt{T_n}} \sum_{t \leq s} \hat{R}_{t,n,m} \right|.$$

If we did not reject any hypotheses, stop. If we did reject at least one hypothesis, continue. At the next step ℓ, denote $(\mathcal{M}_n \times \mathcal{W}_n)^{(\ell)} \subset \mathcal{M}_n \times \mathcal{W}_n$ by the subset of pairs of indices and window numbers in which we did not reject the corresponding hypotheses $H_{0,n,m}^{\mathrm{CI},w}$. Repeat the previous bootstrap procedure until no additional hypotheses are rejected at a given step, or until all hypotheses have been rejected. Let $(\mathcal{M}_n \times \mathcal{W}_n)^{\mathrm{reject}} \subset \mathcal{M}_n \times \mathcal{W}_n$ be the subset of pairs indices and window numbers in which the corresponding hypotheses $H_{0,n,m}^{\mathrm{CI},w}$ were rejected. Similarly, let $(\mathcal{M}_n \times \mathcal{W}_n)^{\mathrm{retain}} \subset \mathcal{M}_n \times \mathcal{W}_n$ be the subset of pairs of indices and window numbers in which the corresponding hypotheses $H_{0,n,m}^{\mathrm{CI},w}$ were not rejected.

The induced simultaneous testing procedure will have the following uniformly asymptotic FWER control

$$\limsup_{n \to \infty} \sup_{\mathcal{M}_n^* \times \mathcal{W}_n^* \subset \mathcal{M}_n \times \mathcal{W}_n} \sup_{P \in \mathcal{P}_{0,n}^*(\mathcal{M}_n^* \times \mathcal{W}_n^*)} \mathbb{P}_P \left(\bigcup_{(m,w) \in \mathcal{M}_n^* \times \mathcal{W}_n^*} \{\text{reject } H_{0,n,m}^{\mathrm{CI},w}\} \right) \leq \alpha,$$

for some $\alpha \in (0, 1)$ if Assumptions 1, 3, 2, 4, 5, 6, 7, 8 all hold for the collection of distributions $\mathcal{P}_{0,n}^*(\mathcal{M}_n^* \times \mathcal{W}_n^*) \subset \mathcal{P}_{0,n}^{\mathrm{CI}}(\mathcal{M}_n^* \times \mathcal{W}_n^*)$ and the predictors.

Let us state what the conclusion of our simultaneous testing procedure would be. For each particular pair $(m, w) \in (\mathcal{M}_n \times \mathcal{W}_n)^{\mathrm{retain}}$, we fail to reject the null hypothesis that conditional independence holds at all times $t \in \mathcal{T}_n^w$ for window number w for index m. Analogously, for each particular pair $(m, w) \in (\mathcal{M}_n \times \mathcal{W}_n)^{\mathrm{reject}}$, we reject the null hypothesis that conditional independence holds at all times $t \in \mathcal{T}_n^w$ for the window number and index pair (m, w).

Next, we informally describe how our test can be used with the inheritance procedure from Goeman and Finos [GF12]. The previously discussed stepdown procedure requires users to preselect the time-windows and to separately test for each index m. This approach may be satisfactory in many settings. However, some problems can arise in noisy, high-dimensional settings. We will focus on just two. First, it is challenging for a practitioner to know how to choose the time-windows to be large enough so that the test has enough power, but small enough so the conclusions are scientifically interesting. Second, it may be difficult to test for conditional independence for all signals at once due to the inherent multiplicity of the problem in high dimensions. We propose using the inheritance procedure of Goeman and Finos [GF12], which is based on and the hierarchical testing procedure from Meinshausen [Mei08] and the sequential rejection principle from Goeman and Solari [GS10].

To deal with first problem of selecting windows, we introduce a procedure for automatically selecting time-windows at appropriate temporal resolutions while controlling the familywise error rate. The main idea is to utilize the temporal hierarchy of the hypotheses when choosing the time-windows. We will start by testing whether conditional independence holds at all times or not at the *coarsest* resolution - that is, by choosing the first time-window to be $\mathcal{T}_n$ itself. If we detect conditional dependencies at some point in time during this time-window, we will then split the time-window and conduct tests on the first and second halves of the time-window. The procedure will continue attempting to identify finer time-windows during which conditional independence does not always hold by successively splitting time-windows.

To deal with second problem of testing multiplicity, we again propose starting at the coarsest resolution of the signals. The procedure will utilize a given hierarchical grouping structure of signals for which conditional independence holds during a particular time-window. The procedure will continue trying to identify smaller sub-groups in the given hierarchy of signals down to the level of individual signals. This solution naturally encourages using groups of time series that are highly correlated and related to the same underlying phenomena. For instance, consider the setting in which we have a hierarchical grouping of cities at the country, state, and county levels. We might not care about identifying the *particular* city for which conditional dependencies were discovered, but only that there were *some* conditional dependencies discovered by the group of time series. Hence, ad hoc dimensionality reduction and spatial averaging techniques to deal with the problem of high-dimensionality can be completely avoided. We envision that the hierarchy of groups of signals will be given by the expert based on domain knowledge, although hierarchical clustering approaches for nonstationary time series could also be explored. See Montero-Manso and Hyndman [MH21] and Wahl, Ninad, and Runge [WNR23a; WNR23b] for more information about groups of time series and frameworks for CI testing, causal discovery, and causal inference within this paradigm.

B Theoretical Framework

B.1 The conditional covariance process

In this subsection, we will introduce the conditional covariance process of the high-dimensional nonstationary time series.

Assumption 3 (Causal representations of the error processes)**.** *We assume that the error processes from Subsection 2.3 have the following causal representations. For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, we can represent the error processes as*

$$\varepsilon_{P,t,n,i,a} = G^{\varepsilon}_{P,t,n,i,a}\left(\mathcal{H}^{\varepsilon}_{t,a}\right) = X_{t,n,i,a} - \mathbb{E}_P(X_{t,n,i,a}|\boldsymbol{Z}_{t,n}),$$

$$\xi_{P,t,n,j,b} = G^{\xi}_{P,t,n,j,b}\left(\mathcal{H}^{\xi}_{t,b}\right) = Y_{t,n,j,b} - \mathbb{E}_P(Y_{t,n,j,b}|\boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}^{\varepsilon}_{t,a} = (\eta^{\varepsilon}_{t,a}, \eta^{\varepsilon}_{t,a-1}, \ldots)$, $\mathcal{H}^{\xi}_{t,b} = (\eta^{\xi}_{t,b}, \eta^{\xi}_{t,b-1}, \ldots)$ and $(\eta^{\varepsilon}_{t,a})_{t\in\mathbb{Z}}$, $(\eta^{\xi}_{t,b})_{t\in\mathbb{Z}}$ are sequences of iid random elements. In particular, $\eta^{\varepsilon}_{t,a} = (\eta^{X}_{t+a}, \eta^{Z}_{t})'$, $\eta^{\xi}_{t,b} = (\eta^{Y}_{t+b}, \eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$ so that the error processes each depend on the inputs for the covariate processes and their respective response process. For each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$, we have that $\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\mathcal{H}^{Z}_{t}) = 0$ and $\mathbb{E}_P(\xi_{P,t,n,j,b}|\mathcal{H}^{Z}_{t}) = 0$, and also that $G^{\varepsilon}_{P,t,n,i,a}(\cdot)$, $G^{\xi}_{P,t,n,j,b}(\cdot)$ are measurable functions such that $G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{s,a})$, $G^{\xi}_{P,t,n,j,b}(\mathcal{H}^{\xi}_{s,b})$ are well-defined random variables for each $s \in \mathbb{Z}$ and $(G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{s,a}))_{s\in\mathbb{Z}}$, $(G^{\xi}_{P,t,n,j,b}(\mathcal{H}^{\xi}_{s,b}))_{s\in\mathbb{Z}}$ are stationary ergodic time series.

In light of the causal representations of the error processes, we have the following causal representation of the high-dimensional nonstationary vector-valued error processes

$$\boldsymbol{\varepsilon}_{P,t,n} = \boldsymbol{G}^{\boldsymbol{\varepsilon}}_{P,t,n}\left(\mathcal{H}^{\boldsymbol{\varepsilon}}_{t}\right) = \left(G^{\varepsilon}_{P,t,n,i,a}\left(\mathcal{H}^{\varepsilon}_{t,a}\right)\right)_{i\in[d_X],a\in A_i},$$

where $\mathcal{H}^{\boldsymbol{\varepsilon}}_{t} = (\eta^{\boldsymbol{\varepsilon}}_{t}, \eta^{\boldsymbol{\varepsilon}}_{t-1}, \ldots)'$ and $\eta^{\boldsymbol{\varepsilon}}_{t} = (\eta^{X}_{t+a_{\max}}, \eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$. Similarly, for $\boldsymbol{\xi}_{P,t,n}$ we write

$$\boldsymbol{\xi}_{P,t,n} = \boldsymbol{G}^{\boldsymbol{\xi}}_{P,t,n}(\mathcal{H}^{\boldsymbol{\xi}}_{t}) = (G^{\xi}_{P,t,n,j,b}(\mathcal{H}^{\xi}_{t,b}))_{j\in[d_Y],b\in B_j},$$

where $\mathcal{H}^{\boldsymbol{\xi}}_{t} = (\eta^{\boldsymbol{\xi}}_{t}, \eta^{\boldsymbol{\xi}}_{t-1}, \ldots)$ and $\eta^{\boldsymbol{\xi}}_{t} = (\eta^{Y}_{t+b_{\max}}, \eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$.

Moreover, for each $m \in \mathcal{M}_n$ the process of error products can be represented as

$$R_{P,t,n,m} = G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{t,m}) = G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{t,a})G^{\xi}_{P,t,n,j,b}(\mathcal{H}^{\xi}_{t,b}),$$

where $\mathcal{H}^{R}_{t,m} = (\eta^{R}_{t,m}, \eta^{R}_{t-1,m}, \ldots)$ and $(\eta^{R}_{t,m})_{t\in\mathbb{Z}}$ is a sequence of iid random elements with $\eta^{R}_{t,m} = (\eta^{X}_{t+a}, \eta^{Y}_{t+b}, \eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$.

For each $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $n \in \mathbb{N}$, and $t \in \mathcal{T}_n$ we have the following causal representation of the high-dimensional nonstationary $\mathbb{R}^{M_n}$-valued process of all the products of errors $\boldsymbol{R}_{P,t,n}$ by

$$\boldsymbol{R}_{P,t,n} = \boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}\left(\mathcal{H}^{\boldsymbol{R}}_{t}\right) = \left(G^{R}_{P,t,n,m}\left(\mathcal{H}^{R}_{t,m}\right)\right)_{m\in\mathcal{M}_n}$$

where $\mathcal{H}_t^{\boldsymbol{R}} = (\eta_t^{\boldsymbol{R}}, \eta_{t-1}^{\boldsymbol{R}}, \ldots)$ and $\eta_t^{\boldsymbol{R}} = (\eta_{t+a_{\max}}^X, \eta_{t+b_{\max}}^Y, \eta_t^Z)'$ for each $t \in \mathbb{Z}$. As with the causal representations of $\boldsymbol{\varepsilon}_{P,t,n}$ and $\boldsymbol{\xi}_{P,t,n}$, for a fixed $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, and $n \in \mathbb{N}$ we have that $\boldsymbol{G}_{P,t,n}^{\boldsymbol{R}}(\mathcal{H}_s^{\boldsymbol{R}})$ is a well-defined high-dimensional random vector for each $s \in \mathbb{Z}$ and $(\boldsymbol{G}_{P,t,n}^{\boldsymbol{R}}(\mathcal{H}_s^{\boldsymbol{R}}))_{s\in\mathbb{Z}}$ is a high-dimensional stationary ergodic $\mathbb{R}^{M_n}$-valued time series.

Next, we define the local long-run covariance matrices and variances of the high-dimensional process of error products.

Definition 1 (Local long-run covariance matrices and variances of process of error products)**.** *For each $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $n \in \mathbb{N}$, define the long-run covariance matrix of the $\mathbb{R}^{M_n}$-valued stationary process $(\boldsymbol{G}_{P,t,n}^{\boldsymbol{R}}(\mathcal{H}_t^{\boldsymbol{R}}))_{t\in\mathbb{Z}}$ by*

$$\boldsymbol{\Sigma}_{P,t,n}^{\boldsymbol{R}} = \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\boldsymbol{G}_{P,t,n}^{\boldsymbol{R}}(\mathcal{H}_0^{\boldsymbol{R}}), \boldsymbol{G}_{P,t,n}^{\boldsymbol{R}}(\mathcal{H}_h^{\boldsymbol{R}})).$$

Similarly, for each $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $n \in \mathbb{N}$, and $m \in \mathcal{M}_n$, denote the long-run variance of the $\mathbb{R}$-valued stationary process $(G_{P,t,n,m}^R(\mathcal{H}_{t,m}^R))_{t\in\mathbb{Z}}$ as

$$\Sigma_{P,t,n,m}^R = \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(G_{P,t,n,m}^R(\mathcal{H}_{0,m}^R), G_{P,t,n,m}^R(\mathcal{H}_{h,m}^R)).$$

B.2 Prediction processes and adaptive learning algorithms

In this subsection, we consider the causal representation of the predictions and the learning algorithms used to construct the predictors. In Delft [Del20] and Delft and Dette [DD24] introduced the idea of causal representations for functional locally stationary time series. First, we introduce the causal representations of adaptive statistical learning algorithms.

Assumption 4 (Causal representations of adaptive statistical learning algorithms)**.** *For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$ let $\mathbb{M}_t(\mathcal{Y}, \mathcal{Z}_n) \subseteq \mathcal{Y}^{\mathcal{Z}_n}$ and $\mathbb{M}_t(\mathcal{X}, \mathcal{Z}_n) \subseteq \mathcal{X}^{\mathcal{Z}_n}$, where $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{Z}_n = \mathbb{R}^{\boldsymbol{d_Z}}$ where $\boldsymbol{d_Z}$ can grow with n. For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$, let $\eta_{t,n,i,a}^{algo}$, $\eta_{t,n,j,b}^{algo}$ be random elements which encode the (possible) stochasticity of the adaptive statistical learning algorithms. If the learning algorithms are not stochastic, then $\eta_{t,n,i,a}^{algo}$, $\eta_{t,n,j,b}^{algo}$ can be ignored without loss of generality. Denote the data used to construct $\hat{f}_{t,n,i,a}$ by $\mathcal{D}_{t,n,i,a}^{\hat{f}} = (X_{s,n,i,a}, \boldsymbol{Z}_{s,n})_{s=\mathbb{T}_n^- - \lambda_n}^t$ and similarly let $\mathcal{D}_{t,n,j,b}^{\hat{g}} = (Y_{s,n,j,b}, \boldsymbol{Z}_{s,n})_{s=\mathbb{T}_n^- - \lambda_n}^t$ be the data for $\hat{g}_{t,n,j,b}$, where $\lambda_n \geq 0$ is the optional late-starting parameter from Subsection 2.1.*

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, we assume that each of the (potentially stochastic) adaptive statistical learning algorithms $\mathcal{A}_{n,i,a}^{\hat{f}}$, $\mathcal{A}_{n,j,b}^{\hat{g}}$ consists of a sequence of measurable functions $\mathcal{A}_{n,i,a}^{\hat{f}} = (\mathcal{A}_{t,n,i,a}^{\hat{f}})_{t\in\mathcal{T}_n}$, $\mathcal{A}_{n,j,b}^{\hat{g}} = (\mathcal{A}_{t,n,j,b}^{\hat{g}})_{t\in\mathcal{T}_n}$, where for each $t \in \mathcal{T}_n$ we have

$$\mathcal{A}_{t,n,i,a}^{\hat{f}} : (\mathcal{X} \times \mathcal{Z}_n)^{t-\mathbb{T}_n^- + \lambda_n + 1} \times [0,1] \to \mathbb{M}_t(\mathcal{X}, \mathcal{Z}_n)$$

$$\mathcal{D}_{t,n,i,a}^{\hat{f}} \times \eta_{t,n,i,a}^{algo} \mapsto \hat{f}_{t,n,i,a},$$

and

$$\mathcal{A}_{t,n,j,b}^{\hat{g}} : (\mathcal{Y} \times \mathcal{Z}_n)^{t-\mathbb{T}_n^- + \lambda_n + 1} \times [0,1] \to \mathbb{M}_t(\mathcal{Y}, \mathcal{Z}_n)$$

$$\mathcal{D}_{t,n,j,b}^{\hat{g}} \times \eta_{t,n,j,b}^{algo} \mapsto \hat{g}_{t,n,j,b},$$

so that the adaptive statistical learning algorithms have the causal representations

$$\mathcal{A}_{t,n,i,a}^{\hat{f}} = G_{t,n,i,a}^{\mathcal{A}^{\hat{f}}}\left(\mathcal{H}_{t,a}^{\mathcal{A}^{\hat{f}}}\right),$$

$$\mathcal{A}_{t,n,j,b}^{\hat{g}} = G_{t,n,j,b}^{\mathcal{A}^{\hat{g}}}\left(\mathcal{H}_{t,b}^{\mathcal{A}^{\hat{g}}}\right).$$

$G_{t,n,i,a}^{\mathcal{A}^{\hat{f}}}(\cdot)$, $G_{t,n,j,b}^{\mathcal{A}^{\hat{g}}}(\cdot)$ are measurable functions such that $G_{t,n,i,a}^{\mathcal{A}^{\hat{f}}}\left(\mathcal{H}_{s,a}^{\mathcal{A}^{\hat{f}}}\right)$, $G_{t,n,j,b}^{\mathcal{A}^{\hat{g}}}\left(\mathcal{H}_{s,b}^{\mathcal{A}^{\hat{g}}}\right)$ are well-defined function-valued random variables for each $s \in \mathbb{Z}$ and $(G_{t,n,i,a}^{\mathcal{A}^{\hat{f}}}\left(\mathcal{H}_{s,a}^{\mathcal{A}^{\hat{f}}}\right))_{s\in\mathbb{Z}}$, $(G_{t,n,j,b}^{\mathcal{A}^{\hat{g}}}\left(\mathcal{H}_{s,b}^{\mathcal{A}^{\hat{g}}}\right))_{s\in\mathbb{Z}}$

are stationary functional time series. The input sequences are $\mathcal{H}_{t,a}^{\mathcal{A}^{\hat{f}}} = (\eta_{t,a}^{\mathcal{A}^{\hat{f}}}, \eta_{t-1,a}^{\mathcal{A}^{\hat{f}}}, \ldots)$, $\mathcal{H}_{t,b}^{\mathcal{A}^{\hat{g}}} = (\eta_{t,b}^{\mathcal{A}^{\hat{g}}}, \eta_{t-1,b}^{\mathcal{A}^{\hat{g}}}, \ldots)$ *where* $(\eta_{t,a}^{\mathcal{A}^{\hat{f}}})_{t\in\mathbb{Z}}$, $(\eta_{t,b}^{\mathcal{A}^{\hat{g}}})_{t\in\mathbb{Z}}$ *are sequences of iid random elements with* $\eta_{t,a}^{\mathcal{A}^{\hat{f}}} = (\eta_{t+a}^{X}, \eta_t^{Z})'$ *and* $\eta_{t,b}^{\mathcal{A}^{\hat{g}}} = (\eta_{t+b}^{Y}, \eta_t^{Z})'$.

We have the following causal representations for all dimensions and time-offsets of the statistical learning algorithms

$$\boldsymbol{\mathcal{A}}_{t,n}^{\hat{\boldsymbol{f}}} = \boldsymbol{G}_{t,n}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}\left(\mathcal{H}_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}\right) = (\mathcal{A}_{t,n,i,a}^{\hat{f}})_{i\in[d_X],a\in A_i},$$

$$\boldsymbol{\mathcal{A}}_{t,n}^{\hat{\boldsymbol{g}}} = \boldsymbol{G}_{t,n}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}\left(\mathcal{H}_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}\right) = (\mathcal{A}_{t,n,j,b}^{\hat{g}})_{j\in[d_Y],b\in B_j},$$

where $\mathcal{H}_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}} = (\eta_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}, \eta_{t-1}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}, \ldots)$ with $\eta_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}} = (\eta_{t+a_{\max}}^{X}, \eta_t^{Z})'$ and $\mathcal{H}_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}} = (\eta_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}, \eta_{t-1}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}, \ldots)$ with $\eta_t^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}} = (\eta_{t+b_{\max}}^{Y}, \eta_t^{Z})'$. The adaptive statistical learning algorithm $\mathcal{A}_{t,n,i,a}^{\hat{f}}$ used to construct the predictor $\hat{f}_{t,n,i,a}$ at time t only uses the covariates up to time t and the prediction target up to time $t+a$. Similarly, the algorithm $\mathcal{A}_{t,n,j,b}^{\hat{g}}$ used to form the predictor $\hat{g}_{t,n,j,b}$ at time t only uses the covariates up to time t and the prediction target up to time $t+b$. Note that we have suppressed the dependence of the predictors on $\eta_{t,n,i,a}^{\text{algo}}$, $\eta_{t,n,j,b}^{\text{algo}}$ for the sake of notational simplicity.

In view of the previous assumption, we discuss the causal representations for the predictions and prediction errors.

Assumption 5 (Causal representations for predictions and prediction errors)**.** *We assume that the statistical learning algorithms are measurable such that we can represent the predictions for each* $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$ *as*

$$\hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}) = G_{t,n,i,a}^{\hat{f}}\left(\mathcal{H}_{t,a}^{\hat{f}}\right) = [\mathcal{A}_{t,n,i,a}^{\hat{f}}(\mathcal{D}_{t,n,i,a}^{\hat{f}}, \eta_{t,n,i,a}^{algo})](\boldsymbol{Z}_{t,n}),$$

$$\hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}) = G_{t,n,j,b}^{\hat{g}}\left(\mathcal{H}_{t,b}^{\hat{g}}\right) = [\mathcal{A}_{t,n,j,b}^{\hat{g}}(\mathcal{D}_{t,n,j,b}^{\hat{g}}, \eta_{t,n,j,b}^{algo})](\boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}_{t,a}^{\hat{f}} = (\eta_{t,a}^{\hat{f}}, \eta_{t-1,a}^{\hat{f}}, \ldots)$ *with* $\eta_{t,a}^{\hat{f}} = (\eta_{t+a}^{X}, \eta_t^{Z})'$ *and* $\mathcal{H}_{t,b}^{\hat{g}} = (\eta_{t,b}^{\hat{g}}, \eta_{t-1,b}^{\hat{g}}, \ldots)$ *with* $\eta_{t,b}^{\hat{g}} = (\eta_{t+b}^{Y}, \eta_t^{Z})'$, *and that the prediction errors can be represented as*

$$\hat{w}_{P,t,n,i,a}^{f}(\boldsymbol{Z}_{t,n}) = G_{P,t,n,i,a}^{\hat{w}^f}\left(\mathcal{H}_{t,a}^{\hat{w}^f}\right) = f_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}) - \hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}),$$

$$\hat{w}_{P,t,n,j,b}^{g}(\boldsymbol{Z}_{t,n}) = G_{P,t,n,j,b}^{\hat{w}^g}\left(\mathcal{H}_{t,b}^{\hat{w}^g}\right) = g_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}) - \hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}_{t,a}^{\hat{w}^f} = (\eta_{t,a}^{\hat{w}^f}, \eta_{t-1,a}^{\hat{w}^f}, \ldots)$ *with* $\eta_{t,a}^{\hat{w}^f} = (\eta_{t+a}^{X}, \eta_t^{Z})'$ *and* $\mathcal{H}_{t,b}^{\hat{w}^g} = (\eta_{t,b}^{\hat{w}^g}, \eta_{t-1,b}^{\hat{w}^g}, \ldots)$ *with* $\eta_{t,b}^{\hat{w}^g} = (\eta_{t+b}^{Y}, \eta_t^{Z})'$.

As usual, $G_{t,n,i,a}^{\hat{f}}(\cdot)$, $G_{t,n,j,b}^{\hat{g}}(\cdot)$ *and* $G_{P,t,n,i,a}^{\hat{w}^f}(\cdot)$, $G_{P,t,n,j,b}^{\hat{w}^g}(\cdot)$ *are measurable functions such that* $G_{t,n,i,a}^{\hat{f}}\left(\mathcal{H}_{s,a}^{\hat{f}}\right)$, $G_{t,n,j,b}^{\hat{g}}\left(\mathcal{H}_{s,b}^{\hat{g}}\right)$ *and* $G_{P,t,n,i,a}^{\hat{w}^f}\left(\mathcal{H}_{s,a}^{\hat{w}^f}\right)$, $G_{P,t,n,j,b}^{\hat{w}^g}\left(\mathcal{H}_{s,b}^{\hat{w}^g}\right)$ *are well-defined real-valued random variables for each* $s \in \mathbb{Z}$ *and* $(G_{t,n,i,a}^{\hat{f}}\left(\mathcal{H}_{s,a}^{\hat{f}}\right))_{s\in\mathbb{Z}}$, $(G_{t,n,j,b}^{\hat{g}}\left(\mathcal{H}_{s,b}^{\hat{g}}\right))_{s\in\mathbb{Z}}$ *and* $\left(G_{P,t,n,i,a}^{\hat{w}^f}\left(\mathcal{H}_{s,a}^{\hat{w}^f}\right)\right)_{s\in\mathbb{Z}}$, $\left(G_{P,t,n,j,b}^{\hat{w}^g}\left(\mathcal{H}_{s,b}^{\hat{w}^g}\right)\right)_{s\in\mathbb{Z}}$ *are real-valued stationary ergodic time series.*

Then we have the following causal representation for all dimensions and time-offsets of the prediction errors

$$\hat{\boldsymbol{w}}_{P,t,n}^{\boldsymbol{f}}(\boldsymbol{Z}_{t,n}) = \boldsymbol{G}_{P,t,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}\left(\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}\right) = (\hat{w}_{P,t,n,i,a}^{f}(\boldsymbol{Z}_{t,n}))_{i\in[d_X],a\in A_i},$$

$$\hat{\boldsymbol{w}}_{P,t,n}^{\boldsymbol{g}}(\boldsymbol{Z}_{t,n}) = \boldsymbol{G}_{P,t,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}\left(\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}\right) = (\hat{w}_{P,t,n,j,b}^{g}(\boldsymbol{Z}_{t,n}))_{j\in[d_Y],b\in B_j},$$

where $\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}} = (\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}, \eta_{t-1}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}, \ldots)$ with $\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}} = (\eta_{t+a_{\max}}^{X}, \eta_t^{Z})'$ and $\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}} = (\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}, \eta_{t-1}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}, \ldots)$ with $\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}} = (\eta_{t+b_{\max}}^{Y}, \eta_t^{Z})'$. Also, we can represent the products of the errors and prediction errors as

$$\hat{w}_{P,t,n,m}^{g,\varepsilon}(\boldsymbol{Z}_{t,n}) = G_{P,t,n,m}^{\hat{w}^{g,\varepsilon}}\left(\mathcal{H}_{t,m}^{\hat{w}^{g,\varepsilon}}\right) = \hat{w}_{P,t,n,j,b}^{g}(\boldsymbol{Z}_{t,n})\varepsilon_{P,t,n,i,a},$$

$$\hat{w}_{P,t,n,m}^{f,\xi}(\boldsymbol{Z}_{t,n}) = G_{P,t,n,m}^{\hat{w}^{f,\xi}}\left(\mathcal{H}_{t,m}^{\hat{w}^{f,\xi}}\right) = \hat{w}_{P,t,n,i,a}^{f}(\boldsymbol{Z}_{t,n})\xi_{P,t,n,j,b},$$

where $\mathcal{H}^{\hat{w}^{g,\varepsilon}}_{t,m} = (\eta^{\hat{w}^{g,\varepsilon}}_{t,m}, \eta^{\hat{w}^{g,\varepsilon}}_{t-1,m}, \ldots)$ with $\eta^{\hat{w}^{g,\varepsilon}}_{t,m} = (\eta^X_{t+a}, \eta^Y_{t+b}, \eta^Z_t)'$ and $\mathcal{H}^{\hat{w}^{f,\xi}}_{t,m} = (\eta^{\hat{w}^{f,\xi}}_{t,m}, \eta^{\hat{w}^{f,\xi}}_{t-1,m}, \ldots)$ with $\eta^{\hat{w}^{f,\xi}}_{t,m} = (\eta^X_{t+a}, \eta^Y_{t+b}, \eta^Z_t)'$. Putting it all together, we have the following causal representation for all dimensions and time-offsets of the products of the errors and prediction errors

$$\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n}(\boldsymbol{Z}_{t,n}) = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_t\right) = (\hat{w}^{g,\varepsilon}_{P,t,n,m}(\boldsymbol{Z}_{t,n}))_{m\in\mathcal{M}_n},$$

$$\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,t,n}(\boldsymbol{Z}_{t,n}) = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_t\right) = (\hat{w}^{f,\xi}_{P,t,n,m}(\boldsymbol{Z}_{t,n}))_{m\in\mathcal{M}_n},$$

where $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_t = (\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_t, \eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t-1}, \ldots)$ with $\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_t = (\eta^X_{t+a_{\max}}, \eta^Y_{t+b_{\max}}, \eta^Z_t)'$ and $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_t = (\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_t, \eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{t-1}, \ldots)$ with $\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_t = (\eta^X_{t+a_{\max}}, \eta^Y_{t+b_{\max}}, \eta^Z_t)'$. We emphasize that $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}(\cdot)$, $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}(\cdot)$ are measurable functions such that $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_s\right)$, $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_s\right)$ are well-defined high-dimensional random vectors for each $s \in \mathbb{Z}$ and $(\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_s\right))_{s\in\mathbb{Z}}$, $(\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}\left(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_s\right))_{s\in\mathbb{Z}}$ are high-dimensional stationary ergodic time series.

B.3 Distribution-uniform functional dependence measure and control of nonstationarity

In this subsection, we discuss regularity conditions for the assumed causal representations introduced previously. The following assumptions build on the framework for high-dimensional nonlinear nonstationary time series from Mies and Steland [MS22] so that the temporal dependence and nonstationarity can be controlled uniformly over a collection of distributions $\mathcal{P}_n$. We consider slightly different measures of dependence than usual, but they are all based on the foundational work of Wu [Wu05].

Denote the set of well-defined tuples of error processes, dimensions, and time-offsets by

$$\mathbb{E} = \{(\varepsilon, i, a) : i \in [d_X], a \in A_i\} \cup \{(\xi, j, b) : j \in [d_Y], b \in B_j\},$$

so that we may write $(e, l, d) \in \mathbb{E}$ to refer to any such combination.

First, we introduce the framework for quantifying temporal dependence via the functional dependence measure of Wu [Wu05].

Definition 2 (Functional dependence measure)**.** *For any tuple $(e, l, d) \in \mathbb{E}$ corresponding to a well-defined combination of an error process, dimension, time-offset, let $(\tilde{\eta}^e_{t,d})_{t\in\mathbb{Z}}$ be an iid copy of $(\eta^e_{t,d})_{t\in\mathbb{Z}}$. Define*

$$\tilde{\mathcal{H}}^e_{t,d,t-h} = (\eta^e_{t,d}, \ldots, \eta^e_{t-h+1,d}, \tilde{\eta}^e_{t-h,d}, \eta^e_{t-h-1,d}, \ldots)$$

to be $\mathcal{H}^e_{t,d}$ with the $(t-h)$-th element $\eta^e_{t-h,d}$ replaced with $\tilde{\eta}^e_{t-h,d}$. Analogously, for $\boldsymbol{e} \in \{\boldsymbol{\varepsilon}, \boldsymbol{\xi}\}$ define $\tilde{\mathcal{H}}^{\boldsymbol{e}}_{t,t-h}$ as $\mathcal{H}^{\boldsymbol{e}}_t$ with the $(t-h)$-th input $\eta^{\boldsymbol{e}}_{t-h}$ replaced with $\tilde{\eta}^{\boldsymbol{e}}_{t-h}$ as in Subsection B.1. Similarly, for the product of errors R define $\tilde{\mathcal{H}}^R_{t,m,t-h}$ as $\mathcal{H}^R_{t,m}$ with the $(t-h)$-th element $\eta^R_{t-h,m}$ replaced with the iid copy $\tilde{\eta}^R_{t-h,m}$. Analogously, define $\tilde{\mathcal{H}}^{\boldsymbol{R}}_{t,t-h}$ as $\mathcal{H}^{\boldsymbol{R}}_t$ with the $(t-h)$-th input $\eta^{\boldsymbol{R}}_{t-h}$ replaced with $\tilde{\eta}^{\boldsymbol{R}}_{t-h}$ as in Subsection B.1.

Define L^∞ versions of the functional dependence measure for the error processes $G^e_{P,t,n,l,d}(\mathcal{H}^e_{t,d})$ for $(e, l, d) \in \mathbb{E}$, $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$ for $h \in \mathbb{N}_0$ as

$$\theta^{e,\infty}_{P,t,n,l,d}(h) = \inf\{K \geq 0 : \mathbb{P}_P(|G^e_{P,t,n,l,d}(\mathcal{H}^e_{t,d}) - G^e_{P,t,n,l,d}(\tilde{\mathcal{H}}^e_{t,d,t-h})| > K) = 0\},$$

and for the vector-valued $\boldsymbol{G}^{\boldsymbol{e}}_{P,t,n}(\mathcal{H}^{\boldsymbol{e}}_t)$ for $\boldsymbol{e} \in \{\boldsymbol{\varepsilon}, \boldsymbol{\xi}\}$, $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$ for $h \in \mathbb{N}_0$, $r \geq 1$ as

$$\theta^{\boldsymbol{e},\infty}_{P,t,n}(h, r) = \inf\{K \geq 0 : \mathbb{P}_P(||\boldsymbol{G}^{\boldsymbol{e}}_{P,t,n}(\mathcal{H}^{\boldsymbol{e}}_t) - \boldsymbol{G}^{\boldsymbol{e}}_{P,t,n}(\tilde{\mathcal{H}}^{\boldsymbol{e}}_{t,t-h})||_r > K) = 0\}.$$

Define the functional dependence measures for the product of error processes $G^R_{P,t,n,m}(\mathcal{H}^R_{t,m})$ for $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$ for each $h \in \mathbb{N}_0$, $q \geq 1$, as

$$\theta^R_{P,t,n,m}(h, q) = [\mathbb{E}_P(|G^R_{P,t,n,m}(\mathcal{H}^R_{t,m}) - G^R_{P,t,n,m}(\tilde{\mathcal{H}}^R_{t,m,t-h})|^q)]^{1/q},$$

and for the vector-valued $\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_t)$ for $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$ for each $h \in \mathbb{N}_0$, $q \geq 1$, $r \geq 1$ as

$$\theta^{\boldsymbol{R}}_{P,t,n}(h, q, r) = [\mathbb{E}_P(||\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_t) - \boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\tilde{\mathcal{H}}^{\boldsymbol{R}}_{t,t-h})||^q_r)]^{1/q}.$$

We make the following assumptions so that the temporal dependence and nonstationarity of the process can be controlled uniformly over a collection of distributions $\mathcal{P}_n$. The following assumption is that there is a polynomial decay of the temporal dependence. To make the effect of the (possibly growing) dimension more transparent, we impose these conditions on each dimension of the process. Note that we will often write the time as 0 when the time of the input sequence does not matter due to stationarity.

Assumption 6 (Distribution-uniform decay of temporal dependence)**.** *We assume that there exist $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 2$ such that for all $t \in \mathcal{T}_n$, $(e, l, d) \in \mathbb{E}$, it holds that*

$$\sup_{P \in \mathcal{P}_n} ||G^{e}_{P,t,n,l,d}(\mathcal{H}^{e}_{0,d})||_{L^{\infty}(P)} \leq \bar{\Theta}^{\infty}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{e,\infty}_{P,t,n,l,d}(h) \leq \bar{\Theta}^{\infty} \cdot (h \vee 1)^{-\bar{\beta}^{\infty}}, \qquad h \geq 0.$$

Also, for additional control in terms of the product of errors alone, we also assume that there exist $\bar{\Theta}^{R} > 0$, $\bar{\beta}^{R} > 2$, $\bar{q}^{R} > 4$, such that for all $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$, it holds that

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(|G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{0,m})|^{\bar{q}^{R}})]^{1/\bar{q}^{R}} \leq \bar{\Theta}^{R}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{R}_{P,t,n,m}(h, \bar{q}^{R}) \leq \bar{\Theta}^{R} \cdot (h \vee 1)^{-\bar{\beta}^{R}}, \qquad h \geq 0.$$

Note that in the previous assumption, it is possible to further upper bound the functional dependence measures and moment bounds for the product of errors in terms of $\bar{\Theta}^{\infty}$ by using the triangle inequality and Hölder's inequality. Also, note that the constants in the previous assumption do not depend on n.

By Jensen's inequality, we have the following functional dependence measures for the corresponding vector-valued processes.

Recall $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 2$. For all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $\boldsymbol{e} \in (\boldsymbol{\varepsilon}, \boldsymbol{\xi})$, we have that

$$\sup_{P \in \mathcal{P}_n} \big|\big|||\boldsymbol{G}^{\boldsymbol{e}}_{P,t,n}(\mathcal{H}^{\boldsymbol{e}}_{0})||_2\big|\big|_{L^{\infty}(P)} \leq M_n^{\frac{1}{2}} \bar{\Theta}^{\infty}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{e},\infty}_{P,t,n}(h, 2) \leq M_n^{\frac{1}{2}} \bar{\Theta}^{\infty} \cdot (h \vee 1)^{-\bar{\beta}^{\infty}}, \qquad h \geq 0.$$

Also, recall $\bar{\Theta}^{R} > 0$, $\bar{\beta}^{R} > 2$, $\bar{q}^{R} > 4$. For all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, it holds that

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(||\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_{0})||_2^{\bar{q}^{R}})]^{1/\bar{q}^{R}} \leq M_n^{\frac{1}{2}} \bar{\Theta}^{R}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{R}}_{P,t,n}(h, \bar{q}^{R}, 2) \leq M_n^{\frac{1}{2}} \bar{\Theta}^{R} \cdot (h \vee 1)^{-\bar{\beta}^{R}}, \qquad h \geq 0.$$

Next, we impose the following regularity conditions to control the nonstationarity uniformly over a collection of distributions $\mathcal{P}_n$ which is based on assumption (G.2) from control from Mies and Steland [MS22]. Again, we impose these regularity conditions on each dimension so that the conditions can be easily verified even the high-dimensional setting.

Assumption 7 (Distribution-uniform total variation condition for nonstationarity)**.** *Recall $\bar{\Theta}^{R} > 0$ from Assumption 6. For each $n \in \mathbb{N}$, we assume that there exist constants $\bar{\Gamma}^{R}_{n} \geq 1$ and powers $\bar{q}^{R} \geq 2$ such that for all $t \in \mathcal{T}_n$, it holds that*

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t=\mathbb{T}_n^{-}+1}^{\mathbb{T}_n^{+}} \max_{m} \left(\mathbb{E}_P |G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{0,m}) - G^{R}_{P,t-1,n,m}(\mathcal{H}^{R}_{0,m})|^{\bar{q}^{R}} \right)^{1/\bar{q}^{R}} \right) \leq \bar{\Theta}^{R} \bar{\Gamma}^{R}_{n}.$$

Recall $\bar{\Theta}^{R} > 0$ from Assumption 6, and $\bar{\Gamma}^{R}_{n} \geq 1$ with $\bar{q}^{R} \geq 2$ from Assumption 7. By Jensen's inequality, we have that

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t=\mathbb{T}_n^{-}+1}^{\mathbb{T}_n^{+}} \left(\mathbb{E}_P ||G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{0,m}) - G^{R}_{P,t-1,n,m}(\mathcal{H}^{R}_{0,m})||_2^{\bar{q}^{R}} \right)^{1/\bar{q}^{R}} \right) \leq M_n^{\frac{1}{2}} \bar{\Theta}^{R} \bar{\Gamma}^{R}_{n}.$$

To avoid specific conditions on the adaptive statistical learning algorithms, we make the following assumption about the decay in temporal dependence of the prediction errors.

Assumption 8 (Distribution-uniform decay of temporal dependence for prediction error)**.** *As in the discussion following Assumption 5, let $\tilde{\mathcal{H}}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t,t-h}$ be $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t}$ with the $(t-h)$-th input replaced with its iid copy. Assume that there exist $\bar{\Theta}^{\diamond} > 0$, $\bar{\beta}^{\diamond} > 2$ such that for all $\boldsymbol{\kappa} \in (\boldsymbol{f}, \boldsymbol{g})$, it holds that*

$$\sum_{t \in \mathcal{T}_n} \mathbb{E}_P[||\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t}) - \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{P,t,n}(\tilde{\mathcal{H}}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t,t-h})||_2^2] \leq \tau_n^2 T_n M_n^{-1} \bar{\Theta}^{\diamond} (h \vee 1)^{-\bar{\beta}^{\diamond}}.$$

That is, the cumulative distance between the original squared prediction errors and the "perturbed" version with the $(t-h)$-th input replaced by its iid copy will decrease as h increases (i.e. through the historical training data for the adaptive predictors and for the covariate at time t). Note that in Assumption 2, we assume that the cumulative squared prediction errors grow sublinearly. It is possible to show that this holds using lower-level dependence assumptions by making additional assumptions, adding and subtracting cross-terms, and applying the triangle inequality. More specifically, the temporal dependence of the cumulative prediction errors from the previous assumption can be upper bounded in terms of (1) how the average prediction errors of a fixed sequence of predictors depend on past inputs for the covariates, and (2) how the predictions of a sequence of predictors (with fixed covariates) depend on past inputs for the historical training data used to estimate that sequence of predictors. Further, under additional assumptions about the statistical learning algorithm (or Lipschitz-type assumptions), it is possible to show that this holds using dependence assumptions on the observed processes.

C The Hardness of Strong Granger Causality Testing

C.1 Hypothesis testing for nonstationary time series

In this subsection, we introduce notation for distribution-uniform hypothesis testing for high-dimensional nonstationary time series. For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, and $t \in \mathcal{T}_n$, let $\mathcal{P}_{0,n,m,t} \subset \mathcal{P}_n$ be a potentially composite null hypothesis consisting of a collection of distributions for the process. We denote the global null hypothesis for the family of null hypotheses $(\mathcal{P}_{0,n,m,t})_{m\in\mathcal{M}_n,t\in\mathcal{T}_n}$ by

$$\mathcal{P}_{0,n} = \bigcap_{t\in\mathcal{T}_n} \bigcap_{m\in\mathcal{M}_n} \mathcal{P}_{0,n,m,t}.$$

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, and $t \in \mathcal{T}_n$, let $\psi_{n,m,t}$ be a potentially randomized test that can be applied to the data such that

$$\psi_{n,m,t} : \mathbb{R}^{(d_X+d_Y+d_Y)\cdot T_n} \times [0,1] \to \{0,1\}$$

is a measurable function where 1 indicates rejecting the null hypothesis $H_{0,n,m,t}$. The last argument is for a uniform random variable $U_{n,m,t} \sim \mathcal{U}[0,1]$ that is independent of the data. Note that the joint distribution of the $(U_{n,m,t})_{m\in\mathcal{M}_n,t\in\mathcal{T}_n}$ can be very complicated. For some $n \in \mathbb{N}$ and $P \in \mathcal{P}_n$, denote the subset of indices at times $t \in \mathcal{T}_n$ in which the null hypothesis is true by

$$\mathcal{M}_{P,n,t} = \{m \in \mathcal{M}_n : H_{0,n,m,t} \text{ is true under } P\}.$$

For some sample size $n \in \mathbb{N}$, level $\alpha \in (0,1)$, global null hypothesis $\mathcal{P}_{0,n}$, and collection of distributions $\mathcal{P}_n$, we say that the family of tests $(\psi_{n,m,t})_{m\in\mathcal{M}_n,t\in\mathcal{T}_n}$ has *valid FWER control in the weak sense at sample size n* if

$$\sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\bigcup_{t\in\mathcal{T}_n} \bigcup_{m\in\mathcal{M}_n} \{\psi_{n,m,t}=1\}\right) \leq \alpha,$$

and *valid FWER control in the strong sense at sample size n* if

$$\sup_{P\in\mathcal{P}_n} \mathbb{P}_P\left(\bigcup_{t\in\mathcal{T}_n} \bigcup_{m\in\mathcal{M}_{P,n,t}} \{\psi_{n,m,t}=1\}\right) \leq \alpha.$$

For some level $\alpha \in (0,1)$, sequence of global null hypotheses $(\mathcal{P}_{0,n})_{n\in\mathbb{N}}$, and sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we say that the sequence of families of tests $((\psi_{n,m,t})_{m\in\mathcal{M}_n,t\in\mathcal{T}_n})_{n\in\mathbb{N}}$ has *uniformly asymptotic FWER control in the weak sense* if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\bigcup_{t\in\mathcal{T}_n} \bigcup_{m\in\mathcal{M}_n} \{\psi_{n,m,t}=1\}\right) \leq \alpha,$$

and *uniformly asymptotic FWER control in the strong sense* if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{P}_P\left(\bigcup_{t\in\mathcal{T}_n}\bigcup_{m\in\mathcal{M}_{P,n,t}} \{\psi_{n,m,t}=1\}\right) \leq \alpha.$$

Lastly, for some level $\alpha \in (0,1)$, sequence of global null hypotheses $(\mathcal{P}_{0,n})_{n\in\mathbb{N}}$, and sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we say that a test ψ_n has *valid level at sample size* n if

$$\sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\{\psi_n=1\}\right) \leq \alpha,$$

and that a sequence of tests $(\psi_n)_{n\in\mathbb{N}}$ has *uniformly asymptotic level* if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\{\psi_n=1\}\right) \leq \alpha.$$

As in Shah and Peters [SP20], we will construct an meta test statistic [Hoc+23] based on some vector norm (e.g. maximum or Euclidean) of individual test statistics $\psi_{n,m,t}$ corresponding to a time t and index m pair. Denote this meta test statistic by ψ_n^{meta}. We will reject the global null hypothesis if our meta test statistic ψ_n^{meta} exceeds an approximated $1-\alpha$ quantile found through a multiplier-esque bootstrap procedure [CCK13]. Crucially, we can only hope to control the *size* of this test uniformly over *subsets of the null hypothesis*. All of the details for the test will be discussed in Subsection 2.4.

C.2 No-free-lunch result

In this subsection, we clarify that the hardness result from Shah and Peters [SP20] also implies that conditional independence testing (and strong Granger causality testing, in particular) for general discrete-time stochastic processes is a hard statistical problem. Our main purpose for doing this is to spread the news about this hardness result to the time series research community and to clarify its relevance. For each $n \in \mathbb{N}$, let $\mathcal{E}_{0,n}^{\text{Leb}}$ be the set of all distributions for general stochastic process $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t\in[n]}$ such that for each $t \in \mathcal{T}_n$ the distribution of $(\boldsymbol{X}_{t,n}, \boldsymbol{Y}_{t,n}, \boldsymbol{Z}_{t,n})$ is absolutely continuous with respect to the Lebesgue measure. Let $\mathcal{P}_{0,n}^{\text{CI, Leb}} \subset \mathcal{E}_{0,n}^{\text{Leb}}$ be the subset of distributions under which $\boldsymbol{X}_{t,n} \perp\!\!\!\perp \boldsymbol{Y}_{t,n} \mid \boldsymbol{Z}_{t,n}$ for all $t \in \mathcal{T}_n$. For any $K \in (0,\infty]$, let $\mathcal{E}_{0,n,K}^{\text{Leb}} \subseteq \mathcal{E}_{0,n}^{\text{Leb}}$ be the subset of all distributions for the process such that for each $t \in \mathcal{T}_n$ the distribution of $(\boldsymbol{X}_{t,n}, \boldsymbol{Y}_{t,n}, \boldsymbol{Z}_{t,n})$ has support contained strictly within an ℓ_∞ ball of radius K, where we take $\mathcal{E}_{0,n,\infty}^{\text{Leb}} = \mathcal{E}_{0,n}^{\text{Leb}}$. Let $\mathcal{Q}_{0,n}^{\text{CD, Leb}} = \mathcal{E}_{0,n}^{\text{Leb}} \setminus \mathcal{P}_{0,n}^{\text{CI, Leb}}$, and set $\mathcal{P}_{0,n,K}^{\text{CI, Leb}} = \mathcal{E}_{0,n,K}^{\text{Leb}} \cap \mathcal{P}_{0,n}^{\text{CI, Leb}}$ and $\mathcal{Q}_{0,n,K}^{\text{CD, Leb}} = \mathcal{E}_{0,n,K}^{\text{Leb}} \cap \mathcal{Q}_{0,n}^{\text{CD, Leb}}$.

Given any $n \in \mathbb{N}$, $\alpha \in (0,1)$, let ψ_n be a potentially randomized test with valid level α for the global null hypothesis $\mathcal{P}_{0,n,K}^{\text{CI, Leb}}$. For some $\mathcal{T}_n^* \times \mathcal{M}_n^* \subseteq \mathcal{T}_n \times \mathcal{M}_n$, consider the subcollection of alternatives $\mathcal{Q}_{0,n,K}^{\text{CD, Leb}}(\mathcal{T}_n^* \times \mathcal{M}_n^*) \subset \mathcal{Q}_{0,n,K}^{\text{CD, Leb}}$ such that $X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n}$ for all $(t,m) \in \mathcal{T}_n^* \times \mathcal{M}_n^*$. Let $I_n^* \times J_n^* \subseteq [d_X] \times [d_Y]$ denote the set of pairs of dimensions (i,j) that appear in at least one dimension/time-offset tuple $m = (i,j,a,b) \in \mathcal{M}_n^*$. Consider the subcollection of alternatives $\mathcal{Q}_{0,n,K}^{\text{CD, Leb, IID}}(I_n^* \times J_n^*)$ such that the stochastic process is an iid process and $X_{t,n,i,a} \not\perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n}$ for all dimension pairs $(i,j) \in \mathcal{I}_n^* \times \mathcal{J}_n^*$ and for all times $t \in \mathcal{T}_n$ and time-offsets $a \in A_i$, $b \in B_j$. Note that the conditional dependence relationships between each of the pairs (i,j) must hold for all times t and time-offsets pairs (a,b) because for iid processes the conditional independence structure cannot vary over time or change with time-offsets. In other words, $\mathcal{Q}_{0,n,K}^{\text{CD, Leb, IID}}(I_n^* \times J_n^*)$ is simply a collection of product distributions $Q = \bigotimes_{t=1}^n Q_t$ of the time-invariant joint distributions Q_t for the triplet of random vectors $(X_{t,n}, Y_{t,n}, Z_{t,n})$ such that the corresponding conditional dependence relationship between the dimension pairs $(i,j) \in \mathcal{I}_n^* \times \mathcal{J}_n^*$ hold for all times and time-offsets. Denote the collection of all alternatives for such iid processes where $X_{t,n} \not\perp\!\!\!\perp Y_{t,n} \mid Z_{t,n}$ for all $t \in \mathcal{T}_n$ by $\mathcal{Q}_{0,n,K}^{\text{CD, Leb, IID}}$.

To avoid specific assumptions about the conditional independence test ψ_n for stochastic processes, we make the following minimal assumption which captures the idea that ψ_n should not gain power if we introduce non-stationarity or temporal dependence. For each $Q \in \mathcal{Q}_{0,n,K}^{\text{CD, Leb}}(\mathcal{T}_n^* \times \mathcal{M}_n^*)$ there is a $Q' \in \mathcal{Q}_{0,n,K}^{\text{CD, Leb, IID}}(I_n^* \times J_n^*)$ such that $\mathbb{P}_Q(\psi_n = 1) \leq \mathbb{P}_{Q'}(\psi_n = 1)$ for all $\mathcal{T}_n^* \times \mathcal{M}_n^* \subseteq \mathcal{T}_n \times \mathcal{M}_n$. Hence, for each $Q \in \mathcal{Q}_{0,n,K}^{\text{CD, Leb}}$ there exists a $Q' \in \mathcal{Q}_{0,n,K}^{\text{CD, Leb, IID}}$ such that $\mathbb{P}_Q(\psi_n = 1) \leq \mathbb{P}_{Q'}(\psi_n = 1)$.

Next, consider the subset of distributions $\mathcal{P}^{\text{CI, Leb, IID}}_{0,n,K} \subset \mathcal{P}^{\text{CI, Leb}}_{0,n,K}$ such that $X_{t,n} \perp\!\!\!\perp Y_{t,n} \mid Z_{t,n}$ for all $t \in \mathcal{T}_n$ and the stochastic process is an iid process under any $P \in \mathcal{P}^{\text{CI, Leb, IID}}_{0,n,K}$. By basic properties of the supremum, if the test has valid level α for the global null hypothesis $\mathcal{P}^{\text{CI, Leb}}_{0,n,K}$ then it also will for $\mathcal{P}^{\text{CI, Leb, IID}}_{0,n,K} \subset \mathcal{P}^{\text{CI, Leb}}_{0,n,K}$. That is, we have

$$\sup_{P \in \mathcal{P}^{\text{CI, Leb}}_{0,n,K}} \mathbb{P}_P(\psi_n = 1) \leq \alpha,$$

and hence we have

$$\sup_{P' \in \mathcal{P}^{\text{CI, Leb, IID}}_{0,n,K}} \mathbb{P}_{P'}(\psi_n = 1) \leq \alpha.$$

Note that $\mathcal{P}^{\text{CI, Leb, IID}}_{0,n,K}$ is simply a collection of product distributions of the form $P = \bigotimes_{t=1}^{n} P_t$, where each $P_t \in \tilde{\mathcal{P}}^{\text{CI, Leb, IID}}_{0,n,K}$ is a time-invariant joint distribution for the triplet of random vectors $(X_{t,n}, Y_{t,n}, Z_{t,n})$ such that $X_{t,n} \perp\!\!\!\perp Y_{t,n} \mid Z_{t,n}$. That is, the distribution of $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t \in [n]}$ under the associated measure $\mathbb{P}_P$ is $P = \bigotimes_{t=1}^{n} P_t$, and the time-invariant distribution of $(X_{t,n}, Y_{t,n}, Z_{t,n})$ under the associated measure $\mathbb{P}_{P_t}$ is P_t. By switching the viewpoint from observing one realization of a vector-valued iid stochastic process to multiple samples of iid random vectors, we have that

$$\sup_{P'_t \in \tilde{\mathcal{P}}^{\text{CI, Leb, IID}}_{0,n,K}} \mathbb{P}_{P'_t}(\psi_n = 1) \leq \alpha.$$

By the *no-free-lunch* result from Theorem 2 in Shah and Peters [SP20], we have that

$$\mathbb{P}_{Q'_t}(\psi_n = 1) \leq \alpha,$$

for all $Q'_t \in \tilde{\mathcal{Q}}^{\text{CD, Leb, IID}}_{0,n,K}$, and hence for all $Q' = \bigotimes_{t=1}^{n} Q'_t \in \mathcal{Q}^{\text{CD, Leb, IID}}_{0,n,K}$ we have that

$$\mathbb{P}_{Q'}(\psi_n = 1) \leq \alpha,$$

again by switching the viewpoint from observing one realization of a vector-valued iid stochastic process to multiple samples of iid random vectors. By the assumption made previously, for all $Q \in \mathcal{Q}^{\text{CD, Leb}}_{0,n,K}$ we can find a $Q' \in \mathcal{Q}^{\text{CD, Leb, IID}}_{0,n,K}$ such that $\mathbb{P}_Q(\psi_n = 1) \leq \mathbb{P}_{Q'}(\psi_n = 1)$, which implies the final result that

$$\mathbb{P}_Q(\psi_n = 1) \leq \alpha$$

for all $Q \in \mathcal{Q}^{\text{CD, Leb}}_{0,n,K}$. That is, the conditional independence test ψ_n for general discrete-time stochastic processes cannot have power against any alternative $Q \in \mathcal{Q}^{\text{CD, Leb}}_{0,n,K}$. Consequently, we must restrict the null hypothesis of conditional independence. Again, this is not surprising in light of Shah and Peters [SP20] but we find it useful to diffuse the awareness about this hardness result and to clarify its relevance to time series analysis.

D Locally Stationary Time Series

D.1 Setting and notation

In this section, we show that our testing framework can be used with a well-studied special class of nonstationary time series. Let us reflect on the purpose for introducing locally stationary time series by recalling Dahlhaus [Dah97]. In regular time series analysis, letting n approach infinity corresponds to getting information about the future. However, if the process of interest is nonstationary, letting n approach infinity does not give us any additional information about the process at earlier points in time. Dahlhaus [Dah97] introduced the idea of rescaling time to the unit interval so that infill asymptotics can be used to study nonstationary processes. In this setting, the sample size n no longer corresponds to getting information about the future, but instead denotes the number of observations we have of a process that changes slowly over time. As n increases, we get more and more data about each *local* structure of the nonstationary process of interest. Zhou and Wu [ZW09] introduced the framework for representing locally stationary time series as nonlinear functions of iid random elements

as in Wu [Wu05]. Before introducing the details for the causal representation, we must introduce the necessary notation.

Recall the subset of original times $\mathcal{T}_n \subseteq \{1, \ldots, n\}$ in which *all* time-offsets of each dimension of $X_{t,n}$, $Y_{t,n}$, and $Z_{t,n}$ are actually observed, where

$$\mathcal{T}_n = \{1 + \lambda_n - \min(\{0\} \cup A \cup B \cup C), n - \max(\{0\} \cup A \cup B \cup C)\},$$

where $\lambda_n \geq 0$ is an optional late-starting parameter that will be used in Subsection D.6.

Similarly, denote the corresponding interval of rescaled times in which all time-offsets are well-defined by

$$\mathcal{U}_n = \left[\frac{1}{n} + \frac{\lambda}{n} - \frac{\min(\{0\} \cup A \cup B \cup C)}{n}, 1 - \frac{\max(\{0\} \cup A \cup B \cup C)}{n}\right] \subset [0, 1].$$

Next, we introduce the causal representation for high-dimensional locally stationary processes, which is most similar to Example 3 in Mies and Steland [MS22]. This representation is different from the previous causal representation from Assumption 1 in that we now assume the nonlinear system is defined at all rescaled times $u \in [0, 1]$.

Assumption 9 (Causal representation of locally stationary processes)**.** *Let $\mathcal{H}_t^X = (\eta_t^X, \eta_{t-1}^X, \ldots)$, $\mathcal{H}_t^Y = (\eta_t^Y, \eta_{t-1}^Y, \ldots)$, $\mathcal{H}_t^Z = (\eta_t^Z, \eta_{t-1}^Z, \ldots)$ where $(\eta_t^X)_{t\in\mathbb{Z}}$, $(\eta_t^Y)_{t\in\mathbb{Z}}$, $(\eta_t^Z)_{t\in\mathbb{Z}}$ are sequences of iid random elements. Assume that we can represent each dimension of the observed sequence as the output of an evolving nonlinear system that was given a sequence of iid inputs:*

$$X_{t,n,i} = \tilde{G}_{n,i}^X(t/n, \mathcal{H}_t^X),\ Y_{t,n,j} = \tilde{G}_{n,j}^Y(t/n, \mathcal{H}_t^Y),\ Z_{t,n,k} = \tilde{G}_{n,k}^Z(t/n, \mathcal{H}_t^Z),$$

where the systems are defined for all $u \in [0, 1]$ by

$$\tilde{X}_{t,n,i}(u) = \tilde{G}_{n,i}^X(u, \mathcal{H}_t^X),\ \tilde{Y}_{t,n,j}(u) = \tilde{G}_{n,j}^Y(u, \mathcal{H}_t^Y),\ \tilde{Z}_{t,n,k}(u) = \tilde{G}_{n,k}^Z(u, \mathcal{H}_t^Z),$$

so that we may write $X_{t,n,i} = \tilde{X}_{t,n,i}(t/n)$, $Y_{t,n,j} = \tilde{Y}_{t,n,j}(t/n)$, $Z_{t,n,k} = \tilde{Z}_{t,n,k}(t/n)$.

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, we assume that $\tilde{G}_{n,i}^X(u, \cdot)$, $\tilde{G}_{n,j}^Y(u, \cdot)$, $\tilde{G}_{n,k}^Z(u, \cdot)$ are each measurable functions from $\mathbb{R}^\infty$ to $\mathbb{R}$ (where we endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections) such that $\tilde{G}_{n,i}^X(u, \mathcal{H}_s^X)$, $\tilde{G}_{n,j}^Y(u, \mathcal{H}_s^Y)$, $\tilde{G}_{n,k}^Z(u, \mathcal{H}_s^Z)$ are each well-defined random variables for each $s \in \mathbb{Z}$ and $(\tilde{G}_{n,i}^X(u, \mathcal{H}_s^X))_{s\in\mathbb{Z}}$, $(\tilde{G}_{n,j}^Y(u, \mathcal{H}_s^Y))_{s\in\mathbb{Z}}$, $(\tilde{G}_{n,k}^Z(u, \mathcal{H}_s^Z))_{s\in\mathbb{Z}}$ are each stationary ergodic time series.

To avoid repeating the same ideas many times, let us state some properties that all causal representations in this subsection will have. These causal representations will all be measurable functions on $\mathbb{R}^\infty$, where we will always endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections. As stated in Assumption 9, the causal mechanism at each time $t \in \mathcal{T}_n$ with the input sequence up to some time $s \in \mathbb{Z}$ is a well-defined r.v., and the process induced by considering the sequence of inputs up to each time $s \in \mathbb{Z}$ is a stationary ergodic time series.

In light of Assumption 9, we have the following causal representations for all dimensions and no time-offsets by

$$\tilde{X}_{t,n}(u) = \tilde{G}_n^X(u, \mathcal{H}_t^X) = (\tilde{G}_{n,i}^X(u, \mathcal{H}_t^X))_{i\in[d_X]},$$
$$\tilde{Y}_{t,n}(u) = \tilde{G}_n^Y(u, \mathcal{H}_t^Y) = (\tilde{G}_{n,j}^Y(u, \mathcal{H}_t^Y))_{j\in[d_Y]},$$

$$\tilde{Z}_{t,n}(u) = \tilde{G}_n^Z(u, \mathcal{H}_t^Z) = (\tilde{G}_{n,k}^Z(u, \mathcal{H}_t^Z))_{k\in[d_Z]},$$

so that we may write the observed sequence as $X_{t,n} = \tilde{X}_{t,n}(t/n)$, $Y_{t,n} = \tilde{Y}_{t,n}(t/n)$, $Z_{t,n} = \tilde{Z}_{t,n}(t/n)$. Also, we have causal representations for dimensions $i \in [d_X]$, $j \in [d_Y]$, $k \in [d_Z]$ with time-offsets $a \in A_i$, $b \in B_j$, $c \in C_k$

$$\tilde{X}_{t,n,i,a}(u) = \tilde{G}_{n,i,a}^X\left(u, \mathcal{H}_{t,a}^X\right) = \tilde{G}_{n,i}^X\left(u + \frac{a}{n}, \mathcal{H}_{t+a}^X\right),$$

$$\tilde{Y}_{t,n,j,b}(u) = \tilde{G}_{n,j,b}^Y\left(u, \mathcal{H}_{t,b}^Y\right) = \tilde{G}_{n,j}^Y\left(u + \frac{b}{n}, \mathcal{H}_{t+b}^Y\right),$$

$$\tilde{Z}_{t,n,k,c}(u) = \tilde{G}^Z_{n,k,c}\left(u, \mathcal{H}^Z_{t,c}\right) = \tilde{G}^Z_{n,k}\left(u + \frac{c}{n}, \mathcal{H}^Z_{t+c}\right),$$

where $\mathcal{H}^X_{t,a} = (\eta^X_{t+a}, \eta^X_{t-1+a}, \ldots)$, $\mathcal{H}^Y_{t,b} = (\eta^Y_{t+b}, \eta^Y_{t-1+b}, \ldots)$, and $\mathcal{H}^Z_{t,c} = (\eta^Z_{t+c}, \eta^Z_{t-1+c}, \ldots)$, so that we may write $X_{t,n,i,a} = \tilde{X}_{t,n,i,a}(t/n)$, $Y_{t,n,j,b} = \tilde{Y}_{t,n,j,b}(t/n)$, $Z_{t,n,k,c} = \tilde{Z}_{t,n,k,c}(t/n)$ for each dimension with time-offset of the observed sequence. We can then write the causal representation of the vectors with all dimensions and time-offsets as

$$\tilde{\boldsymbol{X}}_{t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{X}}_n(u, \mathcal{H}^{\boldsymbol{X}}_t) = (\tilde{G}^X_{n,i,a}(\mathcal{H}^X_{t,a}))_{i\in[d_X], a\in A_i},$$

$$\tilde{\boldsymbol{Y}}_{t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{Y}}_n(u, \mathcal{H}^{\boldsymbol{Y}}_t) = (\tilde{G}^X_{n,j,b}(\mathcal{H}^Y_{t,b}))_{j\in[d_Y], b\in B_j},$$

$$\tilde{\boldsymbol{Z}}_{t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{Z}}_n(u, \mathcal{H}^{\boldsymbol{Z}}_t) = (\tilde{G}^Z_{n,k,c}(\mathcal{H}^Z_{t,c}))_{k\in[d_Z], c\in C_k},$$

where $\mathcal{H}^{\boldsymbol{X}}_t = (\eta^{\boldsymbol{X}}_t, \eta^{\boldsymbol{X}}_{t-1}, \ldots)$, $\mathcal{H}^{\boldsymbol{Y}}_t = (\eta^{\boldsymbol{Y}}_t, \eta^{\boldsymbol{Y}}_{t-1}, \ldots)$, $\mathcal{H}^{\boldsymbol{Z}}_t = (\eta^{\boldsymbol{Z}}_t, \eta^{\boldsymbol{Z}}_{t-1}, \ldots)$, and $\eta^{\boldsymbol{X}}_t = \eta^X_{t+a_{\max}}$, $\eta^{\boldsymbol{Y}}_t = \eta^Y_{t+b_{\max}}$, $\eta^{\boldsymbol{Z}}_t = \eta^Z_{t+c_{\max}}$, so that we may write $\boldsymbol{X}_{t,n} = \tilde{\boldsymbol{X}}_{t,n}(t/n)$, $\boldsymbol{Y}_{t,n} = \tilde{\boldsymbol{Y}}_{t,n}(t/n)$, $\boldsymbol{Z}_{t,n} = \tilde{\boldsymbol{Z}}_{t,n}(t/n)$ for the observed sequence including dimensions and all time-offsets. We assume that for each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $u \in \mathcal{U}_n$, $t \in \mathcal{T}_n$ the distribution of $(\tilde{X}_{t,n,i,a}(u), \tilde{Y}_{t,n,j,b}(u), \tilde{\boldsymbol{Z}}_{t,n}(u))$ is absolutely continuous with respect to the Lebesgue measure.

We introduce the probability space and distributions again for the sake of completeness. Let Ω be a sample space, $\mathcal{B}$ the Borel sigma-algebra, and $(\Omega, \mathcal{B})$ a measurable space. For fixed $n \in \mathbb{N}$, let $(\Omega, \mathcal{B})$ be equipped with a family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that the distribution of the stochastic system

$$(\tilde{G}^X_n(u, \mathcal{H}^X_t), \tilde{G}^Y_n(u, \mathcal{H}^Y_t), \tilde{G}^Z_n(u, \mathcal{H}^Z_t))_{u\in[0,1], t\in\mathbb{Z}}$$

under $\mathbb{P}_P$ is $P \in \mathcal{P}_n$ and $\mathcal{P}_n \subset \text{Prob}[(\mathbb{R}^{d_X+d_Y+d_Z})^{[0,1]\times\mathbb{Z}}]$ is a subset of the set of Borel probability measures on functions from $[0,1] \times \mathbb{Z}$ to $\mathbb{R}^{d_X+d_Y+d_Z}$. Again, the family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}_n}$ is defined with respect to the same measurable space $(\Omega, \mathcal{B})$, but need not have the same dominating measure.

D.2 Hypothesis testing for high-dimensional locally stationary time series

We introduce notation for distribution-uniform simultaneous hypothesis testing for high-dimensional locally stationary time series. For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $u \in \mathcal{U}_n$, define a potentially composite null hypothesis $\mathcal{P}_{0,n,m,u} \subset \mathcal{P}_n$. We denote the global null hypothesis for the family of null hypotheses $(\mathcal{P}_{0,n,m,u})_{m\in\mathcal{M}_n, u\in\mathcal{U}_n}$ by

$$\mathcal{P}_{0,n} = \bigcap_{u\in\mathcal{U}_n} \bigcap_{m\in\mathcal{M}_n} \mathcal{P}_{0,n,m,u}.$$

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $u \in \mathcal{U}_n$, let $\psi_{n,m,u}$ be a potentially randomized test that can be applied to the data such that

$$\psi_{n,m,u} : \mathbb{R}^{(d_X+d_Y+d_Y)\cdot T_n} \times [0,1] \to \{0,1\}$$

is a measurable function where 1 indicates rejecting the null hypothesis $H_{0,n,m,u}$ and where $T_n = |\mathcal{T}_n|$. The last argument is for a uniform random variable $U_{n,m,u} \sim \mathcal{U}[0,1]$ that is independent of the data. Note that the joint distribution of the $(U_{n,m,u})_{m\in\mathcal{M}_n, u\in\mathcal{U}_n}$ can be very complicated. For some $n \in \mathbb{N}$ and $P \in \mathcal{P}_n$, denote the subset of indices at rescaled time $u \in \mathcal{U}_n$ in which the null hypothesis is true by

$$\mathcal{M}_{P,n,u} = \{m \in \mathcal{M}_n : H_{0,n,m,u} \text{ is true under } P\}.$$

For some sample size $n \in \mathbb{N}$, level $\alpha \in (0,1)$, global null hypothesis $\mathcal{P}_{0,n}$, and collection of distributions $\mathcal{P}_n$, we say that the family of tests $(\psi_{n,m,u})_{m\in\mathcal{M}_n, u\in\mathcal{U}_n}$ has *valid FWER control in the weak sense at sample size n* if

$$\sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\bigcup_{u\in\mathcal{U}_n} \bigcup_{m\in\mathcal{M}_n} \{\psi_{n,m,u} = 1\}\right) \le \alpha,$$

and *valid FWER control in the strong sense at sample size n* if

$$\sup_{P\in\mathcal{P}_n} \mathbb{P}_P\left(\bigcup_{u\in\mathcal{U}_n} \bigcup_{m\in\mathcal{M}_{P,n,u}} \{\psi_{n,m,u} = 1\}\right) \le \alpha.$$

For some level $\alpha \in (0,1)$, sequence of global null hypotheses $(\mathcal{P}_{0,n})_{n\in\mathbb{N}}$, and sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we say that the sequence of families of tests $((\psi_{n,m,u})_{m\in\mathcal{M}_n,u\in\mathcal{U}_n})_{n\in\mathbb{N}}$ has *uniformly asymptotic FWER control in the weak sense* if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,n}} \mathbb{P}_P\left(\bigcup_{u\in\mathcal{U}_n}\bigcup_{m\in\mathcal{M}_n}\{\psi_{n,m,u}=1\}\right) \leq \alpha,$$

and *uniformly asymptotic FWER control in the strong sense* if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{P}_P\left(\bigcup_{u\in\mathcal{U}_n}\bigcup_{m\in\mathcal{M}_{P,n,u}}\{\psi_{n,m,u}=1\}\right) \leq \alpha.$$

D.3 Conditional independence for high-dimensional locally stationary time series

Now, let us consider the specific hypothesis of conditional independence. In this section, we discuss a CI test for locally stationary time series that is asymptotically valid as $n\to\infty$, uniformly over a large collection of distributions $\mathcal{P}_n$ for which the null holds. Recall that $(\tilde{X}_{t,n,i,a}(u))_{t\in\mathbb{Z}}$, $(\tilde{Y}_{t,n,j,b}(u))_{t\in\mathbb{Z}}$, $(\tilde{\boldsymbol{Z}}_{t,n}(u))_{t\in\mathbb{Z}}$ are all stationary time series. For this reason, conditional independence holds regardless of the times for the input sequence used. Hence, in the hypotheses of conditional independence for locally stationary time series we may replace the t in the subscripts by any other three times $t_1,t_2,t_3\in\mathbb{Z}$.

In Subsection D.8, we also show that the same class of test statistics considered in Subsection 2.4 can be used in the locally stationary setting to test for the global null hypothesis

$$H_{0,n}^{\mathrm{CI}} : \tilde{X}_{t,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for all } u\in\mathcal{U}_n, \text{ for all } m\in\mathcal{M}_n, \tag{8}$$

for some $n\in\mathbb{N}$. Similar to the discussion in Subsection 2.2, there are four alternative hypotheses $H_{1,n}^{\mathrm{CI}}$ that can be used. We can always use the alternative

$$\tilde{X}_{t,n,i,a}(u) \not\perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for some } u\in\mathcal{U}_n, \text{ for some } m\in\mathcal{M}_n. \tag{9}$$

If the conditional independence relationships can be assumed to be time-invariant, we can use the alternative

$$\tilde{X}_{t,n,i,a}(u) \not\perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for all } u\in\mathcal{U}_n, \text{ for some } m\in\mathcal{M}_n. \tag{10}$$

In the "group of time series" setting in which it can be assumed that all indices have the same conditional independence relationships, we can use the alternative

$$\tilde{X}_{t,n,i,a}(u) \not\perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for all } u\in\mathcal{U}_n, \text{ for all } m\in\mathcal{M}_n. \tag{11}$$

if the conditional independence relationships are time-invariant. We can use the alternative hypothesis

$$\tilde{X}_{t,n,i,a}(u) \not\perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for some } u\in\mathcal{U}_n, \text{ for all } m\in\mathcal{M}_n. \tag{12}$$

if the "group of time series" has time-varying conditional independence relationships.

D.4 Basic setup and main ideas

For a fixed sample size $n\in\mathbb{N}$, distribution $P\in\mathcal{P}_n$, time $t\in\mathcal{T}_n$ and dimension/time-offset index tuple $m=(i,j,a,b)\in\mathcal{M}_n$, we can always decompose

$$X_{t,n,i,a} = f_{P,n,i,a}(t/n,\boldsymbol{Z}_{t,n}) + \varepsilon_{P,t,n,i,a},\ Y_{t,n,j,b} = g_{P,n,j,b}(t/n,\boldsymbol{Z}_{t,n}) + \xi_{P,t,n,j,b},$$

where $f_{P,n,i,a}(u,\boldsymbol{z}) = \mathbb{E}_P(\tilde{X}_{t,n,i,a}(u)|\tilde{\boldsymbol{Z}}_{t,n}(u)=\boldsymbol{z})$ and $g_{P,n,j,b}(u,\boldsymbol{z}) = \mathbb{E}_P(\tilde{Y}_{t,n,j,b}(u)|\tilde{\boldsymbol{Z}}_{t,n}(u)=\boldsymbol{z})$ are the time-varying regression functions. Define the process of error products by

$$R_{P,t,n,m} = \varepsilon_{P,t,n,i,a}\xi_{P,t,n,j,b}.$$

Similarly, denote the process of the residual products by

$$\hat{R}_{t,n,m} = \hat{\varepsilon}_{t,n,i,a}\hat{\xi}_{t,n,j,b},$$

where $\hat{\varepsilon}_{t,n,i,a} = X_{t,n,i,a} - \hat{f}_{n,i,a}(t/n, \boldsymbol{Z}_{t,n})$ and $\hat{\xi}_{t,n,j,b} = Y_{t,n,j,b} - \hat{g}_{n,j,b}(t/n, \boldsymbol{Z}_{t,n})$ and $\hat{f}_{n,i,a}(t/n, \cdot)$, $\hat{g}_{n,j,b}(t/n, \cdot)$ are estimates of $f_{P,n,i,a}(t/n, \cdot)$, $g_{P,n,j,b}(t/n, \cdot)$ created by time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively.

The covariate process $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ is a locally stationary time series and each of the dimensions can depend on one another. As in the previous setting, we can include any lags of any of the dimensions of the original time series $Z_{t,n}$ since these lags are known at time t. The error processes $(\varepsilon_{P,t,n,i,a})_{t\in\mathcal{T}_n}$, $(\xi_{P,t,n,j,b})_{t\in\mathcal{T}_n}$ can also be locally stationary time series that depend on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$, $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$, respectively. In Subsection D.5, we discuss some aspects of the theoretical framework for high-dimensional locally stationary time series.

Let us translate the "weak" conditional independence criterion of Daudin [Dau80] into the locally stationary time series setting. Assume that for each $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$ the joint distribution of $(\tilde{X}_{t,n,i,a}(u), \tilde{Y}_{t,n,j,b}(u), \tilde{\boldsymbol{Z}}_{t,n}(u))$ is absolutely continuous with respect to the Lebesgue measure. For some $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, if $\tilde{X}_{t,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u)$ then $\mathbb{E}_P[\phi(\tilde{X}_{t,n,i,a}(u), \tilde{\boldsymbol{Z}}_{t,n}(u))\varphi(\tilde{Y}_{t,n,j,b}(u), \tilde{\boldsymbol{Z}}_{t,n}(u)] = 0$ for all functions $\phi \in L^2_{\tilde{X}_{t,n,i,a}(u), \tilde{\boldsymbol{Z}}_{t,n}(u)}$ and $\varphi \in L^2_{\tilde{Y}_{t,n,j,b}(u), \tilde{\boldsymbol{Z}}_{t,n}(u)}$ such that $\mathbb{E}_P[\phi(\tilde{X}_{t,n,i,a}(u), \tilde{\boldsymbol{Z}}_{t,n}(u)) \mid \tilde{\boldsymbol{Z}}_{t,n}(u)] = 0$ and $\mathbb{E}_P[\varphi(\tilde{Y}_{t,n,j,b}(u), \tilde{\boldsymbol{Z}}_{t,n}(u))) \mid \tilde{\boldsymbol{Z}}_{t,n}(u))] = 0$ and hence the corresponding *local* expected conditional covariance

$$\rho_{P,t,n,m}(u) = \mathbb{E}_P[\mathrm{Cov}_P(\tilde{X}_{t,n,i,a}(u), \tilde{Y}_{t,n,j,b}(u)|\tilde{\boldsymbol{Z}}_{t,n}(u))]$$

is equal to zero. Hence, the process of error products from the time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ has mean zero. The exact notation will be introduced in Subsection D.4.

In particular, under the global null hypothesis of conditional independence (8), all of the expected conditional covariances $\rho_{P,t,n,m}(t/n)$ are equal to zero. Hence, we aim to detect conditional dependencies by determining whether the local expected conditional covariances $\rho_{P,t,n,m}(u)$ deviate from zero *at any point in time* for any index $m \in \mathcal{M}_n$. As in the previous setting, the test statistic for the locally stationary time series setting is based on the products of residuals from the time-varying nonlinear regressions of $(X_{t,n,i,a})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ and $(Y_{t,n,j,b})_{t\in\mathcal{T}_n}$ on $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively.

With the obvious changes in notation, the same arguments from Subsection C.2 show that conditional independence testing is hard. Consequently, we must restrict the null hypothesis. The next subsections introduce additional concepts and assumptions, and in we introduce our testing procedure in Subsection D.8.

D.5 The conditional covariance process

We will now introduce the causal representations of the locally stationary error processes from Subsection D.4.

Assumption 10 (Causal representations of the error processes)**.** *We assume that the error processes from Subsection D.4 have the following causal representations. For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, $t \in \mathcal{T}_n$, we can represent the error processes as*

$$\varepsilon_{P,t,n,i,a} = \tilde{G}^{\varepsilon}_{P,n,i,a}\left(\frac{t}{n}, \mathcal{H}^{\varepsilon}_{t,a}\right) = X_{t,n,i,a} - \mathbb{E}_P(X_{t,n,i,a}|\boldsymbol{Z}_{t,n}),$$

$$\xi_{P,t,n,j,b} = \tilde{G}^{\xi}_{P,n,j,b}\left(\frac{t}{n}, \mathcal{H}^{\xi}_{t,b}\right) = Y_{t,n,j,b} - \mathbb{E}_P(Y_{t,n,j,b}|\boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}^{\varepsilon}_{t,a} = (\eta^{\varepsilon}_{t,a}, \eta^{\varepsilon}_{t,a-1}, \ldots)$, $\mathcal{H}^{\xi}_{t,b} = (\eta^{\xi}_{t,b}, \eta^{\xi}_{t,b-1}, \ldots)$ and $(\eta^{\varepsilon}_{t,a})_{t\in\mathbb{Z}}$, $(\eta^{\xi}_{t,b})_{t\in\mathbb{Z}}$ are sequences of iid random elements. In particular, $\eta^{\varepsilon}_{t,a} = (\eta^{X}_{t+a}, \eta^{Z}_{t})'$, $\eta^{\xi}_{t,b} = (\eta^{Y}_{t+b}, \eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$ so that the error processes each depend on the inputs for the covariate processes and their respective response process. Also, we have $\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\mathcal{H}^{Z}_{t}) = 0$ and $\mathbb{E}_P(\xi_{P,t,n,j,b}|\mathcal{H}^{Z}_{t}) = 0$. For a given $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $m \in \mathcal{M}_n$, we have that $\tilde{G}^{\varepsilon}_{P,n,i,a}(u, \cdot)$, $\tilde{G}^{\xi}_{P,n,j,b}(u, \cdot)$ are measurable functions

such that $\tilde{G}^{\varepsilon}_{P,n,i,a}(u,\mathcal{H}^{\varepsilon}_{s,a})$, $\tilde{G}^{\xi}_{P,n,j,b}(u,\mathcal{H}^{\xi}_{s,b})$ *are well-defined random variables for each* $s \in \mathbb{Z}$ *and* $(\tilde{G}^{\varepsilon}_{P,n,i,a}(u,\mathcal{H}^{\varepsilon}_{s,a}))_{s\in\mathbb{Z}}$, $(\tilde{G}^{\xi}_{P,n,j,b}(u,\mathcal{H}^{\xi}_{s,b}))_{s\in\mathbb{Z}}$ *are stationary ergodic time series. Also, denote*

$$\tilde{\varepsilon}_{P,t,n,i,a}(u) = \tilde{G}^{\varepsilon}_{P,n,i,a}\left(u,\mathcal{H}^{\varepsilon}_{t,a}\right), \tilde{\xi}_{P,t,n,j,b}(u) = \tilde{G}^{\xi}_{P,n,j,b}\left(u,\mathcal{H}^{\xi}_{t,b}\right),$$

so that we may write $\varepsilon_{P,t,n,i,a} = \tilde{\varepsilon}_{P,t,n,i,a}(t/n)$, $\xi_{P,t,n,j,b} = \tilde{\xi}_{P,t,n,j,b}(t/n)$.

Using the causal representations of error processes, we have the following causal representation of the high-dimensional nonstationary vector-valued error processes

$$\boldsymbol{\varepsilon}_{P,t,n} = \tilde{\boldsymbol{G}}^{\boldsymbol{\varepsilon}}_{P,n}\left(\frac{t}{n},\mathcal{H}^{\boldsymbol{\varepsilon}}_{t}\right) = (\tilde{G}^{\varepsilon}_{P,n,i,a}\left(\frac{t}{n},\mathcal{H}^{\varepsilon}_{t,a}\right))_{i\in[d_X],a\in A_i},$$

where $\mathcal{H}^{\boldsymbol{\varepsilon}}_{t} = (\eta^{\boldsymbol{\varepsilon}}_{t},\eta^{\boldsymbol{\varepsilon}}_{t-1},\ldots)$ and $\eta^{\boldsymbol{\varepsilon}}_{t} = (\eta^{X}_{t+a_{\max}},\eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$. We denote

$$\tilde{\boldsymbol{\varepsilon}}_{P,t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{\varepsilon}}_{P,n}\left(u,\mathcal{H}^{\boldsymbol{\varepsilon}}_{t}\right) = (\tilde{G}^{\varepsilon}_{P,n,i,a}\left(u,\mathcal{H}^{\varepsilon}_{t,a}\right))_{i\in[d_X],a\in A_i},$$

so that we may write $\boldsymbol{\varepsilon}_{P,t,n} = \tilde{\boldsymbol{\varepsilon}}_{P,t,n}(t/n)$. Similarly, for $\boldsymbol{\xi}_{P,t,n}$ we write

$$\boldsymbol{\xi}_{P,t,n} = \tilde{\boldsymbol{G}}^{\boldsymbol{\xi}}_{P,n}\left(\frac{t}{n},\mathcal{H}^{\boldsymbol{\xi}}_{t}\right) = (\tilde{G}^{\xi}_{P,n,j,b}\left(\frac{t}{n},\mathcal{H}^{\xi}_{t,b}\right))_{j\in[d_Y],b\in B_j},$$

where $\mathcal{H}^{\boldsymbol{\xi}}_{t} = (\eta^{\boldsymbol{\xi}}_{t},\eta^{\boldsymbol{\xi}}_{t-1},\ldots)$ and $\eta^{\boldsymbol{\xi}}_{t} = (\eta^{Y}_{t+b_{\max}},\eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$, and

$$\tilde{\boldsymbol{\xi}}_{P,t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{\xi}}_{P,n}\left(u,\mathcal{H}^{\boldsymbol{\xi}}_{t}\right) = (\tilde{G}^{\xi}_{P,n,j,b}\left(u,\mathcal{H}^{\xi}_{t,b}\right))_{j\in[d_Y],b\in B_j},$$

so that we may write $\boldsymbol{\xi}_{P,t,n} = \tilde{\boldsymbol{\xi}}_{P,t,n}(t/n)$.

Moreover, for each $m \in \mathcal{M}_n$ the process of error products can be represented as

$$R_{P,t,n,m} = \tilde{G}^{R}_{P,n,m}\left(\frac{t}{n},\mathcal{H}^{R}_{t,m}\right) = \tilde{G}^{\varepsilon}_{P,n,i,a}\left(\frac{t}{n},\mathcal{H}^{\varepsilon}_{t,a}\right)\tilde{G}^{\xi}_{P,n,j,b}\left(\frac{t}{n},\mathcal{H}^{\xi}_{t,b}\right),$$

where $\mathcal{H}^{R}_{t,m} = (\eta^{R}_{t,m},\eta^{R}_{t-1,m},\ldots)$ and $(\eta^{R}_{t,m})_{t\in\mathbb{Z}}$ is a sequence of iid random elements with $\eta^{R}_{t,m} = (\eta^{X}_{t+a},\eta^{Y}_{t+b},\eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$. We denote

$$\tilde{R}_{P,t,n,m}(u) = \tilde{G}^{R}_{P,n,m}\left(u,\mathcal{H}^{R}_{t,m}\right) = \tilde{G}^{\varepsilon}_{P,n,i,a}\left(u,\mathcal{H}^{\varepsilon}_{t,a}\right)\tilde{G}^{\xi}_{P,n,j,b}\left(u,\mathcal{H}^{\xi}_{t,b}\right),$$

so that we may write $R_{P,t,n,m} = \tilde{R}_{P,t,n,m}(t/n)$.

For each $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $n \in \mathbb{N}$, and $u \in \mathcal{U}_n$ we have the following causal representation of the high-dimensional nonstationary $\mathbb{R}^{M_n}$-valued process of all the products of errors

$$\boldsymbol{R}_{P,t,n} = \tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}\left(\frac{t}{n},\mathcal{H}^{\boldsymbol{R}}_{t}\right) = \left(\tilde{G}^{R}_{P,n,m}\left(\frac{t}{n},\mathcal{H}^{R}_{t,m}\right)\right)_{m\in\mathcal{M}_n}$$

where $\mathcal{H}^{\boldsymbol{R}}_{t} = (\eta^{\boldsymbol{R}}_{t},\eta^{\boldsymbol{R}}_{t-1},\ldots)$ and $\eta^{\boldsymbol{R}}_{t} = (\eta^{X}_{t+a_{\max}},\eta^{Y}_{t+b_{\max}},\eta^{Z}_{t})'$ for each $t \in \mathbb{Z}$. As with the causal representations of $\boldsymbol{\varepsilon}_{P,t,n}$ and $\boldsymbol{\xi}_{P,t,n}$, for a fixed $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, and $n \in \mathbb{N}$ we have that $\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u,\mathcal{H}^{\boldsymbol{R}}_{s})$ is a well-defined high-dimensional random vector for each $s \in \mathbb{Z}$ and $(\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u,\mathcal{H}^{\boldsymbol{R}}_{s}))_{s\in\mathbb{Z}}$ is a high-dimensional stationary ergodic $\mathbb{R}^{M_n}$-valued time series, where we denote

$$\tilde{\boldsymbol{R}}_{P,t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}\left(u,\mathcal{H}^{\boldsymbol{R}}_{t}\right) = \left(\tilde{G}^{R}_{P,n,m}\left(u,\mathcal{H}^{R}_{t,m}\right)\right)_{m\in\mathcal{M}_n},$$

so that we may write $\boldsymbol{R}_{P,t,n} = \tilde{\boldsymbol{R}}_{P,t,n}(t/n)$.

We will now define the local long-run covariance matrices and variances of the high-dimensional process of locally stationary error products.

Definition 3 (Local long-run covariance matrices and variances of process of error products)**.** *For each $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $n \in \mathbb{N}$, define the long-run covariance matrix of the $\mathbb{R}^{M_n}$-valued stationary process $(\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_t))_{t\in\mathbb{Z}}$ by*

$$\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,n}(u) = \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_0), \tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_h)).$$

Similarly, for each $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $n \in \mathbb{N}$, and $m \in \mathcal{M}_n$, denote the long-run variance of the $\mathbb{R}$-valued stationary process $(\tilde{G}^R_{P,n,m}(u, \mathcal{H}^R_t))_{t\in\mathbb{Z}}$ as

$$\Sigma^R_{P,n,m}(u) = \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\tilde{G}^R_{P,n,m}(u, \mathcal{H}^R_{0,m}), \tilde{G}^R_{P,n,m}(u, \mathcal{H}^R_{h,m})).$$

D.6 Prediction processes and adaptive learning algorithms

In this subsection, we make assumptions about the adaptive statistical learning algorithms used to construct the time-varying regression estimators and form the corresponding predictions. Again, the idea of causal representations for locally stationary functional time series was introduced in Delft [Del20] and Delft and Dette [DD24]. Here, we introduce causal representations of adaptive statistical learning algorithms.

In the following assumption we only focus on the causal representations for the times t/n. We do this to simplify the presentation, since we are only concerned with using these predictions to calculate the corresponding residuals. However, the predictions and the corresponding causal representations are well-defined for all rescaled times $u \in \mathcal{U}_n$. These predictions at any $u \in \mathcal{U}_n$ can be formed by utilizing kernel smoothing as in Yousuf and Ng [YN21], for example.

Assumption 11 (Adaptive statistical learning algorithms for predictions)**.** *For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$ let $\mathbb{M}_t(\mathcal{Y}, \mathcal{Z}_n) \subseteq \mathcal{Y}^{\mathcal{Z}_n}$ and $\mathbb{M}_t(\mathcal{X}, \mathcal{Z}_n) \subseteq \mathcal{X}^{\mathcal{Z}_n}$, where $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{Z}_n = \mathbb{R}^{\boldsymbol{dz}}$. For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$, let $\eta^{algo}_{t,n,i,a}$, $\eta^{algo}_{t,n,j,b}$ be random elements which encode the (possible) stochasticity of the adaptive statistical learning algorithms. If the learning algorithms are not stochastic then $\eta^{algo}_{t,n,i,a}$, $\eta^{algo}_{t,n,j,b}$ can be ignored without loss of generality. Denote the data used to construct $\hat{f}_{n,i,a}(t/n, \cdot)$ by $\mathcal{D}^{\hat{f}}_{t,n,i,a} = (X_{s,n,i,a}, \boldsymbol{Z}_{s,n})^t_{s=\mathbb{T}^-_n-\lambda_n}$ and similarly let $\mathcal{D}^{\hat{g}}_{t,n,j,b} = (Y_{s,n,j,b}, \boldsymbol{Z}_{s,n})^t_{s=\mathbb{T}^-_n-\lambda_n}$ be the data for $\hat{g}_{n,j,b}(t/n, \cdot)$, where $\lambda_n \geq 0$ is the optional late-starting parameter from Subsection D.1.*

For each $n \in \mathbb{N}$, $m \in \mathcal{M}_n$, we assume that each of the (potentially stochastic) adaptive statistical learning algorithms $\mathcal{A}^{\hat{f}}_{n,i,a}$, $\mathcal{A}^{\hat{g}}_{n,j,b}$ consists of a sequence of measurable functions $\mathcal{A}^{\hat{f}}_{n,i,a} = (\mathcal{A}^{\hat{f}}_{t,n,i,a})_{t\in\mathcal{T}_n}$, $\mathcal{A}^{\hat{g}}_{n,j,b} = (\mathcal{A}^{\hat{g}}_{t,n,j,b})_{t\in\mathcal{T}_n}$, where for each $t \in \mathcal{T}_n$ we have

$$\mathcal{A}^{\hat{f}}_{t,n,i,a} : (\mathcal{X} \times \mathcal{Z}_n)^{t-\mathbb{T}^-_n+\lambda_n+1} \times [0,1] \to \mathbb{M}_t(\mathcal{X}, \mathcal{Z}_n)$$

$$\mathcal{D}^{\hat{f}}_{t,n,i,a} \times \eta^{algo}_{t,n,i,a} \mapsto \hat{f}_{n,i,a}(t/n, \cdot),$$

and

$$\mathcal{A}^{\hat{g}}_{t,n,j,b} : (\mathcal{Y} \times \mathcal{Z}_n)^{t-\mathbb{T}^-_n+\lambda_n+1} \times [0,1] \to \mathbb{M}_t(\mathcal{Y}, \mathcal{Z}_n)$$

$$\mathcal{D}^{\hat{g}}_{t,n,j,b} \times \eta^{algo}_{t,n,j,b} \mapsto \hat{g}_{n,j,b}(t/n, \cdot).$$

In view of the causal representation of the observed processes from Assumption 9, the adaptive statistical learning algorithms have the causal representations

$$\mathcal{A}^{\hat{f}}_{t,n,i,a} = \tilde{G}^{\mathcal{A}^{\hat{f}}}_{n,i,a}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{f}}}_{t,a}\right),$$

$$\mathcal{A}^{\hat{g}}_{t,n,j,b} = \tilde{G}^{\mathcal{A}^{\hat{g}}}_{n,j,b}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{g}}}_{t,b}\right),$$

where $\tilde{G}^{\mathcal{A}^{\hat{f}}}_{n,i,a}(t/n, \cdot)$, $\tilde{G}^{\mathcal{A}^{\hat{g}}}_{n,j,b}(t/n, \cdot)$ are measurable functions such that $\tilde{G}^{\mathcal{A}^{\hat{f}}}_{n,i,a}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{f}}}_{s,a}\right)$, $\tilde{G}^{\mathcal{A}^{\hat{g}}}_{n,j,b}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{g}}}_{s,b}\right)$ are well-defined function-valued random variables for each $s \in \mathbb{Z}$ and $(\tilde{G}^{\mathcal{A}^{\hat{f}}}_{n,i,a}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{f}}}_{s,a}\right))_{s\in\mathbb{Z}}$,

$(\tilde{G}^{\mathcal{A}^{\hat{g}}}_{n,j,b}\left(t/n, \mathcal{H}^{\mathcal{A}^{\hat{g}}}_{s,b}\right))_{s\in\mathbb{Z}}$ *are stationary ergodic functional time series. The input sequences are* $\mathcal{H}^{\mathcal{A}^{\hat{f}}}_{t,a} = (\eta^{\mathcal{A}^{\hat{f}}}_{t,a}, \eta^{\mathcal{A}^{\hat{f}}}_{t-1,a}, \ldots)$, $\mathcal{H}^{\mathcal{A}^{\hat{g}}}_{t,b} = (\eta^{\mathcal{A}^{\hat{g}}}_{t,b}, \eta^{\mathcal{A}^{\hat{g}}}_{t-1,b}, \ldots)$ *where* $(\eta^{\mathcal{A}^{\hat{f}}}_{t,a})_{t\in\mathbb{Z}}$, $(\eta^{\mathcal{A}^{\hat{g}}}_{t,b})_{t\in\mathbb{Z}}$ *are sequences of iid random elements with* $\eta^{\mathcal{A}^{\hat{f}}}_{t,a} = (\eta^{X}_{t+a}, \eta^{Z}_{t})'$ *and* $\eta^{\mathcal{A}^{\hat{g}}}_{t,b} = (\eta^{Y}_{t+b}, \eta^{Z}_{t})'$.

We have the following causal representations for all dimensions and time-offsets of the statistical learning algorithms

$$\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}_{t,n} = \tilde{\boldsymbol{G}}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{n}\left(t/n, \mathcal{H}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{t}\right) = (\mathcal{A}^{\hat{f}}_{t,n,i,a})_{i\in[d_X], a\in A_i},$$

$$\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}_{t,n} = \tilde{\boldsymbol{G}}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{n}\left(t/n, \mathcal{H}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{t}\right) = (\mathcal{A}^{\hat{g}}_{t,n,j,b})_{j\in[d_Y], b\in B_j},$$

where $\mathcal{H}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{t} = (\eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{t}, \eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{t-1}, \ldots)$ with $\eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{f}}}}_{t} = (\eta^{X}_{t+a_{\max}}, \eta^{Z}_{t})'$ and $\mathcal{H}^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{t} = (\eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{t}, \eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{t-1}, \ldots)$ with $\eta^{\boldsymbol{\mathcal{A}}^{\hat{\boldsymbol{g}}}}_{t} = (\eta^{Y}_{t+b_{\max}}, \eta^{Z}_{t})'$. Again, the adaptive statistical learning algorithms used to construct the predictors at time t only uses the covariates up to time t. Note that we have suppressed the dependence of the predictors on $\eta^{\text{algo}}_{t,n,i,a}$, $\eta^{\text{algo}}_{t,n,j,b}$ for the sake of notational simplicity.

In view of the previous assumption, we discuss the causal representations for the predictions and prediction errors.

Assumption 12 (Causal representations for predictions and prediction errors)**.** *We assume that the statistical learning algorithms are measurable such that we can represent the predictions for each* $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$ *as*

$$\hat{f}_{n,i,a}(t/n, \boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{f}}_{n,i,a}\left(t/n, \mathcal{H}^{\hat{f}}_{t,a}\right) = [\mathcal{A}^{\hat{f}}_{t,n,i,a}(\mathcal{D}^{\hat{f}}_{t,n,i,a}, \eta^{algo}_{t,n,i,a})](\boldsymbol{Z}_{t,n}),$$

$$\hat{g}_{n,j,b}(t/n, \boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{g}}_{n,j,b}\left(t/n, \mathcal{H}^{\hat{g}}_{t,b}\right) = [\mathcal{A}^{\hat{g}}_{t,n,j,b}(\mathcal{D}^{\hat{g}}_{t,n,j,b}, \eta^{algo}_{t,n,j,b})](\boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}^{\hat{f}}_{t,a} = (\eta^{\hat{f}}_{t,a}, \eta^{\hat{f}}_{t-1,a}, \ldots)$ *with* $\eta^{\hat{f}}_{t,a} = (\eta^{X}_{t+a}, \eta^{Z}_{t})'$ *and* $\mathcal{H}^{\hat{g}}_{t,b} = (\eta^{\hat{g}}_{t,b}, \eta^{\hat{g}}_{t-1,b}, \ldots)$ *with* $\eta^{\hat{g}}_{t,b} = (\eta^{Y}_{t+b}, \eta^{Z}_{t})'$, *and that we can represent the prediction errors as*

$$\hat{w}^{f}_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{w}^f}_{P,n,i,a}\left(t/n, \mathcal{H}^{\hat{w}^f}_{t,a}\right) = f_{P,n,i,a}(t/n, \boldsymbol{Z}_{t,n}) - \hat{f}_{n,i,a}(t/n, \boldsymbol{Z}_{t,n}),$$

$$\hat{w}^{g}_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{w}^g}_{P,n,j,b}\left(t/n, \mathcal{H}^{\hat{w}^g}_{t,b}\right) = g_{P,n,j,b}(t/n, \boldsymbol{Z}_{t,n}) - \hat{g}_{n,j,b}(t/n, \boldsymbol{Z}_{t,n}),$$

where $\mathcal{H}^{\hat{w}^f}_{t,a} = (\eta^{\hat{w}^f}_{t,a}, \eta^{\hat{w}^f}_{t-1,a}, \ldots)$ *with* $\eta^{\hat{w}^f}_{t,a} = (\eta^{X}_{t+a}, \eta^{Z}_{t})'$ *and* $\mathcal{H}^{\hat{w}^g}_{t,b} = (\eta^{\hat{w}^g}_{t,b}, \eta^{\hat{w}^g}_{t-1,b}, \ldots)$ *with* $\eta^{\hat{w}^g}_{t,b} = (\eta^{Y}_{t+b}, \eta^{Z}_{t})'$.

As usual, $\tilde{G}^{\hat{f}}_{n,i,a}(t/n, \cdot)$, $\tilde{G}^{\hat{g}}_{n,j,b}(t/n, \cdot)$ *and* $\tilde{G}^{\hat{w}^f}_{P,n,i,a}(t/n, \cdot)$, $\tilde{G}^{\hat{w}^g}_{P,n,j,b}(t/n, \cdot)$ *are measurable functions such that* $\tilde{G}^{\hat{f}}_{n,i,a}\left(t/n, \mathcal{H}^{\hat{f}}_{s,a}\right)$, $\tilde{G}^{\hat{g}}_{n,j,b}\left(t/n, \mathcal{H}^{\hat{g}}_{s,b}\right)$ *and* $\tilde{G}^{\hat{w}^f}_{P,n,i,a}\left(t/n, \mathcal{H}^{\hat{w}^f}_{s,a}\right)$, $\tilde{G}^{\hat{w}^g}_{P,n,j,b}\left(t/n, \mathcal{H}^{\hat{w}^g}_{s,b}\right)$ *are well-defined real-valued random variables for each* $s \in \mathbb{Z}$ *and* $(\tilde{G}^{\hat{f}}_{n,i,a}\left(t/n, \mathcal{H}^{\hat{f}}_{s,a}\right))_{s\in\mathbb{Z}}$, $(\tilde{G}^{\hat{g}}_{n,j,b}\left(t/n, \mathcal{H}^{\hat{g}}_{s,b}\right))_{s\in\mathbb{Z}}$ *and* $\left(\tilde{G}^{\hat{w}^f}_{P,n,i,a}\left(t/n, \mathcal{H}^{\hat{w}^f}_{s,a}\right)\right)_{s\in\mathbb{Z}}$, $\left(\tilde{G}^{\hat{w}^g}_{P,n,j,b}\left(t/n, \mathcal{H}^{\hat{w}^g}_{s,b}\right)\right)_{s\in\mathbb{Z}}$ *are real-valued stationary ergodic time series.*

Similarly, we have the following causal representation for all dimensions and time-offsets of the prediction errors

$$\hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n}(\boldsymbol{Z}_{t,n}) = \tilde{\boldsymbol{G}}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{P,n}\left(t/n, \mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{t}\right) = (\hat{w}^{f}_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}))_{i\in[d_X], a\in A_i},$$

$$\hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n}(\boldsymbol{Z}_{t,n}) = \tilde{\boldsymbol{G}}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\boldsymbol{P},\boldsymbol{n}}\left(t/n, \mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{t}\right) = (\hat{w}^{g}_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}))_{j\in[d_Y], b\in B_j},$$

where $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{t} = (\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{t}, \eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{t-1}, \ldots)$ with $\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{t} = (\eta^{X}_{t+a_{\max}}, \eta^{Z}_{t})'$ and $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{t} = (\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{t}, \eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{t-1}, \ldots)$ with $\eta^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{t} = (\eta^{Y}_{t+b_{\max}}, \eta^{Z}_{t})'$. Also, we can represent the products of the errors and prediction errors as

$$\hat{w}^{g,\varepsilon}_{P,t,n,m}(\boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{w}^{g,\varepsilon}}_{P,n,m}\left(t/n, \mathcal{H}^{\hat{w}^{g,\varepsilon}}_{t,m}\right) = \hat{w}^{g}_{P,t,n,j,b}(\boldsymbol{Z}_{t,n})\varepsilon_{P,t,n,i,a},$$

$$\hat{w}^{f,\xi}_{P,t,n,m}(\boldsymbol{Z}_{t,n}) = \tilde{G}^{\hat{w}^{f,\xi}}_{P,n,m}\left(t/n, \mathcal{H}^{\hat{w}^{f,\xi}}_{t,m}\right) = \hat{w}^{f}_{P,t,n,i,a}(\boldsymbol{Z}_{t,n})\xi_{P,t,n,j,b},$$

where $\mathcal{H}_{t,m}^{\hat{w}^{g,\varepsilon}} = (\eta_{t,m}^{\hat{w}^{g,\varepsilon}}, \eta_{t-1,m}^{\hat{w}^{g,\varepsilon}}, \ldots)$ with $\eta_{t,m}^{\hat{w}^{g,\varepsilon}} = (\eta_{t+a}^X, \eta_{t+b}^Y, \eta_t^Z)'$ and $\mathcal{H}_{t,m}^{\hat{w}^{f,\xi}} = (\eta_{t,m}^{\hat{w}^{f,\xi}}, \eta_{t-1,m}^{\hat{w}^{f,\xi}}, \ldots)$ with $\eta_{t,m}^{\hat{w}^{f,\xi}} = (\eta_{t+a}^X, \eta_{t+b}^Y, \eta_t^Z)'$. Putting it all together, we have the following causal representation for all dimensions and time-offsets of the products of the errors and prediction errors

$$\hat{\boldsymbol{w}}_{P,t,n}^{\boldsymbol{g},\boldsymbol{\varepsilon}}(\boldsymbol{Z}_{t,n}) = \tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\left(t/n, \mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\right) = (\hat{w}_{P,t,n,m}^{g,\varepsilon}(\boldsymbol{Z}_{t,n}))_{m\in\mathcal{M}_n},$$

$$\hat{\boldsymbol{w}}_{P,t,n}^{\boldsymbol{f},\boldsymbol{\xi}}(\boldsymbol{Z}_{t,n}) = \tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\left(t/n, \mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\right) = (\hat{w}_{P,t,n,m}^{f,\xi}(\boldsymbol{Z}_{t,n}))_{m\in\mathcal{M}_n},$$

where $\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}} = (\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}, \eta_{t-1}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}, \ldots)$ with $\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}} = (\eta_{t+a_{\max}}^X, \eta_{t+b_{\max}}^Y, \eta_t^Z)'$ and $\mathcal{H}_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}} = (\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}, \eta_{t-1}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}, \ldots)$ with $\eta_t^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}} = (\eta_{t+a_{\max}}^X, \eta_{t+b_{\max}}^Y, \eta_t^Z)'$. We emphasize that $\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}(t/n, \cdot)$, $\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}(t/n, \cdot)$ are measurable functions such that $\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\left(t/n, \mathcal{H}_s^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\right)$, $\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\left(t/n, \mathcal{H}_s^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\right)$ are well-defined high-dimensional random vectors for each $s \in \mathbb{Z}$ and $(\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\left(t/n, \mathcal{H}_s^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}\right))_{s\in\mathbb{Z}}$, $(\tilde{\boldsymbol{G}}_{P,n}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\left(t/n, \mathcal{H}_s^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}\right))_{s\in\mathbb{Z}}$ are high-dimensional stationary ergodic time series.

D.7 Distribution-uniform functional dependence measure and control of nonstationarity

In this subsection, we introduced distribution-uniform regularity conditions for the assumed causal representations for the high-dimensional nonlinear locally stationary processes that we discussed previously.

Denote the set of well-defined tuples of error processes, dimensions, and time-offsets by

$$\mathbb{E} = \{(\varepsilon, i, a) : i \in [d_X], a \in A_i\} \cup \{(\xi, j, b) : j \in [d_Y], b \in B_j\},$$

so that we may write $(e, l, d) \in \mathbb{E}$ to refer to any such combination.

Again, we quantify temporal dependence using the functional dependence measure of Wu [Wu05] and we impose a uniform polynomial decay of the temporal dependence.

Definition 4 (Functional dependence measure)**.** *For any tuple $(e, l, d) \in \mathbb{E}$ corresponding to a well-defined combination of an error process, dimension, time-offset, let $(\tilde{\eta}_{t,d}^e)_{t\in\mathbb{Z}}$ be an iid copy of $(\eta_{t,d}^e)_{t\in\mathbb{Z}}$. Define*

$$\tilde{\mathcal{H}}_{t,d,t-h}^e = (\eta_{t,d}^e, \ldots, \eta_{t-h+1,d}^e, \tilde{\eta}_{t-h,d}^e, \eta_{t-h-1,d}^e, \ldots)$$

to be $\mathcal{H}_{t,d}^e$ with the $(t-h)$-th element $\eta_{t-h,d}^e$ replaced with $\tilde{\eta}_{t-h,d}^e$. Analogously, for $\boldsymbol{e} \in \{\boldsymbol{\varepsilon}, \boldsymbol{\xi}\}$ define $\tilde{\mathcal{H}}_{t,t-h}^{\boldsymbol{e}}$ as $\mathcal{H}_t^{\boldsymbol{e}}$ with the $(t-h)$-th input $\eta_{t-h}^{\boldsymbol{e}}$ replaced with $\tilde{\eta}_{t-h}^{\boldsymbol{e}}$ as in Subsection D.5. Similarly, for the product of errors R define $\tilde{\mathcal{H}}_{t,m,t-h}^R$ as $\mathcal{H}_{t,m}^R$ with the $(t-h)$-th element $\eta_{t-h,m}^R$ replaced with the iid copy $\tilde{\eta}_{t-h,m}^R$. Analogously, define $\tilde{\mathcal{H}}_{t,t-h}^{\boldsymbol{R}}$ as $\mathcal{H}_t^{\boldsymbol{R}}$ with the $(t-h)$-th input $\eta_{t-h}^{\boldsymbol{R}}$ replaced with $\tilde{\eta}_{t-h}^{\boldsymbol{R}}$ as in Subsection D.5.

Define L^∞ versions of the functional dependence measure for the error processes $\tilde{G}_{P,n,l,d}^e(u, \mathcal{H}_{t,d}^e)$ for $(e, l, d) \in \mathbb{E}$, $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $t \in \mathcal{T}_n$ for $h \in \mathbb{N}_0$, as

$$\theta_{P,u,t,n,l,d}^{e,\infty}(h) = \inf\{K \geq 0 : \mathbb{P}_P(|\tilde{G}_{P,n,l,d}^e(u, \mathcal{H}_{t,d}^e) - \tilde{G}_{P,n,l,d}^e(u, \tilde{\mathcal{H}}_{t,d,t-h}^e)| > K) = 0\},$$

and for the vector-valued $\tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{e}}(u, \mathcal{H}_t^{\boldsymbol{e}}))$ for $\boldsymbol{e} \in \{\boldsymbol{\varepsilon}, \boldsymbol{\xi}\}$, $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $t \in \mathcal{T}_n$ for $h \in \mathbb{N}_0$, $r \geq 1$ as

$$\theta_{P,u,t,n}^{\boldsymbol{e},\infty}(h, r) = \inf\{K \geq 0 : \mathbb{P}_P(||\tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{e}}(u, \mathcal{H}_t^{\boldsymbol{e}}) - \tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{e}}(u, \tilde{\mathcal{H}}_{t,t-h}^{\boldsymbol{e}})||_r > K) = 0\}.$$

Define the functional dependence measures for $\tilde{G}_{P,n,m}^R(u, \mathcal{H}_{t,m}^R)$ for $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $t \in \mathcal{T}_n$, $m \in \mathcal{M}_n$ for $h \in \mathbb{N}_0$, $q \geq 1$ as

$$\theta_{P,u,t,n,m}^R(h, q) = [\mathbb{E}_P(|\tilde{G}_{P,n,m}^R(u, \mathcal{H}_{t,m}^R) - \tilde{G}_{P,n,m}^R(u, \tilde{\mathcal{H}}_{t,m,t-h}^R)|^q)]^{1/q},$$

and for the vector-valued $\tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{R}}(u, \mathcal{H}_t^{\boldsymbol{R}})$ for $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $t \in \mathcal{T}_n$ for $h \in \mathbb{N}_0$, $q \geq 1$, $r \geq 1$ as

$$\theta_{P,u,t,n}^{\boldsymbol{R}}(h, q, r) = [\mathbb{E}_P(||\tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{R}}(u, \mathcal{H}_t^{\boldsymbol{R}}) - \tilde{\boldsymbol{G}}_{P,n}^{\boldsymbol{R}}(u, \tilde{\mathcal{H}}_{t,t-h}^{\boldsymbol{R}})||_r^q)]^{1/q}.$$

We impose the following regularity conditions to control the temporal dependence and nonstationarity uniformly over a collection of distributions $\mathcal{P}_n$. We impose these conditions on each dimension of the processes so that it is easy to verify the conditions even in high-dimensional settings. First, we introduce an assumption imposing a uniform polynomial decay of the temporal dependence. Note that we will often write the time as 0 when the time of the input sequence does not matter because of stationarity.

Assumption 13 (Distribution-uniform decay of temporal dependence)**.** *We assume that there exist $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 2$ such that for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $(e, l, d) \in \mathbb{E}$, it holds that*

$$\sup_{P \in \mathcal{P}_n} ||\tilde{G}^{e}_{P,n,l,d}(u, \mathcal{H}^{e}_{0,d})||_{L^{\infty}(P)} \leq \bar{\Theta}^{\infty}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{e,\infty}_{P,u,0,n,l,d}(h) \leq \bar{\Theta}^{\infty} \cdot (h \vee 1)^{-\bar{\beta}^{\infty}}, \qquad h \geq 0.$$

For additional control in terms of the product of errors alone, we also assume that there exist $\bar{\Theta}^{R} > 0$, $\bar{\beta}^{R} > 2$, $\bar{q}^{R} > 4$, such that for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $m \in \mathcal{M}_n$, it holds that

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(|\tilde{G}^{R}_{P,n,m}(u, \mathcal{H}^{R}_{0,m})|^{\bar{q}^{R}})]^{1/\bar{q}^{R}} \leq \bar{\Theta}^{R}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{R}_{P,u,0,n,m}(h, \bar{q}^{R}) \leq \bar{\Theta}^{R} \cdot (h \vee 1)^{-\bar{\beta}^{R}}, \qquad h \geq 0.$$

Note that in the previous assumption, it is possible to further upper bound the functional dependence measures and moment bounds for the product of errors in terms of $\bar{\Theta}^{\infty}$ by using the triangle inequality and Hölder's inequality. Also, note that the constants in the previous assumption do not depend on n.

By Jensen's inequality, we have the following functional dependence measures for the corresponding vector-valued processes.

Recall $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 2$. For all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $\boldsymbol{e} \in (\boldsymbol{\varepsilon}, \boldsymbol{\xi})$, we have that

$$\sup_{P \in \mathcal{P}_n} \left|\left|||\tilde{\boldsymbol{G}}^{\boldsymbol{e}}_{P,n}(u, \mathcal{H}^{\boldsymbol{e}}_{0})||_2\right|\right|_{L^{\infty}(P)} \leq M_n^{\frac{1}{2}} \bar{\Theta}^{\infty}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{e},\infty}_{P,u,0,n}(h, 2) \leq M_n^{\frac{1}{2}} \bar{\Theta}^{\infty} \cdot (h \vee 1)^{-\bar{\beta}^{\infty}}, \qquad h \geq 0.$$

Recall $\bar{\Theta}^{R} > 0$, $\bar{\beta}^{R} > 2$, $\bar{q}^{R} > 4$. For all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, it holds that

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_{0})||_2^{\bar{q}^{R}})]^{1/\bar{q}^{R}} \leq M_n^{\frac{1}{2}} \bar{\Theta}^{R}, \qquad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{R}}_{P,u,0,n}(h, \bar{q}^{R}, 2) \leq M_n^{\frac{1}{2}} \bar{\Theta}^{R} \cdot (h \vee 1)^{-\bar{\beta}^{R}}, \qquad h \geq 0.$$

We impose the following regularity conditions to control the nonstationarity uniformly over a collection of distributions $\mathcal{P}_n$.

Assumption 14 (Distribution-uniform stochastic Lipschitz condition)**.** *For additional control in terms of the product of errors alone, we also assume that there exist $\bar{\Theta}^{R} > 0$, $\bar{q}^{R} \geq 2$, such that for all $n \in \mathbb{N}$, $u, v \in \mathcal{U}_n$, $m \in \mathcal{M}_n$ for some $\bar{L}^{R} > 0$ it holds that*

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(|\tilde{G}^{R}_{P,n,m}(u, \mathcal{H}^{R}_{0,m}) - \tilde{G}^{R}_{P,n,m}(v, \mathcal{H}^{R}_{0,m})|^{\bar{q}^{R}})]^{1/\bar{q}^{R}} \leq \bar{L}^{R} \bar{\Theta}^{R} |u - v|.$$

To use the test statistic introduced in Subsection 2.4 in the locally stationary setting, it suffices to show that the total variation of the causal mechanism from Assumption 6 in Subsection B.3 can be bounded distribution-uniformly. Indeed, this is straightforward to show by directly applying the stochastic Lipschitz condition. As we discuss in Section E, we can utilize the distribution-uniform strong Gaussian approximation from Lemma E.1 in the locally stationary setting because this result is formulated in terms of triangular arrays.

Again, to avoid specific conditions on the adaptive statistical learning algorithm, we make the following assumption about the decay in temporal dependence of the cumulative squared prediction errors.

Assumption 15 (Distribution-uniform decay of temporal dependence for prediction error)**.** *As in the discussion following Assumption 12, let $\tilde{\mathcal{H}}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t,t-h}$ be $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t}$ with the $(t-h)$-th input replaced with its iid copy. Assume that there exist $\bar{\Theta}^{\diamond} > 0$, $\bar{\beta}^{\diamond} > 2$ such that for all $\boldsymbol{\kappa} \in (\boldsymbol{f}, \boldsymbol{g})$, it holds that*

$$\sum_{t \in \mathcal{T}_n} \mathbb{E}_P[||\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{P,n}(t/n, \mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t}) - \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{P,n}(t/n, \tilde{\mathcal{H}}^{\hat{\boldsymbol{w}}^{\boldsymbol{\kappa}}}_{t,t-h})||_2^2] \leq \tau_n^2 T_n M_n^{-1} \bar{\Theta}^{\diamond} (h \vee 1)^{-\bar{\beta}^{\diamond}}.$$

That is, the cumulative distance between the original squared prediction errors and the "perturbed" version with the $(t-h)$-th input replaced by its iid copy will decrease as h increases (i.e. through the historical training data for the adaptive predictors and for the covariate at time t). Note that in Assumption 16, we assume that the cumulative squared prediction errors grow sublinearly.

D.8 Our practical test in the locally stationary setting

Let

$$\hat{\boldsymbol{R}}_{t,n} = (\hat{R}_{t,n,m})_{m \in \mathcal{M}_n},$$

be the high-dimensional vector process with all indices of the products of residuals, and let $\hat{\boldsymbol{R}}_n$ denote the entire process of residuals. We use the same covariance estimation approach as the test from Section 2. The estimator for the local long-run covariance matrix $\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,n}(t/n)$ is given by

$$\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_n(t/n) = \hat{Q}^{\boldsymbol{R}}_{t,n} - \hat{Q}^{\boldsymbol{R}}_{t-1,n},$$

where $\hat{Q}^{\boldsymbol{R}}_{k,n}$ is the same cumulative covariance estimator from Subsection 2.4 and denote $\hat{Q}^{\boldsymbol{R}}_n = (\hat{Q}^{\boldsymbol{R}}_{t,n})_{t=L_n+\mathbb{T}_n^- - 1}^{\mathbb{T}_n^+}$.

The following result shows that if the time-varying regression estimators satisfy modest time-uniform and distribution-uniform convergence rate requirements, then the quantile $a_\alpha(\hat{Q}^{\boldsymbol{R}}_n)$ closely approximates the $(1-\alpha)$ quantile of the test statistic. Hence, we can use it to calibrate a test.

It is possible to use the same test statistic from Subsection 2.4, but in the locally stationary setting. The essential difference is that this test appeals to infill asymptotics and uses time-varying regression estimators for locally stationary time series.

Define $\hat{\boldsymbol{R}}_n = (\hat{\boldsymbol{R}}_{t,n})_{t=L_n+\mathbb{T}_n^- - 1}^{\mathbb{T}_n^+}$. Let

$$S_{n,p}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- + L_n - 1, \ldots, \mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t \leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_p$$

be the test statistic based on the ℓ_p norm for some $p \geq 2$. For instance, we can use

$$S_{n,\infty}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- + L_n - 1, \ldots, \mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t \leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_\infty,$$

or

$$S_{n,2}(\hat{\boldsymbol{R}}_n) = \max_{s=\mathbb{T}_n^- + L_n - 1, \ldots, \mathbb{T}_n^+} \left\| \frac{1}{\sqrt{T_n}} \sum_{t \leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_2.$$

The same bootstrap-based testing procedure from Subsection 2.4 is used for the locally stationary setting. Define the offsets ν_n and τ_n in the same way as Subsection 2.4.

Assumption 16 (Doubly robust rate requirements for predictors)**.** *For $n \in \mathbb{N}$ and some collection of distributions $\mathcal{P}_n$, assume that the predictors satisfy*

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left\| \hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right\|_2^2 \right) \sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left\| \hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right\|_2^2 \right) \right) = o(T_n \tau_n^2),$$

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left\| \hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right\|_2^2 \right) \right) = o(T_n M_n^{-1} \tau_n^2),$$

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left\| \hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n}(\boldsymbol{Z}_{t,n}) \right\|_2^2 \right) \right) = o(T_n M_n^{-1} \tau_n^2).$$

In practice, these forecasts can be made using any forecasting model as discussed in Subsection 2.5. However, there has also been work theoretically analyzing the statistical guarantees of certain methods. Notably, Yousuf and Ng [YN21] studied forecasting in the high-dimensional locally stationary time series setting using L_2 boosting methods with one-sided kernels. In principle, many of the existing time-varying regression estimators based on kernel smoothing (see Zhang and Wu [ZW15] and Vogt [Vog12] and the references therein) can be used with one-sided kernels for forecasting locally stationary processes.

We will calculate the cumulative covariance process $\hat{Q}_n^{\boldsymbol{R}}$ based on the process of products of residuals. We can approximate the quantile $a_\alpha(\hat{Q}_n^{\boldsymbol{R}})$ via Monte Carlo and we reject the null hypothesis of conditional independence at level α if $S_n(\hat{\boldsymbol{R}}_n) > a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}}) + \tau_n$ for some offsets $\nu_n \to 0$ and $\tau_n \to 0$ defined in Subsection D.6.

Theorem D.1. *Suppose that Assumptions 9, 10, 11, 12, 13, 14, 15, 16 all hold for the collection of distributions $\mathcal{P}_{0,n}^* \subset \mathcal{P}_{0,n}^{CI}$ and the predictors. If the offsets $\tau_n \to 0$ and $\nu_n \to 0$ are chosen such that condition (7) holds, we have that*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_{0,n}^*} \mathbb{P}_P\left(S_{n,p}(\hat{\boldsymbol{R}}_n) > a_{\alpha-\nu_n}(\hat{Q}_n^{\boldsymbol{R}}) + \tau_n\right) \leq \alpha.$$

Proof of Theorem D.1: The result follows by the exact same steps in the proof of Theorem 2.1 with the obvious changes in notation for the locally stationary setting. This is because the results from Section E can be applied to any triangular array framework for high-dimensional nonstationary time series, so the results hold for high-dimensional locally stationary time series in particular. Due to the similarity with the proof of Theorem 2.1, the details are omitted. □

It is also possible to develop a test for the *local* null hypothesis of conditional independence

$$\tilde{X}_{t,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{t,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{t,n}(u) \text{ for all } m \in \mathcal{M}_n$$

for some $u \in \mathcal{U}_n$ and $n \in \mathbb{N}$. As explained in Subsection D.4 we have

$$\rho_{P,t,n,m}(u) = \mathbb{E}_P(\tilde{R}_{P,t,n,m}(u)) = 0$$

under this local null hypothesis. The test statistic would require a kernel function and choosing a bandwidth parameter. The main idea is that for any $n \in \mathbb{N}$ and $m \in \mathcal{M}_n$, we can replace $R_{P,t,n,m}$ by $\tilde{R}_{P,t,n,m}(u)$ with a small approximation for any $t \in \mathcal{T}_n$ close to $u \in \mathcal{U}_n$ by using the stochastic Lipschitz condition from Assumption 14. Hence, the expected conditional covariances of the observed process near rescaled time $u \in \mathcal{U}_n$ must also be close to zero. Despite its simplicity, this locally stationary approximation for the expected conditional covariance has not yet been considered in the related literature. The proof makes use of *distribution-uniform* stationary approximations for locally stationary time series. This requires distribution-uniform extensions of many of the results from Dahlhaus, Richter, and Wu [DRW19], so we leave this for future work.

D.9 A note on simultaneous testing in the locally stationary setting

It is possible to develop simultaneous testing procedures to test for multiple rescaled times $u \in \mathcal{U}_n$ and indices $m \in \mathcal{M}_n$ based on the local test from the last section. However, we omit the details due to the similarity with Section A. Analogously to Theorem 9 in Shah and Peters [SP20], it is also possible to simultaneously test whether conditional independence holds at all times $u \in \mathcal{U}_n$ by creating simultaneous confidence bands for expected conditional covariance curves. The main idea is that the local null hypothesis of conditional independence at rescaled time $u \in \mathcal{U}_n$ for some index $m \in \mathcal{M}_n$ can be rejected if zero is not included in the corresponding confidence interval for the local expected conditional covariance $\rho_{P,t,n,m}(u)$. This can be done using similar arguments as Bai and Wu [BW23], which focuses on inferring time-varying correlation curves.

However, due to the problem of post-selection inference [KKK22], this would require either much stronger assumptions (e.g. Donsker-type), a sample splitting procedure for nonstationary time series which does not yet exist, or two independent realizations of the same nonstationary time series which is rarely possible outside of experimental settings. Hence, we leave this problem of inferring expected conditional covariance curves (and curves based on other functionals of interest in causal inference) for future work. This way, we can focus this manuscript on what can be done with just *one* realization of a nonstationary time series under relatively weak assumptions and without using any form of sample splitting.

An approach for inferring expected conditional covariance curves would have a range of applications outside of testing for conditional independence. The expected conditional covariance appears frequently in the literature on parameter estimation in causal inference [Ken22; Rob+08; Rob+09; Li+11; Rob+17; Rob18; Bic+93]. We suspect that similar approaches can be used to infer curves

based on other functionals of interest in causal inference. Hence, our work in this paper and our future work on inferring expected conditional covariance curves can be of independent interest for the emerging field of causal inference for time series [Sag+20; RGR22; Run+23b; Run+19b; Run18a; RS21; RS19; Bon+21].

E Distribution-Uniform Theory

E.1 Literature review of distribution-uniform inference

First, we discuss the CI testing literature. There has been a lot of recent work on distribution-uniform CI testing frameworks because of the hardness result and subsequent testing framework developed by Shah and Peters [SP20]. For instance, Lundborg, Shah, and Peters [LSP22] introduced many distribution-uniform convergence results for separable Banach and Hilbert spaces. Recently, Christgau, Petersen, and Hansen [CPH22] introduced a distribution-uniform "conditional local independence" testing framework for the setting where n realizations of a point process are observed. Christgau, Petersen, and Hansen [CPH22] also introduce a distribution-uniform extension of Rebolledo's martingale central limit theorem [Reb80] and extend many distribution-uniform convergence results from Lundborg, Shah, and Peters [LSP22] to metric spaces.

Second, we discuss relevant developments in the anytime-valid inference literature. Recently, Waudby-Smith and Ramdas [WR23] introduced a distribution-uniform strong (almost-sure) Gaussian approximation for the full sum of iid random variables, which appears to be the first such result in the literature. The work in Waudby-Smith and Ramdas [WR23] is motivated by prior work on asymptotic anytime-valid inference from Waudby-Smith et al. [Wau+21], in which the authors defined the concept of an "asymptotic confidence sequence" (AsympCS). Moreover, Waudby-Smith et al. [Wau+21] introduced an AsympCS for iid random variables and a Lindeberg-type AsympCS which can capture time-varying means under martingale dependence. To briefly compare, the result from Waudby-Smith and Ramdas [WR23] is a strong (almost-sure) Gaussian approximation for the full sum of iid random variables in a sequential setting, whereas our strong (in probability) Gaussian approximation is for the max of partial sums of a nonstationary high-dimensional random vector in a fixed-n setting.

Third, we mention other areas in which distribution-uniform inference is studied under different names. There is a vast literature discussing the importance of distribution-uniform inference under the name of "honest" or "uniform" inference, see Li [Li89], Kasy [Kas18], Tibshirani et al. [Tib+18], Rinaldo, Wasserman, and G'Sell [RWG19], and Kuchibhotla, Balakrishnan, and Wasserman [KBW23]. Also, there is a plethora of literature on distribution-uniform moment inequality testing Imbens and Manski [IM04], Romano and Shaikh [RS08], Andrews and Guggenberger [AG09], Andrews and Soares [AS10], Andrews and Barwick [AB12], and Romano, Shaikh, and Wolf [RSW14]. Most recently, Li, Liao, and Zhou [LLZ22] developed a distribution-uniform test for general functional inequalities which admits conditional moment inequalities as a special case. Also, in the supplemental appendix of Li, Liao, and Zhou [LLZ22] introduce a distribution-uniform strong Gaussian approximation for the full sum of a high-dimensional mixingale.

E.2 Distribution-uniform strong Gaussian approximation

We introduce a distribution-uniform extension of the strong Gaussian approximation for high-dimensional nonstationary nonlinear time series from Mies and Steland [MS22]. We will consider a process $(W_{t,n})_{t=1}^n$ taking values in $\mathbb{R}^{d_n}$. That being said, to explain our distribution-uniform results we must slightly change some notation. In some terms, we add subscripts P and n so that the dependence on the distribution and sample size explicit. Most results can be read straightforwardly as distribution-uniform extensions of the results as they were stated in Mies and Steland [MS22]. The notable exception is the strong Gaussian approximation, which requires slightly more care. We note that all of the results in this subsection are non-asymptotic and can be applied to triangular arrays.

For each sample size $n \in \mathbb{N}$, consider a $\mathbb{R}^{d_n}$-valued time series $(W_{t,n})_{t\in[n]}$ for some dimension $d_n \in \mathbb{N}$. Let $(\eta_i)_{i\in\mathbb{Z}}$, $(\tilde{\eta}_i)_{i\in\mathbb{Z}}$ be two iid sequences of random elements. Denote

$$\mathcal{H}_t = (\eta_t, \eta_{t-1}, \ldots),$$

$$\tilde{\mathcal{H}}_{t,j} = (\eta_t, \ldots, \eta_{j+1}, \tilde{\eta}_j, \eta_{j-1}, \ldots),$$

$$\bar{\mathcal{H}}_{t,j} = (\eta, \ldots, \eta_{j+1}, \tilde{\eta}_j, \tilde{\eta}_{j-1}, \ldots).$$

We assume that for each $n \in \mathbb{N}$ and $t \in [n]$, $W_{t,n}$ can be represented as a function of these iid random elements

$$W_{t,n} = G_{t,n}(\mathcal{H}_t),$$

where $G_{t,n} : \mathbb{R}^\infty \to \mathbb{R}^{d_n}$ is a measurable function (where we endow $\mathbb{R}^\infty$ with the σ-algebra generated by all finite projections) so that $G_{t,n}(\mathcal{H}_s)$ is a well-defined high-dimensional random vector for every $s \in \mathbb{Z}$ and $(G_{t,n}(\mathcal{H}_s))_{s\in\mathbb{Z}}$ is a high-dimensional stationary ergodic time series.

As discussed in in Examples 2 and 3 in Mies and Steland [MS22], locally stationary time series are a special case of this triangular array framework for nonstationary time series. For high-dimensional locally stationary time series, the causal representation

$$\tilde{W}_{t,n}(u) = \tilde{G}_n(u, \mathcal{H}_t),$$

is well-defined for all rescaled times $u \in [0, 1]$, so that we may write

$$W_{t,n} = G_{t,n}(\mathcal{H}_t) = \tilde{G}_n(t/n, \mathcal{H}_t) = \tilde{W}_{t,n}(t/n).$$

Let us briefly discuss some differences in the theoretical framework introduced in this section and those we introduced in Sections 2 and D. One of the main differences is that we use sequences of iid uniform random variables $\mathcal{H}_t$ (as in Mies and Steland [MS22]) in this section instead of sequences of iid random elements in the causal representations from Sections 2 and D. By standard results (for example, see Lemma 4.21, Lemma 4.22, and the previous discussion in section 4 of Kallenberg [Kal97]), it is possible to define the measurable functions $G_{t,n}(\cdot)$ in such a way that the causal representations used in the theoretical frameworks from Section 2 and Section D are special cases.

Let us introduce the setting rigorously. Let Ω be a sample space, $\mathcal{B}$ the Borel sigma-algebra, and $(\Omega, \mathcal{B})$ a measurable space. For fixed $n \in \mathbb{N}$, let $(\Omega, \mathcal{B})$ be equipped with a family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that the distribution of the stochastic system

$$(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}},$$

or, in the locally stationary setting,

$$(\tilde{G}_n(u, \mathcal{H}_s))_{u\in[0,1],s\in\mathbb{Z}},$$

under $\mathbb{P}_P$ is $P \in \mathcal{P}_n$ where $\mathcal{P}_n$ is a collection of such distributions. The family of probability measures $(\mathbb{P}_P)_{P\in\mathcal{P}_n}$ is defined with respect to the same measurable space $(\Omega, \mathcal{B})$, but need not have the same dominating measure. Denote a family of probability spaces by $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$ and a sequence of such families of probability spaces by $((\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n})_{n\in\mathbb{N}}$. When we say that the process $(W_{t,n})_{t\in[n]}$ is defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$ for some $n \in \mathbb{N}$, we mean that $(W_{t,n})_{t\in[n]}$ is defined on the probability space $(\Omega, \mathcal{B}, \mathbb{P}_P)$ for each $P \in \mathcal{P}_n$.

Next, we define our measure of temporal dependence for the process.

Definition 5 (Functional dependence measure)**.** *For $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, define the functional dependence measure of the process $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ as*

$$\theta_{P,t,n}(j, q, r) = (\mathbb{E}_P||G_{t,n}(\mathcal{H}_t) - G_{t,n}(\tilde{\mathcal{H}}_{t,t-j})||_r^q)^{\frac{1}{q}},$$

with $h \in \mathbb{N}_0$, $q \geq 1$, $r \geq 1$.

We make the following distribution-uniform assumptions on the temporal dependence and nonstationarity of $(W_{t,n})_{t\in[n]}$.

Assumption 17 (Distribution-uniform decay of temporal dependence)**.** *We assume that there exist $\beta > 0$, $q \geq 2$ and a constant $\Theta_n > 0$ for each $n \in \mathbb{N}$, such that for all times $t \in [n]$ it holds*

$$\sup_{P\in\mathcal{P}_n} \theta_{P,t,n}(j, q, r) \leq \Theta_n \cdot (j \vee 1)^{-\beta},$$

for $j \geq 0$, and that

$$\sup_{P\in\mathcal{P}_n} (\mathbb{E}_P||G_{t,n}(\mathcal{H}_0)||_2^q)^{1/q} \leq \Theta_n.$$

Assumption 18 (Distribution-uniform total variation condition for nonstationarity)**.** *Recall Θ_n from Assumption 17. For each $n \in \mathbb{N}$, we also assume that there exists some $\Gamma_n \geq 1$ such that it holds*

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t=2}^{n} (\mathbb{E}_P ||G_{t,n}(\mathcal{H}_0) - G_{t-1,n}(\mathcal{H}_0)||_2^2)^{\frac{1}{2}} \right) \leq \Gamma_n \cdot \Theta_n.$$

Note that the distribution-uniform assumptions stated in Subsection B.3 satisfy conditions 17 and 18. Thus, we can use the results from this subsection for our tests in Subsection 2.4. Similarly, our framework for high-dimensional locally stationary time series from Subsection D.7 is a special case of this framework and it also satisfies conditions 17 and 18.

Define the two rates

$$\chi(q, \beta) = \begin{cases} \frac{q-2}{6q-4}, & \beta \geq \frac{3}{2}, \\ \frac{(\beta-1)(q-2)}{q(4\beta-3)-2}, & \beta \in (1, \frac{3}{2}), \end{cases}$$

and

$$\xi(q, \beta) = \begin{cases} \frac{q-2}{6q-4}, & \beta \geq 3, \\ \frac{(\beta-2)(q-2)}{(4\beta-6)q-4}, & \frac{3+\frac{2}{q}}{1+\frac{2}{q}} < \beta < 3, \\ \frac{1}{2} - \frac{1}{\beta}, & 2 < \beta \leq \frac{3+\frac{2}{q}}{1+\frac{2}{q}}, \end{cases}$$

which will appear in the results in this section. Note that we allow $M_n = O(T_n^{\frac{1-\delta}{1+\frac{1}{2\xi(q,\beta)}}})$ for some $\delta > 0$ so that the error of the strong Gaussian approximation for high-dimensional nonstationary time series from Mies and Steland [MS22] is negligible. As discussed in Mies and Steland [MS22], in the limiting case when $\beta \geq 3$ and $q \to \infty$ we have $M_n = O(T_n^{\frac{1}{4}-\delta})$ for some $\delta > 0$.

The following theorem is a distribution-uniform version of the strong Gaussian approximation from Theorem 3.1 in Mies and Steland [MS22].

Lemma E.1. *For some $n \in \mathbb{N}$, let the process $(W_{t,n})_{t \in [n]}$ be defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P \in \mathcal{P}_n}$ such that Assumption 17 is satisfied for $\mathcal{P}_n$ with some $q > 2$, $\beta > 1$ and constant $\Theta_n > 0$. For each $P \in \mathcal{P}_n$ and $t \in [n]$, let $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ with $\mathbb{E}_P(W_{t,n}) = 0$ and suppose $d_n < cn$ for some $c > 0$. Let $(\Omega', \mathcal{B}', \mathbb{P}'_P)_{P \in \mathcal{P}_n}$ be a new collection of probability spaces on which there exist random vectors $(W'_{t,n})_{t \in [n]}$ such that $(W'_{t,n})_{t \in [n]} \overset{d}{=} (W_{t,n})_{t \in [n]}$ for each $P \in \mathcal{P}_n$ and independent, mean zero, Gaussian random vectors $V'_{t,n}$ such that*

$$\sup_{P \in \mathcal{P}_n} \left(\mathbb{E}_P \max_{k \leq n} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{k} (W'_{t,n} - V'_{t,n}) \right\|_2^2 \right)^{\frac{1}{2}} \leq C \Theta_n \sqrt{\log(n)} \left(\frac{d_n}{n} \right)^{\chi(q,\beta)}$$

for some universal constant C depending only on q, c, and β.

If $\beta > 2$, then the local long-run covariance matrix $\Sigma_{P,t,n} = \sum_{h=-\infty}^{\infty} \mathrm{Cov}_P(G_{t,n}(\mathcal{H}_0), G_{t,n}(\mathcal{H}_h))$ is well defined for each $P \in \mathcal{P}_n$ and $n \in \mathbb{N}$. If Assumption 18 is also satisfied, then there exist random vectors $(W'_{t,n})_{t \in [n]}$ such that $(W'_{t,n})_{t \in [n]} \overset{d}{=} (W_{t,n})_{t \in [n]}$ for each $P \in \mathcal{P}_n$ and independent, mean zero, Gaussian random vectors $V^*_{t,n} \sim \mathcal{N}(0, \Sigma_{P,t,n})$ such that

$$\sup_{P \in \mathcal{P}_n} \left(\mathbb{E}_P \max_{k \leq n} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{k} (W'_{t,n} - V^*_{t,n}) \right\|_2^2 \right)^{\frac{1}{2}} \leq C \Theta_n \Gamma_n^{\frac{1}{2} \frac{\beta-2}{\beta-1}} \sqrt{\log(n)} \left(\frac{d_n}{n} \right)^{\xi(q,\beta)}$$

for some universal constant C depending only on q, c, and β.

Proof of Lemma E.1: Assumptions 17 and 18 are distribution-uniform versions of conditions (G.1) and (G.2) from Mies and Steland [MS22]. Hence, under the assumptions of the Lemma related to Assumptions 17 and 18, the distribution-pointwise inequalities from Theorem 3.1 in Mies and Steland [MS22] hold for all $P \in \mathcal{P}_n$. Since the suprema (over all distributions in the collection) of the upper bounds are always finite, the distribution-uniform inequalities from the Lemma hold for $\mathcal{P}_n$ by basic properties of the supremum. □

The following result is a distribution-uniform version of Theorem 3.2 from Mies and Steland [MS22].

Lemma E.2. *For some $n \in \mathbb{N}$, let the process $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ for $t \in [n]$ be defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$, and suppose $W_{t,n} \in L_q(P)$ for each $P \in \mathcal{P}_n$ for some $2 \le r \le q < \infty$. Define $\theta_{P,t,n}(j,q,r)$ as in Definition 5.*

There exists a universal constant $C = C(q,r)$, such that for all $n \in \mathbb{N}$, we have

$$\sup_{P\in\mathcal{P}_n}\left(\mathbb{E}_P \max_{k\le n}\left\|\sum_{t=1}^{k}(W_{t,n}-\mathbb{E}_P(W_{t,n}))\right\|_r^q\right)^{\frac{1}{q}} \le \sup_{P\in\mathcal{P}_n}\left(C\, n^{\frac{1}{2}-\frac{1}{q}}\sum_{j=1}^{\infty}\left(\sum_{t=1}^{n}\theta_{P,t,n}^q(j,q,r)\right)^{\frac{1}{q}}\right)$$
$$\le \sup_{P\in\mathcal{P}_n}\left(C\, n^{\frac{1}{2}}\sum_{j=1}^{\infty}\max_{t\le n}\theta_{P,t,n}(j,q,r)\right).$$

In the special case $r = 2$, the inequality may be improved to

$$\sup_{P\in\mathcal{P}_n}\left(\mathbb{E}_P \max_{k\le n}\left\|\sum_{t=1}^{k}(W_{t,n}-\mathbb{E}_P(W_{t,n}))\right\|_2^q\right)^{\frac{1}{q}}$$
$$\le \sup_{P\in\mathcal{P}_n}\left(C\sum_{j=1}^{\infty}(j\wedge n)^{\frac{1}{2}-\frac{1}{q}}\left(\sum_{t=1}^{n}\theta_{P,t,n}^q(j,q,2)\right)^{\frac{1}{q}} + C\sum_{j=1}^{n}\left(\sum_{t=1}^{n}\theta_{P,t,n}^2(j,2,2)\right)^{\frac{1}{2}}\right).$$

Proof of Lemma E.2: Under the assumptions of the Lemma, the distribution-pointwise inequalities from Theorem 3.2 in Mies and Steland [MS22] hold for all $P \in \mathcal{P}_n$. Since the suprema (over all distributions in the collection) of the upper bounds are always finite, the distribution-uniform inequalities from the Lemma hold for $\mathcal{P}_n$ by basic properties of the supremum. □

E.3 Distribution-uniform feasible Gaussian approximation

In this subsection, we introduce distribution-uniform extensions of Theorem 4.1 and Proposition 4.2 from Mies and Steland [MS22] so that the distribution-uniform strong Gaussian approximation from Subsection E.2 can be used for statistical inference. The key is a distribution-uniform cumulative covariance estimator $\hat{Q}_{k,n}$ of the cumulative long-run covariance process $Q_{k,n} = \sum_{t=1}^{k}\Sigma_{P,t,n}$ where $\Sigma_{P,t,n} = \sum_{h=-\infty}^{\infty}\mathrm{Cov}_P(G_{t,n}(\mathcal{H}_0), G_{t,n}(\mathcal{H}_h))$ and $W_{t,n} = G_{t,n}(\mathcal{H}_t)$. We will prove these guarantees for the same estimator from Mies and Steland [MS22], namely

$$\hat{Q}_{k,n} = \sum_{r=L_n}^{k}\frac{1}{L_n}\left(\sum_{s=r-L_n+1}^{r}W_{s,n}\right)\left(\sum_{s=r-L_n+1}^{r}W_{s,n}\right)^{\top}$$

for some window size $L_n \asymp n^{\zeta}$ for some $\zeta \in (0, \frac{1}{2})$.

In practice, we select the lag window parameter L_n with the minimum volatility method suggested by Luo and Wu [LW23]. First, select $H \in \mathbb{N}$ candidate window sizes $l_1 < l_2 < \ldots < l_H$. For each index $h = 1, \ldots, H$, let

$$\hat{\Sigma}_{t,n,l_h} = \frac{1}{l_h}\left(\sum_{s=t-l_h+1}^{t}W_{s,n}\right)\left(\sum_{s=t-l_h+1}^{t}W_{s,n}\right)^{\top}$$

be the local long-run covariance matrix with window size $l_h \in \mathbb{N}$ corresponding to the index h. Second, calculate the minimum volatility criterion for each $j = 1, \ldots, H$,

$$\mathbf{MV}(j) = \max_{t=l_H,\ldots,n}\mathbf{se}[(\hat{\Sigma}_{t,n,l_h})_{h=1\vee(j-\Delta)}^{H\wedge(j+\Delta)}],$$

where the Δ is chosen heuristically (such as $\Delta = 3$) and can be replaced with other reasonable choices, and where

$$\mathbf{se}[(\hat{\Sigma}_{t,n,l_h})_{h=h_1}^{h_2}] = \mathrm{tr}\left[\frac{1}{h_2-h_1+1}\sum_{h=h_1}^{h_2}\left(\hat{\Sigma}_{t,n,l_h} - \frac{1}{h_2-h_1+1}\sum_{l=h_1}^{h_2}\hat{\Sigma}_{t,n,l_h}\right)^2\right]^{1/2}.$$

Third, select the window size L_n^* that corresponds to the index j^* which yields the smallest minimum volatility criterion

$$j^* = \underset{j=1,\ldots,H}{\arg\min}\ \mathbf{MV}(j).$$

The following theorem is a distribution-uniform extension of Theorem 4.1 from Mies and Steland [MS22].

Lemma E.3. *For some $n \in \mathbb{N}$, let the process $(W_{t,n})_{t\in[n]}$ be defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$. Let $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ satisfy Assumptions 17 and 18 for $\mathcal{P}_n$ with $q \geq 4$ and $\beta > 2$. Then*

$$\sup_{P\in\mathcal{P}_n}\left(\mathbb{E}_P \max_{k=L_n,\ldots,n}\left|\left|\hat{Q}_{k,n} - \sum_{t=1}^{k}\Sigma_{P,t,n}\right|\right|_{\mathrm{tr}}\right) \leq C\Theta_n^2\left(\Gamma_n\sqrt{L_n} + \sqrt{nd_nL_n} + nL_n^{-1} + nL_n^{2-\beta}\right)$$

for some universal constant C depending only on β and q.

Proof of Lemma E.3: Assumptions 17 and 18 are distribution-uniform versions of conditions (G.1) and (G.2) from Mies and Steland [MS22]. Hence, under the assumptions of the Lemma related to Assumptions 17 and 18, the distribution-pointwise inequalities from Theorem 4.1 in Mies and Steland [MS22] hold for all $P \in \mathcal{P}_n$. Since the supremum (over all distributions in the collection) of the upper bound is always finite, the distribution-uniform inequality from the Lemma holds for $\mathcal{P}_n$ by basic properties of the supremum. □

The following theorem is a distribution-uniform extension of Proposition 4.2 from Mies and Steland [MS22].

Lemma E.4. *For each $n \in \mathbb{N}$ and $P \in \mathcal{P}_n$, let $\Sigma_{P,t,n}, \Sigma'_{P,t,n} \in \mathbb{R}^{d_n\times d_n}$ be symmetric, positive definite matrices for $t \in [n]$. For some $n \in \mathbb{N}$, let the independent, mean zero, Gaussian random vectors $V_{t,n} \sim \mathcal{N}(0, \Sigma_{P,t,n})$ for $t \in [n]$ be defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$. Let $(\Omega', \mathcal{B}', \mathbb{P}'_P)_{P\in\mathcal{P}_n}$ be a new collection of probability spaces on which there exist independent, mean zero, Gaussian random vectors $V'_{t,n} \sim \mathcal{N}(0, \Sigma'_{P,t,n})$ for $t \in [n]$ such that*

$$\sup_{P\in\mathcal{P}_n}\left(\mathbb{E}_P \max_{k=1,\ldots,n}\left|\left|\sum_{t=1}^{k}V_{t,n} - \sum_{t=1}^{k}V'_{t,n}\right|\right|_2^2\right) \leq \sup_{P\in\mathcal{P}_n}\left(C\ \log(n)\ [\sqrt{n\delta_{P,n}\rho_{P,n}} + \rho_{P,n}]\right),$$

where

$$\delta_{P,n} = \max_{k=1,\ldots,n}\left|\left|\sum_{t=1}^{k}\Sigma_{P,t,n} - \sum_{t=1}^{k}\Sigma'_{P,t,n}\right|\right|_{\mathrm{tr}},$$

$$\rho_{P,n} = \max_{t=1,\ldots,n}||\Sigma_{P,t,n}||_{\mathrm{tr}}.$$

Proof of Lemma E.4: The distribution-pointwise inequalities from Proposition 4.2 in Mies and Steland [MS22] hold for all $P \in \mathcal{P}_n$. Since the supremum (over all distributions in the collection) of the upper bound is always finite, the distribution-uniform inequality from the Lemma holds for $\mathcal{P}_n$ by basic properties of the supremum. □

F Auxiliary Lemmas

The following result is a distribution-uniform version of Proposition 5.4 from Mies and Steland [MS22].

Lemma F.1. *Let $G_{t,n}$ satisfy Assumption 17 with $q \geq 2$. Denote*

$$\gamma_{P,t,n}(h) = \mathrm{Cov}_P[G_{t,n}(\mathcal{H}_0), G_{t,n}(\mathcal{H}_h)] \in \mathbb{R}^{d_n\times d_n}.$$

Then

$$\sup_{P\in\mathcal{P}_n}||\gamma_{P,t,n}(h)||_{\mathrm{tr}} \leq \Theta_n^2\sum_{j=h}^{\infty}j^{-\beta},$$

where $||\cdot||_{\mathrm{tr}}$ denotes the trace norm. Hence, if $\beta > 2$, then the long-run covariance matrix $\gamma_{P,t,n} = \sum_{h=-\infty}^{\infty}\gamma_{P,t,n}(h)$ is well-defined for all $P \in \mathcal{P}_n$.

Proof of Lemma : The result follows by the same steps in the proof from Proposition 5.4 in Mies and Steland [MS22] while carrying the supremum over $\mathcal{P}_n$. □

The following result is similar to the bounded convergence lemma from Lemma 25 in Shah and Peters [SP20].

Lemma F.2. *For each $n \in \mathbb{N}$, let X_n be a real-valued random variable with distribution determined by $P \in \mathcal{P}_n$ where the collection of distributions $\mathcal{P}_n$ can change with n. Let $C > 0$ and suppose that $|X_n| \leq C$ for all $n \in \mathbb{N}$ and $X_n = o_{\mathcal{P}}(1)$. Then $\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|X_n|) = o(1)$.*

Proof of Lemma F.2: For any given $\epsilon > 0$,

$$|X_n| = |X_n| \mathbb{1}_{\{|X_n| > \epsilon\}} + |X_n| \mathbb{1}_{\{|X_n| \leq \epsilon\}} \leq C \mathbb{1}_{\{|X_n| > \epsilon\}} + \epsilon.$$

By the assumption that $X_n = o_{\mathcal{P}}(1)$, we can find some $N \in \mathbb{N}$ such that $\sup_{P \in \mathcal{P}_n} \mathbb{P}_P(|X_n| > \epsilon) < \epsilon / C$ for $n \geq N$. Hence, for $n \geq N$ we have

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|X_n|) \leq C \sup_{P \in \mathcal{P}_n} \mathbb{P}_P(|X_n| > \epsilon) + \epsilon < 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary, we obtain the desired result. □