Conditional Independence Testing in the Presence of Heteroskedasticity, Autocorrelation, and Nonstationarity

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

February 13, 2025

Abstract

Understanding the structure of complex temporal systems is a central challenge in many scientific and engineering disciplines. There has been a growing recognition that the standard toolkit for multivariate time series analysis, largely dominated by linear vector autoregressive models, is insufficient for capturing the richness of the nonlinear dependencies present in many real-world systems. As a result, many algorithms for causal discovery, variable selection, and graphical modeling have been developed for nonlinear time series. Accurately detecting conditional dependencies is a fundamental aspect of many of these algorithms. Unfortunately, most nonparametric conditional independence tests rely heavily on the iid assumption. At best, some of these tests allow for weak temporal dependence, yet they still require stationarity. In our simulations, we find that even slight deviations from the iid setting lead to disastrous results for some tests.

Hence, there is a need for conditional independence tests which can be applied to real-world non-iid data and safely deployed in downstream statistical applications. To meet this need, we introduce a framework for robustly detecting conditional dependencies. As far as we know, this is the first framework to enable conditional independence testing with a *single realization* of a nonstationary nonlinear time series. The key technical ingredients are time-varying regression estimation, time-varying covariance estimation, and a distribution-uniform strong Gaussian approximation. Notably, our approach avoids data splitting and kernel smoothing, which results in a practical, user-friendly test.

Contents

1 Introduction **2**

2 The Dynamic Generalized Covariance Measure (dGCM) **5**
2.1 Setting and notation . . . 5
2.2 The null hypothesis of conditional independence . . . 6
2.3 Time-varying regression functions . . . 8
2.4 Main ideas of our work and the algorithm . . . 8

3 Assumptions and the Main Theoretical Result for dGCM **10**
3.1 Nonstationary observed processes . . . 11
3.2 Prediction processes . . . 12
3.3 Nonstationary error processes . . . 14
3.4 Assumptions on dependence and nonstationarity . . . 15
3.5 Main theoretical result for dGCM . . . 16

*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.

4 dGCM with Sieve Time-Varying Regression (Sieve-dGCM) 18
4.1 Setting and notation . . . 18
4.2 Locally stationary observed processes . . . 19
4.3 Sieve time-varying nonlinear regression estimator . . . 20
4.4 Locally stationary error processes . . . 23
4.5 Assumptions on dependence and nonstationarity . . . 24
4.6 Assumptions on local long-run covariances . . . 27
4.7 Theoretical result for Sieve-dGCM . . . 29

5 Numerical Simulations 29
5.1 Parameter selection via subsampling and minimum volatility . . . 29
5.2 Analysis of level and power . . . 31
5.3 Discussion of simulation results . . . 39

6 Future Work 39

A Extensions 48
A.1 Alternative test statistics . . . 48
A.2 Other approaches for estimating time-varying regression functions . . . 49
A.3 Simplifications under stationarity . . . 49
A.4 Local CI testing for locally stationary time series . . . 50
A.5 Piecewise locally stationary time series . . . 52
A.6 Weakening the temporal dependence assumptions for error processes . . . 52

B Distribution-Uniform Theory 52
B.1 Literature review of distribution-uniform inference . . . 52
B.2 Distribution-uniform strong Gaussian approximation . . . 53
B.3 Distribution-uniform feasible Gaussian approximation . . . 56
B.4 Auxiliary Lemmas . . . 57

C Proofs 58
C.1 Proof of Theorem 3.1 . . . 58
C.2 Proof of Theorem 4.1 . . . 69

1 Introduction

A great deal of work has been dedicated to developing tests for conditional independence (henceforth CI). That is, testing whether two random vectors X and Y are independent given a third random vector Z. For example, there are CI tests based on conditional densities [SW08], characteristic functions [SW07], empirical likelihood ratios [SW14], discretization [Mar05; Hua10], permutation [Dor+14; Sen+17], kernels [Fuk+07; Zha+11; SP11], copulas [BRT12], and conditional mutual information [Run18b]. Also, there are many CI tests based on regressing X on Z and Y on Z followed by testing for independence between the residuals [Pat+09; Pet+14; Ram14; FFX20; ZZG17; Zha+19].

Unfortunately, CI tests oftentimes struggle to control Type-I error in finite samples, as demonstrated through the examples and simulations presented by Shah and Peters [SP20]. In fact, Shah and Peters [SP20] show that CI testing is *fundamentally impossible* without making further assumptions. This issue has sparked significant interest in CI testing over the last several years. We begin by providing an overview of these recent advances in CI testing. Afterwards, we discuss how our work addresses limitations in the existing literature. Finally, we motivate our work by reviewing key applications of CI tests for time series in areas such as variable selection, causal discovery, and Granger causality.

The hardness of conditional independence testing. The no-free-lunch result from Shah and Peters [SP20] states that if one wants to have a CI test with Type-I error control for all absolutely continuous (with respect to the Lebesgue measure) triplets of random vectors (X, Y, Z), then this CI test cannot have power against any alternative. To make the CI testing problem feasible, we must consider a smaller subset of the null hypothesis and use domain knowledge to select an appropriate CI

test. This hardness result was revisited by Neykov, Balakrishnan, and Wasserman [NBW21] and Kim et al. [Kim+22], and was extended to the time series setting by Bodik and Pasche [BP24].

Shah and Peters [SP20] proposed a CI test based on the *generalized covariance measure* (GCM), which is a suitably normalized sum of the products of the residuals from the regressions of X on Z and Y on Z. In this case, the practitioner's domain knowledge is used to select appropriate regression methods for the problem at hand. In contrast to the CI tests previously mentioned, Shah and Peters [SP20] demonstrate that the GCM test has asymptotic Type-I error control, *uniformly* over a large collection of distributions for which the null hypothesis of CI holds.

Since then, numerous tests have been developed which draw inspiration from the original GCM test [Lun+22; SHB22; LSP22; CPH22; CZK24; WR23; KKR24]. Our CI test can be considered a GCM-type test for the nonstationary nonlinear time series setting. As we will discuss, moving to this setting introduces several complexities and requires completely different theoretical tools than the original GCM test. Although we develop a GCM-type CI test in this work, we believe it is important to point the reader to another influential literature about CI testing based on versions of the model-X assumption [Can+18; Liu+22; Niu+24; BCS20; HJ20; BJ22; SMR23; GHL24].

Limitations of the existing literature. Most of the previously discussed CI tests lack Type-I error control guarantees outside the iid setting. Furthermore, the literature on CI testing when given only a single realization of a nonstationary time series remains strikingly limited. To the best of our knowledge, only two CI tests have been proposed for this setting.

First, Malinsky and Spirtes [MS19] introduce a CI test for nonstationary linear vector autoregressions with iid Gaussian errors. Specifically, they study time series that exhibit "stochastic trends" so that the first difference of the process is stable. In contrast, our CI test allows for nonlinear time series with very general forms of nonstationarity and time-varying regression functions with non-iid, non-Gaussian errors. Moreover, we demonstrate that our CI test possesses uniformly asymptotic Type-I error control, as established for the GCM test from Shah and Peters [SP20].

Second, Flaxman, Neill, and Smola [FNS15] develop a CI testing framework for non-iid data based on Gaussian process regression. The main idea is to pre-whiten the non-iid data using Gaussian process regression to control for dependencies (e.g. spatial, temporal, or network), which should yield iid residuals. The next step is to test for independence between these residuals using the Hilbert-Schmidt Independence Criterion (HSIC) [Gre+07]. The authors mention that their framework could be used with nonstationary covariance functions, although this idea was not developed further.

We also mention some CI tests designed for the setting in which multiple realizations of a stochastic process are available. Manten et al. [Man+24] develop a CI test for stochastic processes using the signature kernel. Christgau, Petersen, and Hansen [CPH22] introduced a framework for testing so-called "conditional local independence" relationships between point processes. Lundborg, Shah, and Peters [LSP22] introduce a CI test for function-valued random variables. Also, we note that there is a growing literature on (unconditional) independence testing for nonstationary time series. Liu et al. [Liu+23] develop independence tests based on the Hilbert-Schmidt Independence Criterion (HSIC) [Gre+07]. These tests require multiple realizations of the nonstationary time series, whereas the independence tests for locally stationary time series from Brunotte [Bru22] and Beering [Bee21] only require one realization of the process.

Causal discovery. The discovery of time-lagged causal relationships (see Figure 1) from observational time series is an important problem in numerous scientific domains [Run+19b]. CI tests for time series are a core component of constraint-based and hybrid causal discovery algorithms designed for temporally correlated data. For example, Runge et al. [Run+19a] used the conditional mutual information-based CI test from Runge [Run18b] in a causal discovery algorithm for time series called PCMCI, which builds on the classical PC algorithm [SGS01]. Over the last several years, causal discovery for *nonstationary* time series has become an increasingly active area of research [MS19; Hua+20; FHG23; Don+23; SGF24]. We stress that the CI test used as the foundation for the causal discovery algorithm must be appropriately tailored to the characteristics of the data. If the underlying CI test fails to account for temporal dependence or nonstationarity, then the causal discovery algorithm may produce incorrect conclusions about the causal structure of process. Our work fills a gap in the literature on causal discovery for nonstationary nonlinear time series by providing a practical CI test for this setting.

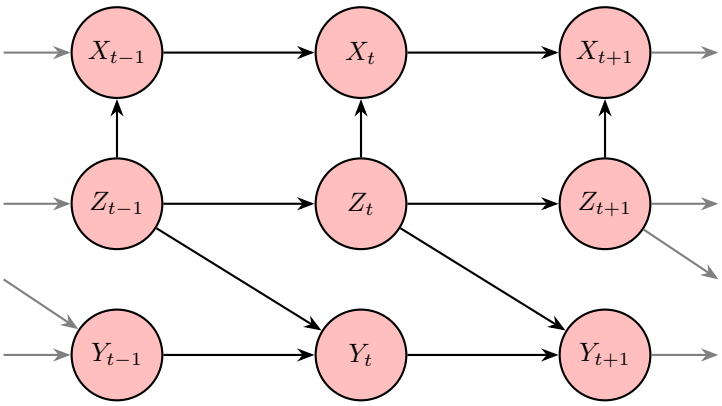

Figure 1: Causal time series graph depicting the time-lagged causal relationships among the processes $X = (X_t)_{t=1}^n$, $Y = (Y_t)_{t=1}^n$, and $Z = (Z_t)_{t=1}^n$ for some sample size $n \in \mathbb{N}$. In this example, Z is a common cause of both X and Y, directly influencing X in the same period and affecting Y with a one time-step delay. Although the underlying dynamics of the processes may be nonstationary, the causal structure remains time-invariant, as indicated by the gray arrows on the ends of the graph.

Variable selection. A central problem in statistics and machine learning is variable selection. CI tests can be used for variable selection when paired with multiple testing procedures to control the false discovery rate (FDR) [BH95; BY01]. In the familiar context of time series forecasting, the goal is to identify a minimal subset $S \subseteq \{1, ..., p\}$ out of p signals (including relevant lags) such that, for all times t, the forecasting target Y_{t+h} at horizon h is conditionally independent of the other signals $(X_{t,i})_{i \notin S}$ given $(X_{t,i})_{i \in S}$. This minimal subset S is usually called a Markov blanket for Y_{t+h}; see Pearl [Pea14] and Candés et al. [Can+18] for more discussion. We contribute to this literature by providing a CI test flexible enough to be used for identifying relevant forecasting signals in unstable environments.

Granger causality. Lastly, we discuss a commonly used framework for assessing relationships between time series, which goes by the name of Granger causality. Unfortunately, the numerous definitions of Granger causality have caused a great deal of confusion. Recently, Shojaie and Fox [SF22] have written a comprehensive review to help clarify matters.

The original definition of Granger causality from Granger [Gra69] is about prediction. Informally, a time series X is said to be Granger causal of another time series Y if the variance of the *optimal prediction* of Y_t at time t using all the relevant information up to time $t-1$ is increased by removing the history of X up to time $t-1$. See Section 2 in Shojaie and Fox [SF22] for the exact definition and the stringent conditions under which this predictive definition corresponds to genuine causality as in Pearl [Pea09]. While this original definition does not assume linear dynamics, much of the following methodology revolves around the identification of coefficients in linear vector autoregressive (VAR) models with p time series [Gra80; Lüt05; BSM15].

Another definition of Granger causality, referred to as *strong* Granger causality [FM82], is stated in terms of conditional independence relationships between (possibly nonlinear) time series. Let $(X_i)_{i \in [p]}$ be p signals used to predict the target Y. The time series X_i is said to be (strongly) Granger noncausal of Y if, for all times t, Y_t is conditionally independent of the history of the signal X_i up to time $t-1$ given the history of the other signals $(X_j)_{j \in [p] \setminus \{i\}}$ up to time $t-1$. See Definition 2 in Shojaie and Fox [SF22] for the exact definition, and the rest of Section 4 therein for more discussion.

Notably, Eichler [Eic12] introduced a comprehensive graphical modeling framework for time series based on strong Granger causality, which can be detected using CI tests for nonlinear time series [SP11; BRT12]. In a similar vein, our proposed CI test can be used to detect strong Granger causality for nonlinear time series with *nonstationary* dynamics. This can be incorporated into graphical modeling frameworks for nonstationary nonlinear time series, analogous to Basu and Rao [BR23].

We note that there are also various techniques for assessing nonlinear Granger causality that do not utilize CI testing. For instance, the neural Granger causality method from Tank et al. [Tan+21] extracts Granger causal structures by using sparsity-inducing penalties on the weights of structured multilayer perceptrons (MLPs) and recurrent neural networks (RNNs). Additionally, there is an influential strand

of literature connecting Granger causality and directed information theory [AM12; QKC15]. See Section 4 of Shojaie and Fox [SF22] for more discussion of nonlinear Granger causality.

Paper outline. The rest of the paper is structured as follows. In Section 2, we discuss the main ideas and implementation of our proposed CI test. In Section 3, we introduce the details of our theoretical framework. In Section 4, we develop the theoretical justifications for our test in the context of the sieve time-varying nonlinear regression estimator from Ding and Zhou [DZ21] within the well-studied framework of locally stationary time series. In Section 5, we demonstrate the satisfactory performance of our test through comprehensive numerical simulations. Along the way, in Subsection 5.1, we introduce a novel cross-validation procedure based on subsampling for selecting the parameters of "global" estimators of time-varying regression functions. In Section 6, we discuss promising avenues for future work. In Section A, we consider several extensions of our CI test. In Section B, we state distribution-uniform versions of the results from Mies and Steland [MS23], such as the strong Gaussian approximation for high-dimensional nonstationary nonlinear time series.

2 The Dynamic Generalized Covariance Measure (dGCM)

In this section, we give a high-level overview of our work. Specifically, we introduce the notation, main ideas, and implementation of our proposed *dynamic generalized covariance measure* (dGCM) test. For expository purposes, we delay the technical details of our theoretical framework until Section 3.

2.1 Setting and notation

We work in a triangular array framework for high-dimensional nonstationary time series. Let $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t\in[n]}$ be the observed sequence of length $n \in \mathbb{N}$, where $[n] = \{1, \dots, n\}$. We use the notation $X_n = (X_{t,n})_{t\in[n]}$, $Y_n = (Y_{t,n})_{t\in[n]}$, $Z_n = (Z_{t,n})_{t\in[n]}$ to refer to sequences of length n, and we use the notation X, Y, Z to generically refer to the processes with any length. Let $d_X = d_{X,n}$, $d_Y = d_{Y,n}$, $d_Z = d_{Z,n}$ denote the dimensions, which can grow with n. 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 each dimension of $X_{t,n}$, $Y_{t,n}$, $Z_{t,n}$ because we want to infer time-lagged conditional dependencies. Negative time-offsets are called *lags* of the process, and positive time-offsets are called *leads* of the process. Time-offsets of zero are allowed so that contemporaneous conditional dependencies can be considered. Let

$$A_i \subset \{-n+1, \dots, n-1\},\ B_j \subset \{-n+1, \dots, n-1\},\ C_k \subset \{-n+1, \dots, 0\},$$

be the sets of time-offsets of $X_{t,n,i}$, $Y_{t,n,j}$, $Z_{t,n,k}$ under consideration. We require the time-offsets C_k to be non-positive so that the conditioning variables are known at time t for purposes that will be made clear in the next few subsections. In practice, the largest (in magnitude) time-offsets should be selected small enough so that there is a sufficient amount of data to conduct the test.

Denote 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 $Z_{t,n,k,c} = Z_{t+c,n,k}$. Denote the sets of all time-offsets by $A = \bigcup_{i=1}^{d_X} A_i$, $B = \bigcup_{j=1}^{d_Y} B_j$, $C = \bigcup_{k=1}^{d_Z} C_k$, and largest (signed) time-offsets by $a_{\max} = \max(A)$, $b_{\max} = \max(B)$, $c_{\max} = \max(C)$, and the smallest (signed) time-offsets by $a_{\min} = \min(A)$, $b_{\min} = \min(B)$, $c_{\min} = \min(C)$.

Since we are interested in time-lagged conditional independence relationships, it is often useful to refer to the subset of original times,

$$\mathcal{T}_n = \{1 - \min(a_{\min}, b_{\min}, c_{\min}), n - \max(a_{\max}, b_{\max}, c_{\max})\} \subseteq \{1, \dots, n\},$$

in which *all* time-offsets of each dimension of $X_{t,n}$, $Y_{t,n}$, $Z_{t,n}$ are actually observed. Going forward, we will write $t \in \mathcal{T}_n$ instead of $t \in [n]$ because we are only using the subset of times in which all time-offsets are observed. Denote the first time of $\mathcal{T}_n$ by $\mathbb{T}_n^- = \min(\mathcal{T}_n)$, the last time of $\mathcal{T}_n$ by $\mathbb{T}_n^+ = \max(\mathcal{T}_n)$, and the cardinality of $\mathcal{T}_n$ by $T_n = |\mathcal{T}_n|$. Note that if no negative time-offsets (i.e. lags) are used then $\min(a_{\min}, b_{\min}, c_{\min}) = 0$, and if no positive time-offsets (i.e. leads) are used then $\max(a_{\max}, b_{\max}, c_{\max}) = 0$. Hence, if only time-offsets of zero are used, then $\mathcal{T}_n = [n]$.

For all $t \in \mathcal{T}_n$, denote the vectors with all dimensions and time-offsets of interest 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 entire 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 to grow with n, that is, $A_i = A_{i,n}$, $B_j = B_{j,n}$, $C_k = C_{k,n}$ and $A = A_n$, $B = B_n$, $C = C_n$. However, we require that the largest (in magnitude) time-offset grows at a slower rate than n such that as $n \to \infty$ we have $\min(a_{\min}, b_{\min}, c_{\min})/n \to 0$ and $\max(a_{\max}, b_{\max}, c_{\max})/n \to 0$ so that the number of observed times $T_n \to \infty$ arbitrarily slowly.

Since we allow *both* the number of time-offsets and the number of dimensions to grow with n, we introduce the index set

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

which contains all of the indices for the dimensions and time-offsets of interest. Note that $\mathcal{D}_n$ is specified by the user, and need not contain all possible combinations. Going forward, we will often refer to the dimension/time-offset tuple by $m = (i, j, a, b) \in \mathcal{D}_n$ to lighten the notation. The index set $\mathcal{D}_n$ depends on the sample size n through the dimensions *and* the time-offsets, so its cardinality $D_n = |\mathcal{D}_n|$ may grow with n. Note that D_n reflects the *intrinsic dimensionality* of the problem and will appear frequently in the rest of the paper. In the "best case" scenario, we allow $D_n = O(T_n^{\frac{1}{6}})$. See (18) and the rest of Subsection B.2 for the full details about how quickly D_n can grow.

For each $n \in \mathbb{N}$, let $\mathcal{P}_n$ be a collection of distributions for the processes, which we allow to change with n. For expository purposes, we delay the technical details about $\mathcal{P}_n$ until the end of Subsection 3.1. Next, we state the null hypothesis.

2.2 The null hypothesis of conditional independence

Our univariate test is for the null hypothesis

$$X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \tag{1}$$

for a single dimensions/time-offset tuple $(i, j, a, b) \in \mathcal{D}_n$. If domain knowledge suggests that we can restrict $\mathcal{P}_n$ to consist of distributions in which the conditional dependencies are time-invariant, then we can use the alternative hypothesis

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

We begin by focusing on time-invariant conditional independence relationships and address the time-varying case at the conclusion of this subsection.

Consider the following forecasting example in the univariate setting with $d_X = 1$, $d_Y = 1$, $\boldsymbol{d_Z} \geq 1$.

Example 2.1 (Univariate CI test for time series forecasting)**.** *Suppose we are interested in determining whether the current value (time-offset $a = 0$) of a signal (dimension $i = 1$) is relevant for forecasting a target (dimension $j = 1$) seven time steps into the future (time-offset $b = 7$) after accounting for the existing forecasting signals $\boldsymbol{Z}_{t,n}$. Note that $\boldsymbol{Z}_{t,n}$ can consist of current values and lags of each of the forecasting signals. In this case, we would use the univariate version of our test with the null hypothesis*

$$X_{t,n,1,0} \perp\!\!\!\perp Y_{t,n,1,7} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n.$$

Naturally, our CI test can be paired with multiple testing procedures so that we can conduct several of these univariate tests while controlling the false discovery rate. For example, we can individually test for conditional independence between $X_{t,n,i,a}$ and $Y_{t,n,j,b}$ given $\boldsymbol{Z}_{t,n}$ for any combination of forecasting horizons $b \in \{7, 14, 21\}$, lags $a \in \{0, -7, -14\}$, targets $j \in \{1, 2, 3\}$, and signals $i \in \{1, 2, 3, 4, 5\}$.

The following null hypothesis of "no causal effect at all times" is pertinent to the growing literature on causal inference for time series [Sag+20; RGR22; Run+23; Run+19a; Run18a], particularly for the setting in which just one realization of a nonstationary time series is available.

Example 2.2 (Univariate CI test for time series causal inference)**.** *Suppose we are interested in determining whether the current value (time-offset $a = 0$) of a continuous treatment (dimension $i = 1$) has any causal effect on an outcome of interest (dimension $j = 1$) one time step into the future (time-offset $b = 1$) after accounting for the confounders $\boldsymbol{Z}_{t,n}$. Crucially, we assume there are no unobserved confounders. In this case, we would use the univariate version of our test with the null hypothesis*

$$X_{t,n,1,0} \perp\!\!\!\perp Y_{t,n,1,1} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n.$$

If we do not require a p-value for each conditional independence relationship, then we can use the multivariate version of our test. In this case, we use the null hypothesis

$$X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } (i,j,a,b) \in \mathcal{D}_n. \tag{3}$$

By grouping together highly correlated dimensions and consecutive time-offsets, one can often construct a multivariate test that is more powerful than a univariate test based on a single dimension/time-offset tuple. Note that when using the multivariate test, different alternative hypotheses can be used depending on whether it is reasonable to restrict $\mathcal{P}_n$ to consist of distributions in which the conditional dependencies are dimension-invariant. As we will explain next, there are many other situations in which grouping together related time series and using our multivariate test is useful.

Suppose we have time series data from nearby cities, countries, or sensors, as is common in the social sciences, earth sciences, and epidemiology. Write $X^{\ell}_{t,n,i,a}$, $Y^{\ell}_{t,n,j,b}$, $\boldsymbol{Z}^{\ell}_{t,n}$ to denote $X_{t,n,i,a}$, $Y_{t,n,j,b}$, $\boldsymbol{Z}_{t,n}$ at index $\ell \in \mathcal{L}_n$, where $\mathcal{L}_n$ is an index set (e.g. different locations). We can often gain power by grouping together the time series in $\mathcal{L}_n$ and using the multivariate test with the null hypothesis

$$X^{\ell}_{t,n,i,a} \perp\!\!\!\perp Y^{\ell}_{t,n,j,b} \mid \boldsymbol{Z}^{\ell}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } \ell \in \mathcal{L}_n, \tag{4}$$

for a single dimension/time-offset tuple $(i,j,a,b) \in \mathcal{D}_n$. Crucially, we allow each of the time series to be correlated with one another across $\mathcal{L}_n$ and to have different distributions. Note that this is the same as the null hypothesis (1) but for all indices $\ell \in \mathcal{L}_n$. We can use the alternative hypothesis

$$X^{\ell}_{t,n,i,a} \not\perp\!\!\!\perp Y^{\ell}_{t,n,j,b} \mid \boldsymbol{Z}^{\ell}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for some } \ell \in \mathcal{L}_n, \tag{5}$$

since we have restricted the collection of distributions $\mathcal{P}_n$ to consist of those in which the conditional dependencies are time-invariant. If we can further restrict $\mathcal{P}_n$ so that it consists of distributions with time-invariant *and* index-invariant conditional dependencies, then we can use the alternative hypothesis

$$X^{\ell}_{t,n,i,a} \not\perp\!\!\!\perp Y^{\ell}_{t,n,j,b} \mid \boldsymbol{Z}^{\ell}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } \ell \in \mathcal{L}_n. \tag{6}$$

The following example is about forecasting a group of time series with $d_X = 1$, $d_Y = 1$, $\boldsymbol{d_Z} \geq 1$, and $|\mathcal{L}_n| > 1$. Similar hypotheses arise in causal inference with groups of time series (i.e. multivariate analogies of Example 2.2); see Section 6 of Wahl, Ninad, and Runge [WNR24]. In many cases, the multivariate test used here will have more power than the univariate test used in Example 2.1.

Example 2.3 (Multivariate CI test for forecasting a group of time series)**.** *As in Example 2.1, we are interested in forecasting a target seven time steps ahead. We want to determine whether the current value of a new forecasting signal is relevant or not after accounting for the existing forecasting signals. As before, we have dimensions $i = 1$, $j = 1$ and time-offsets $a = 0$, $b = 7$. However, now we have access to the same set of forecasting signals and targets at each location index $\ell \in \mathcal{L}_n$. In this case, we would use the multivariate version of our test with the null hypothesis*

$$X^{\ell}_{t,n,1,0} \perp\!\!\!\perp Y^{\ell}_{t,n,1,7} \mid \boldsymbol{Z}^{\ell}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } \ell \in \mathcal{L}_n.$$

Going forward, we suppress the superscript $\ell \in \mathcal{L}_n$ and revert back to the original notation (i.e. from $X^{\ell}_{t,n,i,a}$, $Y^{\ell}_{t,n,j,b}$, $\boldsymbol{Z}^{\ell}_{t,n}$ to $X_{t,n,i,a}$, $Y_{t,n,j,b}$, $\boldsymbol{Z}_{t,n}$). Note that this superscript can always be ignored outside of the "groups of time series" context, such as when there is only one index (e.g. one location).

To deal with the problem of time-varying conditional dependencies, we suggest modeling the conditional dependencies as though they are stable during certain time windows. If the breakpoints separating these time windows are known, then we can simply use our CI test on each of these time windows and use multiple testing procedures to control the false discovery rate. However, this becomes a bit more challenging if the breakpoints are unknown. In future work, we will develop a procedure for identifying time windows during which stable conditional dependencies hold while controlling the false discovery rate. That way, we can focus this manuscript on the main testing procedure. In Subsection A.4, we discuss how to test for time-varying conditional independence relationships at particular points in time by using the framework of locally stationary time series.

2.3 Time-varying regression functions

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 tuple $(i, j, a, b) \in \mathcal{D}_n$, 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. The observed processes and error processes can all be nonstationary time series; see Section 3 for the details. Denote the product of errors at time t by

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

for $m = (i, j, a, b) \in \mathcal{D}_n$.

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

$$\hat{R}_{t,n,m} = \hat{\varepsilon}_{t,n,i,a}\hat{\xi}_{t,n,j,b}, \tag{7}$$

for $m = (i, j, a, b) \in \mathcal{D}_n$. Let $\hat{\boldsymbol{R}}_{t,n} = (\hat{R}_{t,n,m})_{m\in\mathcal{D}_n}$ be the high-dimensional vector process containing all the residual products for all dimension/time-offset combinations in $\mathcal{D}_n$.

2.4 Main ideas of our work and the algorithm

We show how to conduct tests for each of the null hypotheses from Subsection 2.2. Our proposed CI tests are asymptotically valid as $n \to \infty$, *uniformly* over a large collection of distributions for which the null holds. As mentioned in Section 1, our CI test is inspired by the generalized covariance measure (GCM) test from Shah and Peters [SP20]. The original GCM test is based on expected conditional covariances between iid random variables, whereas our proposed dynamic GCM (dGCM) test is based on time-varying expected conditional covariances between nonstationary nonlinear time series. While our dGCM test shares similarities with the original GCM test from Shah and Peters [SP20], adapting it to this complex setting necessitates the use of advanced theoretical tools for nonstationary nonlinear time series, many of which have only been developed in recent years. Now, let us briefly summarize the main ideas behind the univariate version of the original GCM test.

For this paragraph, momentarily redefine X, Y to be two random variables and Z to be a random vector. 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. As discussed in Shah and Peters [SP20], this can be seen as a generalization of the fact that the partial correlation coefficient, defined as the correlation between the residuals of linear regressions of X on Z and Y on Z, is equal to zero if and only if $X \perp\!\!\!\perp Y \mid Z$ when (X, Y, Z) are jointly Gaussian. The GCM test from Shah and Peters [SP20] is based on the normalized sum of the products of residuals from the regressions of X on Z and Y on Z.

Now, let us translate the "weak" conditional independence criterion of Daudin [Dau80] into our setting using the notation from Subsection 2.1. 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 $\mathbb{E}_P[\varphi(Y_{t,n,j,b}, \boldsymbol{Z}_{t,n}) \mid \boldsymbol{Z}_{t,n}] = 0$. Hence, under the null, the expected conditional covariances,

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

are always equal to zero for the dimension/time-offset combination $m = (i, j, a, b) \in \mathcal{D}_n$. Equivalently, the mean of the error products $\mathbb{E}_P(R_{P,t,n,m})$ 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 $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ from Subsection 2.3 will always be zero under the null. This can be seen as a generalization of the partial correlation coefficient being equal to zero under conditional independence in the linear-Gaussian time series context; see the related discussion about Gaussian graphical models for nonstationary time series in Basu and Rao [BR23].

Crucially, the expected conditional covariances $\rho_{P,t,n,m}$ can be zero at all times, even under alternatives in which the corresponding conditional dependencies always hold. Consequently, we can only hope to have power against alternatives in which the (time-varying) expected conditional covariances $\rho_{P,t,n,m}$ are non-zero for at least *some* times. Hence, we design our test statistic so that it can detect non-zero covariances between the errors $\varepsilon_{P,t,n,i,a}$ and $\xi_{P,t,n,j,b}$ *at any point in time along the path.*

Our test statistic is based on the empirical covariances between the residuals from the time-varying 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 $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$. Define the set of times

$$\mathcal{T}_{n,L} = \{L_n + \mathbb{T}_n^- - 1, \ldots, \mathbb{T}_n^+ - 1, \mathbb{T}_n^+\},$$

for some integer $L_n \in \mathbb{N}$ which is defined in (9). Also, denote its cardinality by $T_{n,L} = |\mathcal{T}_{n,L}|$. Denote the entire process containing the residual products by $\hat{\boldsymbol{R}}_n = (\hat{\boldsymbol{R}}_{t,n})_{t\in\mathcal{T}_{n,L}}$. The test statistic based on the maximum ℓ_p-norm ($p \geq 2$) achieved by the partial sum process is given by

$$S_{n,p}(\hat{\boldsymbol{R}}_n) = \max_{s\in\mathcal{T}_{n,L}} \left\| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_p . \tag{8}$$

Using this test statistic, we will reject the null hypothesis of conditional independence if the ℓ_p norm of the partial sum process of residual products ever becomes "too large" *at any point in time along the path.* For example, we can use the ℓ_∞-type or ℓ_2-type test statistics

$$S_{n,\infty}(\hat{\boldsymbol{R}}_n) = \max_{s\in\mathcal{T}_{n,L}} \left\| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_\infty , \; S_{n,2}(\hat{\boldsymbol{R}}_n) = \max_{s\in\mathcal{T}_{n,L}} \left\| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} \hat{\boldsymbol{R}}_{t,n} \right\|_2 ,$$

to achieve high power against either sparse or dense alternatives, respectively.

In the univariate case with dimensions $d_X = 1$, $d_Y = 1$, $\boldsymbol{d_Z} \geq 1$ and time-offsets $A = \{a\}$, $B = \{b\}$, the test statistic reduces to the absolute value of the partial sum process of residual products

$$S_n(\hat{R}_{n,m}) = \max_{s\in\mathcal{T}_{n,L}} \left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} \hat{R}_{t,n,m} \right| ,$$

where $m = (1, 1, a, b) \in \mathcal{D}_n$ and $\hat{R}_{n,m} = (\hat{R}_{t,n,m})_{t\in\mathcal{T}_{n,L}}$. See Subsections A.1 and A.4 for discussions of alternative test statistics, namely those based on the full sum and those employing kernel smoothing.

We use a bootstrap-based testing procedure which appeals to the strong Gaussian approximation in Section B. The key ingredient of this bootstrap procedure is the time-varying covariance structure of the approximating nonstationary Gaussian process. Define the rolling window estimate of the time-varying covariance matrices of the vectors of error products $(\boldsymbol{R}_{t,n})_{t\in\mathcal{T}_{n,L}}$ by

$$\hat{\boldsymbol{\Sigma}}_{t,n}^{\boldsymbol{R}} = \frac{1}{L_n} \left(\sum_{s=t-L_n+1}^{t} \hat{\boldsymbol{R}}_{s,n} \right)^{\otimes 2} , \tag{9}$$

where $L_n \in \mathbb{N}$ is a lag-window size parameter and the outer product is denoted by $v^{\otimes 2} = vv^T$. In the univariate case with dimensions $d_X = 1$, $d_Y = 1$, $\boldsymbol{d_Z} \geq 1$ and time-offsets $A = \{a\}$, $B = \{b\}$, we have the rolling-window estimate of the time-varying variances of the error products $(R_{t,n,m})_{t\in\mathcal{T}_{n,L}}$

$$\hat{\sigma}_{t,n}^{R} = \frac{1}{L_n} \left(\sum_{s=t-L_n+1}^{t} \hat{R}_{s,n,m} \right)^2 ,$$

where $m = (1, 1, a, b) \in \mathcal{D}_n$. We postpone the details about these covariances until Subsection 3.5.

The dGCM test is given by Algorithm 1 below. The main steps are time-varying nonlinear regression, time-varying covariance estimation, and a bootstrap procedure justified by the distribution-uniform strong Gaussian approximation in Section B. We present the testing procedure in the multivariate setting, and the algorithm for the univariate setting is obtained by replacing $S_{n,p}(\cdot)$, $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n}$, $\hat{\boldsymbol{R}}_{t,n}$, $\hat{\boldsymbol{R}}_n$, $\breve{\boldsymbol{R}}^{(r)}_{t,n}$, $\breve{\boldsymbol{R}}^{(r)}_n$ with $S_n(\cdot)$, $\hat{\sigma}^R_{t,n}$, $\hat{R}_{t,n}$, $\hat{R}_n$, $\breve{R}^{(r)}_{t,n}$, $\breve{R}^{(r)}_n$, respectively.

Algorithm 1 The dynamic generalized covariance measure (dGCM) test

Input: Dimensions and time-offsets of interest $\mathcal{D}_n$, time points $\mathcal{T}_n$, data $(X_{t,n}, Y_{t,n}, Z_{t,n})_{t\in\mathcal{T}_n}$, test statistic $S_{n,p}(\cdot)$, α for the significance level, α' for the quantile $\hat{q}^{\text{boot}}_{1-\alpha'}$, number of simulations s
for each time $t \in \mathcal{T}_n$ and dimension/time-offset tuple $m = (i, j, a, b) \in \mathcal{D}_n$ **do**
    Estimate the time-varying regression functions $\hat{f}_{t,n,i,a}$ and $\hat{g}_{t,n,j,b}$
    Calculate the product of residuals $\hat{R}_{t,n,m} = \hat{\varepsilon}_{t,n,i,a}\hat{\xi}_{t,n,j,b}$ from (7)
end for
Select the lag-window size L_n for covariance estimation according to Subsection 5.1
for each time $t \in \mathcal{T}_{n,L}$ **do**
    Calculate the rolling-window estimates $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n}$ of the time-varying covariance matrices from (9)
end for
for each simulation $r = 1, \ldots, s$ **do**
    for each time $t \in \mathcal{T}_{n,L}$ **do**
        Simulate independent Gaussian random vectors $\breve{\boldsymbol{R}}^{(r)}_{t,n} \sim \mathcal{N}(0, \hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n})$
    end for
    Calculate the test statistic $S_{n,p}(\breve{\boldsymbol{R}}^{(r)}_n)$ from (8) using the Gaussian process $\breve{\boldsymbol{R}}^{(r)}_n = (\breve{\boldsymbol{R}}^{(r)}_{t,n})_{t\in\mathcal{T}_{n,L}}$
end for
Calculate the $1-\alpha'$ empirical quantile $\hat{q}^{\text{boot}}_{1-\alpha'}$ of $(S_{n,p}(\breve{\boldsymbol{R}}^{(r)}_n))_{r=1}^s$
Calculate the test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ from (8) using the residual products $\hat{\boldsymbol{R}}_n = (\hat{\boldsymbol{R}}_{t,n})_{t\in\mathcal{T}_{n,L}}$
if $S_{n,p}(\hat{\boldsymbol{R}}_n) > \hat{q}^{\text{boot}}_{1-\alpha'}$ **then**
    Reject the null hypothesis at level α
end if
Output: Decision to reject the null hypothesis at level α or not

As with the GCM test from Shah and Peters [SP20], the Type-I error control guarantee for the dGCM test is asymptotic; see Theorem 3.1 for the details. With large sample sizes, we can straightforwardly reject the null hypothesis when the test statistic exceeds the quantile $\hat{q}^{\text{boot}}_{0.95}$ to obtain an approximately level 0.05 test. On the other hand, with small sample sizes, we may be conservative and only reject the null hypothesis at level 0.05 if the test statistic exceeds the quantile $\hat{q}^{\text{boot}}_{0.975}$.

3 Assumptions and the Main Theoretical Result for dGCM

In this section, we introduce a triangular array framework for high-dimensional nonstationary nonlinear time series. Specifically, our framework is designed to enable hypothesis testing based on residuals formed from time-varying nonlinear regression estimates. It provides a foundational template for future research in causal inference, graphical modeling, and variable selection within this complex setting.

We allow the processes to have long-range temporal dependence and very complicated forms of nonstationarity which can be both abrupt and smooth. We control the temporal dependence and nonstationarity of the processes *uniformly* over collections of distributions by employing versions of the functional dependence measure of Wu [Wu05] and the "bounded variation nonstationarity" condition of Mies and Steland [MS23]; see Assumptions 3.5 and 3.6 for the details. These distribution-uniform assumptions are needed for the uniform level guarantee in Theorem 3.1, which is our main theoretical result. We discuss the importance of this uniform guarantee in Subsection 5.3.

The framework we introduce in this section nests several well-studied classes of processes. In Section 4, we show how our framework encompasses a subclass of nonstationary processes called *locally stationary time series* [Dah97]. Specifically, we employ the functional dependence measure and a stochastic Lipschitz continuity condition to control the temporal dependence and nonstationarity;

see Assumptions 4.5 and 4.6. For more discussion of locally stationary time series, we refer readers to Dahlhaus [Dah12] and Dahlhaus, Richter, and Wu [DRW19] for the linear and nonlinear cases, respectively. In Subsection A.5, we explain how our framework also nests a more general class of nonstationary time series called *piecewise locally stationary time series* [Zho13]. This class extends locally stationary time series by permitting both smooth and abrupt changes; specifically, the stochastic Lipschitz continuity condition is only required to hold within subintervals of time separated by finitely many breakpoints. Naturally, our framework also encompasses the case of stationary time series, as well as the fundamental setting of iid sequences. In Subsection A.3, we discuss how our test can be simplified in the case of stationary time series.

3.1 Nonstationary observed processes

In this subsection, we introduce the so-called "causal representations" of the processes. That is, for each $n \in \mathbb{N}$, we will 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 inputs. This type of representation for time series has a long history, tracing back to at least Rosenblatt [Ros61] and Wiener [Wie66], though its importance for the statistical analysis of time series was first elucidated by Wu [Wu05]. What follows is most similar to the framework for high-dimensional nonstationary nonlinear time series from Mies and Steland [MS23], which in turn builds on the framework from Zhou and Wu [ZW09]. For the following assumption, 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 $(\eta_t^X, \eta_t^Y, \eta_t^Z)_{t\in\mathbb{Z}}$ is a sequence of iid random vectors. Denote the dimensions of $\eta_t^X = \eta_{t,n}^X$, $\eta_t^Y = \eta_{t,n}^Y$, $\eta_t^Z = \eta_{t,n}^Z$ respectively by $d_X^\eta = d_{X,n}^\eta$, $d_Y^\eta = d_{Y,n}^\eta$, $d_Z^\eta = d_{Z,n}^\eta$, which can change with n.

Assumption 3.1 (Causal representation of the observed processes)**.** *Assume that for each time $t \in \mathcal{T}_n$ we can represent each dimension of each of the observed processes 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 $n \in \mathbb{N}$, $(i,j,a,b) \in \mathcal{D}_n$, $t \in \mathcal{T}_n$, we assume that $G_{t,n,i}^X(\cdot)$, $G_{t,n,j}^Y(\cdot)$, $G_{t,n,k}^Z(\cdot)$ are each measurable functions from $(\mathbb{R}^{d_X^\eta})^\infty$, $(\mathbb{R}^{d_Y^\eta})^\infty$, $(\mathbb{R}^{d_Z^\eta})^\infty$ to $\mathbb{R}$ — where we endow $(\mathbb{R}^{d_X^\eta})^\infty$, $(\mathbb{R}^{d_Y^\eta})^\infty$, $(\mathbb{R}^{d_Z^\eta})^\infty$ 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.

To simplify the notation, we have not defined the input sequences for the observed processes separately for each dimension. Without loss of generality, we can define the measurable functions $G_{t,n,i}^X(\cdot)$, $G_{t,n,j}^Y(\cdot)$, $G_{t,n,k}^Z(\cdot)$ and the inputs η_t^X, η_t^Y, η_t^Z so that each dimension of the observed processes can have idiosyncratic inputs.

We will introduce several more causal representations throughout this paper. Let us state some properties that all causal representations will have to avoid repeating the same ideas each time. The causal representations will all be measurable functions on $(\mathbb{R}^{d^\eta})^\infty$ for some $d^\eta \in \mathbb{N}$, where we will always endow $(\mathbb{R}^{d^\eta})^\infty$ with the σ-algebra generated by all finite projections. The causal mechanism at a particular time $t \in \mathcal{T}_n$ with the input sequence up to a particular $s \in \mathbb{Z}$ is a well-defined random variable or vector. Similarly, the process induced by considering the input sequence up to each $s \in \mathbb{Z}$ with a fixed causal mechanism is a stationary ergodic time series, as in the previous assumption.

In view of Assumption 3.1, we have the following causal representations for the observed processes with all dimensions

$$\begin{aligned} X_{t,n} &= G_{t,n}^X(\mathcal{H}_t^X) = (G_{t,n,i}^X(\mathcal{H}_t^X))_{i\in[d_X]}, \\ Y_{t,n} &= G_{t,n}^Y(\mathcal{H}_t^Y) = (G_{t,n,j}^Y(\mathcal{H}_t^Y))_{j\in[d_Y]}, \\ Z_{t,n} &= G_{t,n}^Z(\mathcal{H}_t^Z) = (G_{t,n,k}^Z(\mathcal{H}_t^Z))_{k\in[d_Z]}. \end{aligned}$$

Also, we have causal representations for each of the dimensions $i \in [d_X]$, $j \in [d_Y]$, $k \in [d_Z]$ with

time-offsets $a \in A_i$, $b \in B_j$, $c \in C_k$

$$X_{t,n,i,a} = G^X_{t,n,i,a}(\mathcal{H}^X_{t,a}) = G^X_{t+a,n,i}(\mathcal{H}^X_{t+a}),$$
$$Y_{t,n,j,b} = G^Y_{t,n,j,b}(\mathcal{H}^Y_{t,b}) = G^Y_{t+b,n,j}(\mathcal{H}^Y_{t+b}),$$
$$Z_{t,n,k,c} = G^Z_{t,n,k,c}(\mathcal{H}^Z_{t,c}) = G^Z_{t+c,n,k}(\mathcal{H}^Z_{t+c}),$$

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 $\mathcal{H}^Z_{t,c} = (\eta^Z_{t+c}, \eta^Z_{t-1+c}, \ldots)$. We can then write the causal representation of the vectors with all dimensions and time-offsets as

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

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 $\eta^{\boldsymbol{X}}_t = \eta^X_{t+a_{\max}}$, $\eta^{\boldsymbol{Y}}_t = \eta^Y_{t+b_{\max}}$, $\eta^{\boldsymbol{Z}}_t = \eta^Z_{t+c_{\max}}$.

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 joint distribution of the nonlinear stochastic systems

$$(G^X_{t,n}(\mathcal{H}^X_s))_{t\in[n], s\in\mathbb{Z}},\ (G^Y_{t,n}(\mathcal{H}^Y_s))_{t\in[n], s\in\mathbb{Z}},\ (G^Z_{t,n}(\mathcal{H}^Z_s))_{t\in[n], s\in\mathbb{Z}}$$

under $\mathbb{P}_P$ is $P \in \mathcal{P}_n$, where the collection of distributions $\mathcal{P}_n$ can change with n. 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. Let $\mathbb{P}_P(E)$ denote the probability of an event $E \in \mathcal{B}$. 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 $\mathcal{P}_n$ which 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 $R_{P,n} = O_{\mathcal{P}}(1)$.

In the next few subsections, we will state distribution-uniform assumptions with respect to a *generic* sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$ for the observed process. Let $\mathcal{P}^{\text{CI}}_{0,n}$ be a collection of distributions for the observed process such that the null hypothesis is true, and let $(\mathcal{P}^{\text{CI}}_{0,n})_{n\in\mathbb{N}}$ be a sequence of such collections of distributions. In our main result, which we state as Theorem 3.1, we will assume that these distribution-uniform assumptions hold for a sequence of collections of distributions $(\mathcal{P}^*_{0,n})_{n\in\mathbb{N}}$, where $\mathcal{P}^*_{0,n} \subset \mathcal{P}^{\text{CI}}_{0,n}$ for each $n \in \mathbb{N}$. That is, we will make these assumptions for a subcollection of distributions for which the global null hypothesis of conditional independence holds.

3.2 Prediction processes

We introduce causal representations for the prediction processes in this subsection. For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $(i, j, a, b) \in \mathcal{D}_n$, let $\eta^{\text{algo}}_{t,n,i,a}$, $\eta^{\text{algo}}_{t,n,j,b}$ be random variables that encode the (possible) stochasticity of the statistical learning algorithms. If the learning algorithms are not stochastic, then these random variables can be ignored without loss of generality. Going forward, we will suppress the dependence of the predictors on $\eta^{\text{algo}}_{t,n,i,a}$, $\eta^{\text{algo}}_{t,n,j,b}$ to simplify the notation.

Let $\mathfrak{D}^{\hat{f}}_{t,n,i,a}$, $\mathfrak{D}^{\hat{g}}_{t,n,j,b}$ be the datasets containing the observations used to form the predictors $\hat{f}_{t,n,i,a}$, $\hat{g}_{t,n,j,b}$, and let $\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}$, $\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}$ be the corresponding input sequences. For example, if only the observations in $\mathcal{T}_n$ up to time $t \in \mathcal{T}_n$ are used to form the predictor $\hat{g}_{t,n,j,b}$, then $\mathfrak{D}^{\hat{g}}_{t,n,j,b} = (Y_{s,n,j,b}, \boldsymbol{Z}_{s,n})_{s \leq t}$ and $\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b} = (\mathcal{H}^{Y}_{t,b}, \mathcal{H}^{\boldsymbol{Z}}_{t})$. Similarly, if all of the observations in $\mathcal{T}_n$ are used (i.e. to time $\mathbb{T}^+_n$) to form the predictor $\hat{g}_{t,n,j,b}$, then $\mathfrak{D}^{\hat{g}}_{t,n,j,b} = (Y_{t,n,j,b}, \boldsymbol{Z}_{t,n})_{t \in \mathcal{T}_n}$ and $\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b} = (\mathcal{H}^{Y}_{\mathbb{T}^+_n,b}, \mathcal{H}^{\boldsymbol{Z}}_{\mathbb{T}^+_n})$.

Denote the sets of times corresponding to $\mathfrak{D}^{\hat{f}}_{t,n,i,a}$, $\mathfrak{D}^{\hat{g}}_{t,n,j,b}$ by $\mathcal{T}^{\hat{f}}_{t,n,i,a}$, $\mathcal{T}^{\hat{g}}_{t,n,j,b}$, respectively, and let $T^{\hat{f}}_{t,n,i,a} = |\mathcal{T}^{\hat{f}}_{t,n,i,a}|$, $T^{\hat{g}}_{t,n,j,b} = |\mathcal{T}^{\hat{g}}_{t,n,j,b}|$ be the cardinalities. For each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$ let $\mathcal{M}(\mathcal{Z}, \mathcal{Y}) \subseteq \mathcal{Y}^{\mathcal{Z}}$ and $\mathcal{M}(\mathcal{Z}, \mathcal{X}) \subseteq \mathcal{X}^{\mathcal{Z}}$, where $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$, and $\mathcal{Z} = \mathbb{R}^{\boldsymbol{d_Z}}$. Note that $\boldsymbol{d_Z}$ can grow with n as discussed in Subsection 2.1, although we suppress this in the notation.

Assumption 3.2 (Causal representations of predictors)**.** *For each $n \in \mathbb{N}$, $(i,j,a,b) \in \mathcal{D}_n$, assume that the sequences of statistical learning algorithms $\mathcal{A}^{\hat{f}}_{n,i,a} = (\mathcal{A}^{\hat{f}}_{t,n,i,a})_{t \in \mathcal{T}_n}$, $\mathcal{A}^{\hat{g}}_{n,j,b} = (\mathcal{A}^{\hat{g}}_{t,n,j,b})_{t \in \mathcal{T}_n}$ consist of the (Borel) measurable functions*

$$\mathcal{A}^{\hat{f}}_{t,n,i,a} : \begin{cases} (\mathcal{Z} \times \mathcal{X})^{T^{\hat{f}}_{t,n,i,a}} \to \mathcal{M}(\mathcal{Z}, \mathcal{X}) \\ \mathfrak{D}^{\hat{f}}_{t,n,i,a} \to \hat{f}_{t,n,i,a}, \end{cases}$$

and

$$\mathcal{A}^{\hat{g}}_{t,n,j,b} : \begin{cases} (\mathcal{Z} \times \mathcal{Y})^{T^{\hat{g}}_{t,n,j,b}} \to \mathcal{M}(\mathcal{Z}, \mathcal{Y}) \\ \mathfrak{D}^{\hat{g}}_{t,n,j,b} \mapsto \hat{g}_{t,n,j,b}, \end{cases}$$

such that the predictors have the causal representations

$$\hat{f}_{t,n,i,a} = G^{\mathcal{A}^{\hat{f}}}_{t,n,i,a}(\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}),$$
$$\hat{g}_{t,n,j,b} = G^{\mathcal{A}^{\hat{g}}}_{t,n,j,b}(\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}),$$

in view of Assumption 3.1. $G^{\mathcal{A}^{\hat{f}}}_{t,n,i,a}(\cdot)$, $G^{\mathcal{A}^{\hat{g}}}_{t,n,j,b}(\cdot)$ are measurable functions so that $G^{\mathcal{A}^{\hat{f}}}_{t,n,i,a}(\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a})$, $G^{\mathcal{A}^{\hat{g}}}_{t,n,j,b}(\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b})$ are well-defined function-valued random variables.

We make the following assumption for the predictions and prediction errors for some sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$.

Assumption 3.3 (Causal representations for predictions and prediction errors)**.** *Assume that the predictors $\hat{f}_{t,n,i,a}$, $\hat{g}_{t,n,j,b}$ are (Borel) measurable functions from $\mathbb{R}^{\boldsymbol{d_Z}}$ to $\mathbb{R}$ such that for each $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $(i,j,a,b) \in \mathcal{D}_n$ we can represent the predictions and prediction errors as*

$$\hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}) = G^{\hat{f}}_{t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a}) = [\mathcal{A}^{\hat{f}}_{t,n,i,a}(X_{n,i,a}, \boldsymbol{Z}_n)](\boldsymbol{Z}_{t,n}),$$
$$\hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}) = G^{\hat{g}}_{t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b}) = [\mathcal{A}^{\hat{g}}_{t,n,j,b}(Y_{n,j,b}, \boldsymbol{Z}_n)](\boldsymbol{Z}_{t,n}),$$

and

$$\hat{w}^{f}_{P,t,n,i,a} = G^{\hat{w}^f}_{P,t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a}) = f_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}) - \hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}),$$
$$\hat{w}^{g}_{P,t,n,j,b} = G^{\hat{w}^g}_{P,t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b}) = g_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}) - \hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}),$$

in view of Assumptions 3.1, 3.2, where the input sequences are

$$\mathcal{H}^{\hat{f}}_{t,a} = (\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}, \mathcal{H}^{\boldsymbol{Z}}_{t}), \ \mathcal{H}^{\hat{g}}_{t,b} = (\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}, \mathcal{H}^{\boldsymbol{Z}}_{t}). \tag{10}$$

Also, assume that for all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $(i,j,a,b) \in \mathcal{D}_n$ there exists some $q \geq 2$ such that

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|\hat{w}^{f}_{P,t,n,i,a}|^q) < \infty, \ \sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|\hat{w}^{g}_{P,t,n,j,b}|^q) < \infty.$$

$G^{\hat{f}}_{t,n,i,a}(\cdot)$, $G^{\hat{w}^f}_{P,t,n,i,a}(\cdot)$ and $G^{\hat{g}}_{t,n,j,b}(\cdot)$, $G^{\hat{w}^g}_{P,t,n,j,b}(\cdot)$ are measurable functions such that $G^{\hat{f}}_{t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a})$, $G^{\hat{g}}_{t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b})$ and $G^{\hat{w}^f}_{P,t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a})$, $G^{\hat{w}^g}_{P,t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b})$ are well-defined real-valued random variables.

In view of Assumption 3.3, we have the following causal representation for all dimensions and time-offsets of the prediction errors

$$\hat{\boldsymbol{w}}^{\boldsymbol{f}}_{P,t,n} = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{f}}}_t) = (\hat{w}^{f}_{P,t,n,i,a})_{i\in[d_X],a\in A_i},$$
$$\hat{\boldsymbol{w}}^{\boldsymbol{g}}_{P,t,n} = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{g}}}_t) = (\hat{w}^{g}_{P,t,n,j,b})_{j\in[d_Y],b\in B_j},$$

where $\mathcal{H}^{\hat{\boldsymbol{f}}}_t = (\mathcal{H}^{\hat{f}}_{t,a})_{a\in A}$ and $\mathcal{H}^{\hat{\boldsymbol{g}}}_t = (\mathcal{H}^{\hat{g}}_{t,b})_{b\in B}$.

3.3 Nonstationary error processes

In this subsection, we will introduce the causal representations of the error processes of the high-dimensional nonstationary time series. For the next assumption, for each $a \in A$, $b \in B$, define the input sequences

$$\mathcal{H}^{\varepsilon}_{t,a} = (\eta^{\varepsilon}_{t,a}, \eta^{\varepsilon}_{t,a-1}, \ldots),\ \mathcal{H}^{\xi}_{t,b} = (\eta^{\xi}_{t,b}, \eta^{\xi}_{t,b-1}, \ldots), \tag{11}$$

where $(\eta^{\varepsilon}_{t,a}, \eta^{\xi}_{t,b})_{t\in\mathbb{Z}}$ is a sequence of iid random vectors. For the following assumption, denote the dimension of $\eta^{\varepsilon}_{t,a} = \eta^{\varepsilon}_{t,a,n}$ by $d^{\eta}_{\varepsilon} = d^{\eta}_{\varepsilon,n}$, and the dimension of $\eta^{\xi}_{t,b} = \eta^{\xi}_{t,b,n}$ by $d^{\eta}_{\xi} = d^{\eta}_{\xi,n}$, both of which can change with n.

Assumption 3.4 (Causal representations of the error processes)**.** *Assume that for each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $(i,j,a,b) \in \mathcal{D}_n$, $t \in \mathcal{T}_n$, we can represent the error processes from Subsection 2.3 as*

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

with $\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{g}}}_t) = 0$ and $\mathbb{E}_P(\xi_{P,t,n,j,b}|\mathcal{H}^{\hat{\boldsymbol{f}}}_t) = 0$, where the input sequences $\mathcal{H}^{\hat{\boldsymbol{g}}}_t$, $\mathcal{H}^{\hat{\boldsymbol{f}}}_t$ are from Subsection 3.2. $G^{\varepsilon}_{P,t,n,i,a}(\cdot)$ and $G^{\xi}_{P,t,n,j,b}(\cdot)$ are measurable functions from $(\mathbb{R}^{d^{\eta}_{\varepsilon}})^{\infty}$ and $(\mathbb{R}^{d^{\eta}_{\xi}})^{\infty}$, respectively, to $\mathbb{R}$ — where we endow $(\mathbb{R}^{d^{\eta}_{\varepsilon}})^{\infty}$ and $(\mathbb{R}^{d^{\eta}_{\xi}})^{\infty}$ with the σ-algebra generated by all finite projections — so 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.

We have not defined the input sequences for the error processes separately for each dimension. Without loss of generality, the measurable functions $G^{\varepsilon}_{P,t,n,i,a}(\cdot)$, $G^{\xi}_{P,t,n,j,b}(\cdot)$ and inputs $\eta^{\varepsilon}_{t,a}, \eta^{\xi}_{t,b}$ can be defined so that each dimension of the error processes has idiosyncratic inputs.

In view of the causal representations of the univariate error processes, we have the following causal representations for the high-dimensional nonstationary vector-valued error processes

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

where $\mathcal{H}^{\boldsymbol{\varepsilon}}_t = (\eta^{\boldsymbol{\varepsilon}}_t, \eta^{\boldsymbol{\varepsilon}}_{t-1}, \ldots)$, $\mathcal{H}^{\boldsymbol{\xi}}_t = (\eta^{\boldsymbol{\xi}}_t, \eta^{\boldsymbol{\xi}}_{t-1}, \ldots)$ with $\eta^{\boldsymbol{\varepsilon}}_t = (\eta^{\varepsilon}_{t,a})_{a\in A}$, $\eta^{\boldsymbol{\xi}}_t = (\eta^{\xi}_{t,b})_{b\in B}$ for each $t \in \mathbb{Z}$. Similarly, for each dimension/time-offset tuple $m = (i,j,a,b) \in \mathcal{D}_n$ the error products at time t 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 $\mathcal{H}^{R}_{t,m} = (\eta^{R}_{t,m}, \eta^{R}_{t-1,m}, \ldots)$ with $\eta^{R}_{t,m} = (\eta^{\varepsilon}_{t,a}, \eta^{\varepsilon}_{t,b})^{\top}$ for each $t \in \mathbb{Z}$. Also, we have the following representation for the high-dimensional nonstationary $\mathbb{R}^{D_n}$-valued process of all the products of errors

$$\boldsymbol{R}_{P,t,n} = \boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_t) = (G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{t,m}))_{m=(i,j,a,b)\in\mathcal{D}_n}$$

where $\mathcal{H}^{\boldsymbol{R}}_t = (\eta^{\boldsymbol{R}}_t, \eta^{\boldsymbol{R}}_{t-1}, \ldots)$ and $\eta^{\boldsymbol{R}}_t = (\eta^{\boldsymbol{\varepsilon}}_t, \eta^{\boldsymbol{\xi}}_t)^{\top}$ for each $t \in \mathbb{Z}$. Note that for a fixed $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, and $n \in \mathbb{N}$ we have that $\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_s)$ is a well-defined high-dimensional random vector for each $s \in \mathbb{Z}$ and $(\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_s))_{s\in\mathbb{Z}}$ is a high-dimensional stationary ergodic $\mathbb{R}^{D_n}$-valued time series.

In view of Assumptions 3.3 and 3.4, for $m = (i,j,a,b) \in \mathcal{D}_n$, we can represent the products of the errors and prediction errors as

$$\hat{w}^{g,\varepsilon}_{P,t,n,m} = G^{\hat{w}^{g,\varepsilon}}_{P,t,n,m}(\mathcal{H}^{\hat{w}^{g,\varepsilon}}_{t,m}) = \hat{w}^{g}_{P,t,n,j,b}\varepsilon_{P,t,n,i,a},$$
$$\hat{w}^{f,\xi}_{P,t,n,m} = G^{\hat{w}^{f,\xi}}_{P,t,n,m}(\mathcal{H}^{\hat{w}^{f,\xi}}_{t,m}) = \hat{w}^{f}_{P,t,n,i,a}\xi_{P,t,n,j,b},$$

with $\mathcal{H}^{\hat{w}^{g,\varepsilon}}_{t,m} = (\mathcal{H}^{\hat{g}}_{t,b}, \mathcal{H}^{\varepsilon}_{t,a})$ and $\mathcal{H}^{\hat{w}^{f,\xi}}_{t,m} = (\mathcal{H}^{\hat{f}}_{t,a}, \mathcal{H}^{\xi}_{t,b})$. 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

$$\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t}) = (\hat{w}^{g,\varepsilon}_{P,t,n,m})_{m=(i,j,a,b)\in\mathcal{D}_n},$$
$$\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,t,n} = \boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{t}) = (\hat{w}^{f,\xi}_{P,t,n,m})_{m=(i,j,a,b)\in\mathcal{D}_n},$$

with $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t} = (\mathcal{H}^{\hat{\boldsymbol{g}}}_{t}, \mathcal{H}^{\boldsymbol{\varepsilon}}_{t})$ and $\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{t} = (\mathcal{H}^{\hat{\boldsymbol{f}}}_{t}, \mathcal{H}^{\boldsymbol{\xi}}_{t})$, where we have suppressed the dependence on n. $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}(\cdot)$ and $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}(\cdot)$ are measurable functions such that $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{s})$, $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{s})$ are well-defined high-dimensional random vectors for each $s \in \mathbb{Z}$ and $(\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{s}))_{s\in\mathbb{Z}}$, $(\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{P,t,n}(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}}_{s}))_{s\in\mathbb{Z}}$ are high-dimensional stationary ergodic time series.

3.4 Assumptions on dependence and nonstationarity

In this subsection, we impose mild assumptions on the rate of decay in temporal dependence and the degree of nonstationarity of the *error processes*. Crucially, these assumptions are stated in a distribution-uniform manner, which is essential for applying the strong Gaussian approximation in Section B. This will be further elaborated upon in Subsection 3.5.

We quantify temporal dependence using the functional dependence measure of Wu [Wu05]. Let $(\tilde{\eta}^{\varepsilon}_{t,a}, \tilde{\eta}^{\xi}_{t,b})_{t\in\mathbb{Z}}$ be an iid copy of $(\eta^{\varepsilon}_{t,a}, \eta^{\xi}_{t,b})_{t\in\mathbb{Z}}$. Denote the set of well-defined tuples of error processes, dimensions, and time-offsets by

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

For any tuple $(e, l, d) \in \mathbb{E}$ corresponding to a well-defined combination of an error process, dimension, and time-offset, define

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

to be $\mathcal{H}^{e}_{t,d}$ with the input $\eta^{e}_{t-h,d}$ replaced with the iid copy $\tilde{\eta}^{e}_{t-h,d}$. Similarly, define $\tilde{\mathcal{H}}^{R}_{t,m,h}$ as $\mathcal{H}^{R}_{t,m}$ with the input $\eta^{R}_{t-h,m}$ replaced with the iid copy $\tilde{\eta}^{R}_{t-h,m}$ for $m = (i, j, a, b) \in \mathcal{D}_n$, and define $\tilde{\mathcal{H}}^{\boldsymbol{R}}_{t,h}$ as $\mathcal{H}^{\boldsymbol{R}}_{t}$ with the input $\eta^{\boldsymbol{R}}_{t-h}$ replaced with the iid copy $\tilde{\eta}^{\boldsymbol{R}}_{t-h}$. Next, we define the functional dependence measures.

Definition 3.1 (Functional dependence measure)**.** *We define the following measures of temporal dependence for each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, and $t \in \mathcal{T}_n$. First, define the L^∞ version of the functional dependence measure for the error processes $G^{e}_{P,t,n,l,d}(\mathcal{H}^{e}_{t,d})$ for each $(e, l, d) \in \mathbb{E}$ with $h \in \mathbb{N}_0$ as*

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

Second, define the functional dependence measures for the processes of error products $G^{R}_{P,t,n,m}(\mathcal{H}^{R}_{t,m})$ for each $m = (i, j, a, b) \in \mathcal{D}_n$ with $h \in \mathbb{N}_0$, $q \geq 1$, as

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

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

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

For some sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we make the following assumption about the temporal dependence. We only require the relatively mild assumption that it decays polynomially, rather than geometrically. Note that we will often write the time of the input sequence as 0 when it does not matter due to stationarity.

Assumption 3.5 (Distribution-uniform decay of temporal dependence)**.** *Assume that there exist $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 1$ such that for all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $(e, l, d) \in \mathbb{E}$, it holds that*

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

For additional control in terms of the product of errors alone, also assume that there exist $\bar{\Theta}^R > 0$, $\bar{\beta}^R > 3$, $\bar{q}^R > 4$, *such that for all* $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $m = (i, j, a, b) \in \mathcal{D}_n$, *it holds that*

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

By Jensen's inequality, we have for all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, that

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(||\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_0)||_2^{\bar{q}^R})]^{1/\bar{q}^R} \leq D_n^{\frac{1}{2}} \bar{\Theta}^R, \qquad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{R}}_{P,t,n}(h, \bar{q}^R, 2) \leq D_n^{\frac{1}{2}} \bar{\Theta}^R \cdot (h \vee 1)^{-\bar{\beta}^R}, \qquad h \geq 0.$$

A few remarks about Assumption 3.5 are in order. First, note that the constants in the previous assumption do not depend on n. Second, the dependence assumption for the errors in terms of $\theta^{e,\infty}_{P,t,n,l,d}(h)$ can be weakened; see Subsection A.6 for more discussion.

Next, for some sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$, we make the following assumption to control the nonstationarity of the process of error products.

Assumption 3.6 (Distribution-uniform total variation condition for nonstationarity)**.** *Recall* $\bar{\Theta}^R > 0$ *from Assumption 3.5. Assume that for each* $n \in \mathbb{N}$, *there exists a constant* $\bar{\Gamma}^R_n \geq 1$ *such that*

$$\sup_{P \in \mathcal{P}_n} \left(\sum_{t=\mathbb{T}_n^- + 1}^{\mathbb{T}_n^+} \left(\mathbb{E}_P ||\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_0) - \boldsymbol{G}^{\boldsymbol{R}}_{P,t-1,n}(\mathcal{H}^{\boldsymbol{R}}_0)||_2^2\right)^{1/2} \right) \leq \bar{\Theta}^R \bar{\Gamma}^R_n.$$

3.5 Main theoretical result for dGCM

In this subsection, we present the theoretical result that justifies the bootstrap procedure described in Algorithm 1. This result relies on time-varying nonlinear regression between nonstationary time series and the distribution-uniform strong Gaussian approximation from Section B applied to the process of error products. The approximating nonstationary Gaussian process has a time-varying covariance structure, which is explicitly characterized by *local long-run* covariance matrices.

Definition 3.2 (Local long-run covariance matrices of process of error products)**.** *For each* $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $n \in \mathbb{N}$, *define the local long-run covariance matrix* $\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n} \in \mathbb{R}^{D_n \times D_n}$ *of the* $\mathbb{R}^{D_n}$*-valued stationary process* $(\boldsymbol{G}^{\boldsymbol{R}}_{P,t,n}(\mathcal{H}^{\boldsymbol{R}}_t))_{t \in \mathbb{Z}}$ *by*

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

In view of the Gaussian approximation theory developed in Mies and Steland [MS23], we only require an estimator of the *cumulative* covariance matrices of the error products

$$Q^{\boldsymbol{R}}_{P,t,n} = \sum_{s=\mathbb{T}_n^-}^{t} \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,s,n},$$

rather than the local long-run covariance matrices at each time individually. This is critical for the practical applicability of our method, as estimating individual local long-run covariance matrices can be extremely challenging in practice. Specifically, we use the estimator

$$\hat{Q}^{\boldsymbol{R}}_{t,n} = \sum_{s=L_n+\mathbb{T}_n^- - 1}^{t} \frac{1}{L_n} \left(\sum_{r=s-L_n+1}^{s} \hat{\boldsymbol{R}}_{r,n} \right)^{\otimes 2}, \tag{12}$$

for some lag-window size $L_n \in \mathbb{N}$. We discuss how to select L_n in practice using the minimum volatility method in Subsection 5.1. Going forward, denote $Q^{\boldsymbol{R}}_{P,n} = (Q^{\boldsymbol{R}}_{P,t,n})_{t \in \mathcal{T}_{n,L}}$ and $\hat{Q}^{\boldsymbol{R}}_n = (\hat{Q}^{\boldsymbol{R}}_{t,n})_{t \in \mathcal{T}_{n,L}}$, where

$$\mathcal{T}_{n,L} = \{L_n + \mathbb{T}_n^- - 1, \ldots, \mathbb{T}_n^+ - 1, \mathbb{T}_n^+\}.$$

To account for the estimation errors for the time-varying regression functions and the cumulative covariance matrices, as well as the error for the Gaussian approximation, we introduce offsets $\tau_n \to 0$, $\nu_n \to 0$ so that $\tau_n = o(\log^{-(1+\delta)}(T_n))$ for some $\delta > 0$ and

$$\nu_n \gg \log(T_n) D_n \left[\left(\frac{D_n}{T_n} \right)^{2\xi(\bar{q}^R, \bar{\beta}^R)} + \tau_n^{-2} \left(\varphi_{n,1} + \varphi_{n,2} \right) \right], \tag{13}$$

where

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

comes from the covariance estimation error,

$$\varphi_{n,2} = \tau_n^{\frac{7}{2}} D_n^{-\frac{5}{4}} + \tau_n^7 D_n^{-\frac{5}{2}},$$

comes from the time-varying regression function estimation errors. Also, the lag-window size L_n from (12) must satisfy $L_n \asymp T_n^{\zeta}$ for some $\zeta \in (0, \frac{1}{2})$ so that $\tau_n^{-6} D_n^2 L_n^{-1} = o(1)$ and $\bar{\Gamma}_n^R T_n^{-1} D_n^2 \tau_n^{-6} L_n^{\frac{1}{2}} = o(1)$, where $\bar{\Gamma}_n^R$ is from Assumption 3.6. We see that the offsets depend on the number of observations T_n from Subsection 2.1, the intrinsic dimensionality D_n from Subsection 2.1, the degree of nonstationarity $\bar{\Gamma}_n^R$ from Assumption 3.6, and the lag-window parameter L_n from (12). $\xi(\bar{q}^R, \bar{\beta}^R)$ is a rate defined in Section B that depends on the constants $\bar{\beta}^R$, $\bar{q}^R$ from Assumption 3.5.

The following result establishes the theoretical validity of our bootstrap-based testing procedure described in Algorithm 1, provided that the previously stated assumptions hold and the prediction errors

$$\hat{w}^f_{P,t,n,i,a} = f_{P,t,n,i,a}(\boldsymbol{Z}_{t,n}) - \hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n}),$$
$$\hat{w}^g_{P,t,n,j,b} = g_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}) - \hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n}),$$

converge to zero sufficiently fast, in some sense. If it were known, we could correctly calibrate our test with the (random) quantile function $\hat{q}$ of $S_{n,p}(\breve{\boldsymbol{R}}_n)$, where $\breve{\boldsymbol{R}}_n = (\breve{\boldsymbol{R}}_{t,n})_{t \in \mathcal{T}_{n,L}}$ and $\breve{\boldsymbol{R}}_{t,n} \sim \mathcal{N}(0, \hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n})$ for all $t \in \mathcal{T}_{n,L}$. In practice, $\hat{q}$ is numerically approximated by conducting a large number of Monte Carlo simulations, and we use $\hat{q}^{\text{boot}}$ from Algorithm 1 in its place.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 related to the temporal dependence and nonstationarity of the processes all hold for the sequence of collections of distributions $(\mathcal{P}^*_{0,n})_{n \in \mathbb{N}}$, where $\mathcal{P}^*_{0,n} \subset \mathcal{P}^{\text{CI}}_{0,n}$ for each $n \in \mathbb{N}$. Further, suppose that*

$$\sup_{P \in \mathcal{P}^*_{0,n}} \max_{(i,j,a,b) \in \mathcal{D}_n} \max_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left| \hat{w}^f_{P,t,n,i,a} \right|^2 \right)^{\frac{1}{2}} \mathbb{E}_P \left(\left| \hat{w}^g_{P,t,n,j,b} \right|^2 \right)^{\frac{1}{2}} = o(T_n^{-\frac{1}{2}} \tau_n^7 D_n^{-\frac{3}{2}}),$$

$$\sup_{P \in \mathcal{P}^*_{0,n}} \max_{i \in [d_X], a \in A_i} \max_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left| \hat{w}^f_{P,t,n,i,a} \right|^2 \right)^{\frac{1}{2}} = o(\tau_n^7 D_n^{-\frac{5}{2}}),$$

$$\sup_{P \in \mathcal{P}^*_{0,n}} \max_{j \in [d_Y], b \in B_j} \max_{t \in \mathcal{T}_n} \mathbb{E}_P \left(\left| \hat{w}^g_{P,t,n,j,b} \right|^2 \right)^{\frac{1}{2}} = o(\tau_n^7 D_n^{-\frac{5}{2}}).$$

If the offsets $\tau_n \to 0$ and $\nu_n \to 0$ are chosen such that condition (13) holds, then we have

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(S_{n,p}(\hat{\boldsymbol{R}}_n) > \hat{q}_{1-\alpha+\nu_n} + \tau_n \right) \leq \alpha.$$

The previous result demonstrates that the dGCM test possesses a property known as *rate double robustness*. This property means that we only require modest convergence rates for the *products* of the prediction errors, rather than for each prediction error individually. This feature of the dGCM test can be especially useful in the contexts of causal discovery for time-lagged effects and variable selection in time series forecasting. In these applications, a faster convergence rate of a nowcasting model can compensate for a slower convergence rate of a forecasting model.

4 dGCM with Sieve Time-Varying Regression (Sieve-dGCM)

The purpose of this section is to demonstrate that the convergence rates required by Theorem 3.1 for estimating the time-varying regression functions can actually be achieved. To show this, we consider an instantiation of the dynamic generalized covariance measure (dGCM) test based on the sieve time-varying nonlinear regression estimator from Ding and Zhou [DZ21]. We refer to this instantiation of the dGCM test from Algorithm 1 that uses the predictions from the sieve estimator as the Sieve-dGCM test.

We prove that, under mild assumptions about the temporal dependence and nonstationarity of the processes, the sieve estimator achieves the required convergence rates and the Sieve-dGCM test has asymptotic Type-I error control. In Section 5, we study the finite sample performance of the Sieve-dGCM test. Along the way, in Subsection 5.1, we introduce a novel cross-validation scheme which we use for selecting the parameters of the sieve estimator.

In this section, we use the framework of *locally stationary time series* [Dah97; ZW09; Dah12; DRW19]. This is a well-studied class of nonstationary time series that fits within the general triangular array framework for nonstationary time series from Section 3. We note that there are several other time-varying regression estimators for locally stationary time series; see Zhang and Wu [ZW15], Yousuf and Ng [YN21], and Chen, Smetanina, and Wu [CSW22] and the references therein. We discuss extensions to *piecewise* locally stationary time series in Subsection A.5.

4.1 Setting and notation

We follow Dahlhaus [Dah97] in rescaling time to the unit interval $t/n \in [0, 1]$, 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. Instead, as n increases we get more observations about each *local structure* of the nonstationary process. Zhou and Wu [ZW09] introduced the framework for representing locally stationary time series as nonlinear functions of iid inputs as in Wu [Wu05].

We use the same notation as Subsection 2.1, with the only difference being that we fix the number of dimensions d_Z and time-offsets C_k for each dimension $k \in [d_Z]$. We still allow the number of dimensions $d_X = d_{X,n}$, $d_Y = d_{Y,n}$ and time-offsets A_i, B_j for each $i \in [d_X]$, $j \in [d_Y]$ to grow with n. Define A, B, C as the collection of all time-offsets as in Subsection 2.1, where $A = A_n$, $B = B_n$ and C is fixed. We emphasize that there is no inherent necessity for fixing the number of dimensions d_Z and time-offsets C_k. Our reason for doing this is because we want to leverage the *existing* theoretical results for the sieve estimator from Ding and Zhou [DZ21]. Future investigations can study the performance of the sieve estimator in the high-dimensional setting, so that we can allow the number of dimensions d_Z and time-offsets C_k to grow with n.

We will still use the notation $\mathcal{T}_n$ for the subset of original times in which *all* time-offsets of each dimension of $X_{t,n}$, $Y_{t,n}$, and $Z_{t,n}$ are actually observed,

$$\mathcal{T}_n = \{1 - \min(a_{\min}, b_{\min}, c_{\min}), n - \max(a_{\max}, b_{\max}, c_{\max})\} \subseteq \{1, \dots, n\}.$$

Also, we will still denote $T_n = |\mathcal{T}_n|$, $\mathbb{T}_n^- = \min(\mathcal{T}_n)$, and $\mathbb{T}_n^+ = \max(\mathcal{T}_n)$. Similarly, denote the corresponding interval of rescaled times in which all time-offsets are well-defined by

$$\mathcal{U}_n = \left[\frac{1}{n} - \frac{\min(a_{\min}, b_{\min}, c_{\min})}{n}, 1 - \frac{\max(a_{\max}, b_{\max}, c_{\max})}{n}\right] \subset [0, 1],$$

and denote $\mathbb{U}_n^- = \min(\mathcal{U}_n)$, and $\mathbb{U}_n^+ = \max(\mathcal{U}_n)$.

Recall the index set containing the dimensions and time-offsets of interest

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

where A_i, B_j are the time-offsets for dimensions $i \in [d_X]$, $j \in [d_Y]$. Again, we will often refer to the dimension/time-offset tuple by $m = (i, j, a, b) \in \mathcal{D}_n$ to lighten the notation. Denote cardinality $D_n = |\mathcal{D}_n|$ which may grow with n.

4.2 Locally stationary observed processes

Next, we introduce the causal representation of locally stationary processes, which is most similar to Ding and Zhou [DZ21] and Example 3 in Mies and Steland [MS23]. This representation is different than the previous causal representation from Assumption 3.1, because we now assume that the nonlinear stochastic system is well-defined for *all* rescaled times. For the following assumption, 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 $(\eta_t^X, \eta_t^Y, \eta_t^Z)_{t\in\mathbb{Z}}$ is an iid sequence of random vectors. Denote the dimensions of $\eta_t^X = \eta_{t,n}^X$, $\eta_t^Y = \eta_{t,n}^Y$, $\eta_t^Z = \eta_{t,n}^Z$ respectively by $d_X^\eta = d_{X,n}^\eta$, $d_Y^\eta = d_{Y,n}^\eta$, $d_Z^\eta = d_{Z,n}^\eta$, which can change with n.

Assumption 4.1 (Causal representation of locally stationary processes)**.** *Assume that we can represent each dimension of each of the observed processes 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 have $X_{t,n,i} = \tilde{X}_{t,n,i}(t/n)$, $Y_{t,n,j} = \tilde{Y}_{t,n,j}(t/n)$, $Z_{t,n,k} = \tilde{Z}_{t,n,k}(t/n)$.

For each $n \in \mathbb{N}$, $(i,j,a,b) \in \mathcal{D}_n$, $t \in \mathcal{T}_n$, 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 measurable functions from $(\mathbb{R}^{d_X^\eta})^\infty$, $(\mathbb{R}^{d_Y^\eta})^\infty$, $(\mathbb{R}^{d_Z^\eta})^\infty$, respectively, to $\mathbb{R}$ — where we endow $(\mathbb{R}^{d_X^\eta})^\infty$, $(\mathbb{R}^{d_Y^\eta})^\infty$, $(\mathbb{R}^{d_Z^\eta})^\infty$ 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.

As in Subsection 3.1, we have not defined the input sequences for the observed processes separately for each dimension. However, without loss of generality, we can define the measurable functions $\tilde{G}_{n,i}^X(u,\cdot)$, $\tilde{G}_{n,j}^Y(u,\cdot)$, $\tilde{G}_{n,k}^Z(u,\cdot)$ and the inputs η_t^X, η_t^Y, η_t^Z so that each dimension of the observed processes can have idiosyncratic inputs.

In light of Assumption 4.1, we have the following causal representations for all dimensions with no time-offsets

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

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

$$\begin{aligned}
\tilde{X}_{t,n,i,a}(u) &= \tilde{G}_{n,i,a}^X(u, \mathcal{H}_{t,a}^X) = \tilde{G}_{n,i}^X(u + a/n, \mathcal{H}_{t+a}^X),\\
\tilde{Y}_{t,n,j,b}(u) &= \tilde{G}_{n,j,b}^Y(u, \mathcal{H}_{t,b}^Y) = \tilde{G}_{n,j}^Y(u + b/n, \mathcal{H}_{t+b}^Y),\\
\tilde{Z}_{t,n,k,c}(u) &= \tilde{G}_{n,k,c}^Z(u, \mathcal{H}_{t,c}^Z) = \tilde{G}_{n,k}^Z(u + c/n, \mathcal{H}_{t+c}^Z),
\end{aligned}$$

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

$$\begin{aligned}
\tilde{\boldsymbol{X}}_{t,n}(u) &= \tilde{\boldsymbol{G}}_n^{\boldsymbol{X}}(u, \mathcal{H}_t^{\boldsymbol{X}}) = (\tilde{G}_{n,i,a}^X(u, \mathcal{H}_{t,a}^X))_{i\in[d_X], a\in A_i},\\
\tilde{\boldsymbol{Y}}_{t,n}(u) &= \tilde{\boldsymbol{G}}_n^{\boldsymbol{Y}}(u, \mathcal{H}_t^{\boldsymbol{Y}}) = (\tilde{G}_{n,j,b}^Y(u, \mathcal{H}_{t,b}^Y))_{j\in[d_Y], b\in B_j},\\
\tilde{\boldsymbol{Z}}_{t,n}(u) &= \tilde{\boldsymbol{G}}_n^{\boldsymbol{Z}}(u, \mathcal{H}_t^{\boldsymbol{Z}}) = (\tilde{G}_{n,k,c}^Z(u, \mathcal{H}_{t,c}^Z))_{k\in[d_Z], c\in C_k},
\end{aligned}$$

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 $\eta_t^{\boldsymbol{X}} = \eta_{t+a_{\max}}^{X}$, $\eta_t^{\boldsymbol{Y}} = \eta_{t+b_{\max}}^{Y}$, $\eta_t^{\boldsymbol{Z}} = \eta_{t+c_{\max}}^{Z}$, so that we have $\boldsymbol{X}_{t,n} = \tilde{\boldsymbol{X}}_{t,n}(t/n)$, $\boldsymbol{Y}_{t,n} = \tilde{\boldsymbol{Y}}_{t,n}(t/n)$, $\boldsymbol{Z}_{t,n} = \tilde{\boldsymbol{Z}}_{t,n}(t/n)$ for the observed sequence including all dimensions and time-offsets.

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 joint distribution of the nonlinear stochastic systems

$$(\tilde{G}_n^X(u, \mathcal{H}_t^X))_{u \in [0,1], t \in \mathbb{Z}}, \ (\tilde{G}_n^Y(u, \mathcal{H}_t^Y))_{u \in [0,1], t \in \mathbb{Z}}, \ (\tilde{G}_n^Z(u, \mathcal{H}_t^Z))_{u \in [0,1], t \in \mathbb{Z}}$$

under $\mathbb{P}_P$ is $P \in \mathcal{P}_n$, where the collection of distributions $\mathcal{P}_n$ can change with n. 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.

We use the same null hypotheses of conditional independence as those in Subsection 2.2. Again, for each $n \in \mathbb{N}$, we denote the collection of distributions such that the null hypothesis is true by $\mathcal{P}_{0,n}^{\mathrm{CI}}$. In the locally stationary time series setting, the null hypothesis

$$X_{t,n,i,a} \perp\!\!\!\perp Y_{t,n,j,b} \mid \boldsymbol{Z}_{t,n} \text{ for all } t \in \mathcal{T}_n, \text{ for all } (i,j,a,b) \in \mathcal{D}_n, \tag{14}$$

can be written equivalently as

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

where $\mathcal{D}_n$ only contains a single dimension/time-offset tuple in the univariate setting.

We will state more assumptions in the next several subsections for a generic sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$. In Theorem 4.1, we will assume that these conditions hold for the sequence of collections of distributions $(\mathcal{P}_{0,n}^*)_{n \in \mathbb{N}}$, where $\mathcal{P}_{0,n}^* \subset \mathcal{P}_{0,n}^{\mathrm{CI}}$ for each $n \in \mathbb{N}$. Note that we make stronger assumptions in this section than in Section 3 to ensure that the sieve estimator satisfies the convergence rate requirement of Theorem 3.1.

4.3 Sieve time-varying nonlinear regression estimator

For a given sample size $n \in \mathbb{N}$, distribution $P \in \mathcal{P}_n$, time $t \in \mathcal{T}_n$, and dimension/time-offset tuple $(i,j,a,b) \in \mathcal{D}_n$, we consider the time-varying nonlinear regression model

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

where $f_{P,n,i,a}(u, \boldsymbol{z})$ and $g_{P,n,j,b}(u, \boldsymbol{z})$ are smooth functions of rescaled time u and covariate values $\boldsymbol{z}$ with $f_{P,n,i,a}(t/n, \boldsymbol{z}) = \mathbb{E}_P(X_{t,n,i,a} | \boldsymbol{Z}_{t,n} = \boldsymbol{z})$ and $g_{P,n,j,b}(t/n, \boldsymbol{z}) = \mathbb{E}_P(Y_{t,n,j,b} | \boldsymbol{Z}_{t,n} = \boldsymbol{z})$. We emphasize that the functions $f_{P,n,i,a}(u, \boldsymbol{z})$ and $g_{P,n,j,b}(u, \boldsymbol{z})$ depend on rescaled time u rather than "real time" t, as in the literature on nonparametric regression for locally stationary time series [Vog12; ZW15; YN21; CSW22; DZ21]. For $m = (i,j,a,b) \in \mathcal{D}_n$, denote the error products at time t by

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

and the corresponding residual products by

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

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

The estimates $\hat{f}_{t,n,i,a}$ and $\hat{g}_{t,n,j,b}$ of the functions $f_{P,n,i,a}$ and $g_{P,n,j,b}$ are formed by regressing $(X_{t,n,i,a})_{t \in \mathcal{T}_n}$ 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, using the time-varying nonlinear sieve regression estimator introduced below. The subscript t in $\hat{f}_{t,n,i,a}$ and $\hat{g}_{t,n,j,b}$ is to indicate that we allow for *sequential estimation*, which will be discussed in Remark 4.1. Let $\hat{\boldsymbol{R}}_{t,n} = (\hat{R}_{t,n,m})_{m \in \mathcal{D}_n}$ be the high-dimensional vector process containing the residual products for all dimension/time-offset combinations in $\mathcal{D}_n$. The observed processes X, Y, Z and error processes ϵ, ξ can all be locally stationary time series; see Subsection 4.5 for the details.

For some sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$, we make the following assumption.

Assumption 4.2 (Additive form and regularity)**.** *For each sample size $n \in \mathbb{N}$, distribution $P \in \mathcal{P}_n$, rescaled time $u \in \mathcal{U}_n$, and dimension/time-offset tuple $(i, j, a, b) \in \mathcal{D}_n$, assume that*

$$f_{P,n,i,a}(u, \boldsymbol{z}) = \sum_{k=1}^{d_Z} \sum_{c=1}^{C_k} f_{P,n,i,a,k,c}(u, z_{k,c}),$$
$$g_{P,n,j,b}(u, \boldsymbol{z}) = \sum_{k=1}^{d_Z} \sum_{c=1}^{C_k} g_{P,n,j,b,k,c}(u, z_{k,c}),$$

where $f_{P,n,i,a,k,c} : \mathcal{U}_n \times \mathbb{R} \to \mathbb{R}$ and $g_{P,n,j,b,k,c} : \mathcal{U}_n \times \mathbb{R} \to \mathbb{R}$ are time-varying partial response functions, so that we have

$$\mathbb{E}_P(X_{t,n,i,a} | \boldsymbol{Z}_{t,n} = \boldsymbol{z}) = \sum_{k=1}^{d_Z} \sum_{c=1}^{C_k} f_{P,n,i,a,k,c}(t/n, z_{k,c}),$$
$$\mathbb{E}_P(Y_{t,n,j,b} | \boldsymbol{Z}_{t,n} = \boldsymbol{z}) = \sum_{k=1}^{d_Z} \sum_{c=1}^{C_k} g_{P,n,j,b,k,c}(t/n, z_{k,c}),$$

for each time $t \in \mathcal{T}_n$.

Further, assume for all $n \in \mathbb{N}$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, $k \in [d_Z]$, $c \in C_k$, $u \in \mathcal{U}_n$, there exists some $q \geq 2$ such that

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|f_{P,n,i,a,k,c}(u, Z_{t,n,k,c})|^q) < \infty,$$
$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|g_{P,n,j,b,k,c}(u, Z_{t,n,k,c})|^q) < \infty.$$

To fix ideas, we use the algebraic mapping $h : [-1, 1] \to \mathbb{R}$ from Example 3.1 in Ding and Zhou [DZ21] with positive scaling factor $s = 1$,

$$h(\tilde{z}) = \begin{cases} -\infty, & \tilde{z} = -1, \\ \frac{\tilde{z}}{\sqrt{1-\tilde{z}^2}}, & \tilde{z} \in (-1, 1), \\ \infty, & \tilde{z} = 1. \end{cases}$$

See the discussion preceding Definition 3.1 in Ding and Zhou [DZ21] for additional details. For some sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$, for each $n \in \mathbb{N}$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, $k \in [d_Z]$, $c \in C_k$, and $P \in \mathcal{P}_n$, we relate the time-varying partial response functions $f_{P,n,i,a,k,c} : \mathcal{U}_n \times \mathbb{R} \to \mathbb{R}$ and $g_{P,n,j,b,k,c} : \mathcal{U}_n \times \mathbb{R} \to \mathbb{R}$ to $\tilde{f}_{P,n,i,a,k,c} : [0, 1] \times [0, 1] \to \mathbb{R}$ and $\tilde{g}_{P,n,j,b,k,c} : [0, 1] \times [0, 1] \to \mathbb{R}$, respectively, where

$$\tilde{f}_{P,n,i,a,k,c}(u^*, z^*) = f_{P,n,i,a,k,c}(\mathbb{U}_n^- + u^*(\mathbb{U}_n^+ - \mathbb{U}_n^-), h(2z^* - 1)),$$
$$\tilde{g}_{P,n,j,b,k,c}(u^*, z^*) = g_{P,n,j,b,k,c}(\mathbb{U}_n^- + u^*(\mathbb{U}_n^+ - \mathbb{U}_n^-), h(2z^* - 1)),$$

with $\mathbb{U}_n^- = \min(\mathcal{U}_n)$ and $\mathbb{U}_n^+ = \max(\mathcal{U}_n)$.

For some sequence of collections of distributions $(\mathcal{P}_n)_{n \in \mathbb{N}}$, we make the following assumption.

Assumption 4.3 (Smoothness)**.** *For each $n \in \mathbb{N}$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, $k \in [d_Z]$, $c \in C_k$, and $P \in \mathcal{P}_n$, assume that for each fixed $u^* \in [0, 1]$ we have*

$$\tilde{f}_{P,n,i,a,k,c}(u^*, \cdot) \in C^\infty([0, 1]), \quad \tilde{g}_{P,n,j,b,k,c}(u^*, \cdot) \in C^\infty([0, 1]),$$

and for each fixed $z^* \in [0, 1]$ we have

$$\tilde{f}_{P,n,i,a,k,c}(\cdot, z^*) \in C^\infty([0, 1]), \quad \tilde{g}_{P,n,j,b,k,c}(\cdot, z^*) \in C^\infty([0, 1]),$$

where $C^\infty([0, 1])$ denotes the space of functions on $[0, 1]$ that are infinitely differentiable.

If Assumption 4.3 holds, then by Theorem 3.1 of Ding and Zhou [DZ21] we can approximate the time-varying partial response functions by

$$f_{P,n,i,a,k,c}(u,z) \approx \sum_{\ell_1=1}^{\tilde{c}_n} \sum_{\ell_2=1}^{\tilde{d}_n} \beta^f_{P,n,i,a,k,c,\ell_1,\ell_2} b_{\ell_1,\ell_2}(u,z),$$

$$g_{P,n,j,b,k,c}(u,z) \approx \sum_{\ell_1=1}^{\tilde{c}_n} \sum_{\ell_2=1}^{\tilde{d}_n} \beta^g_{P,n,j,b,k,c,\ell_1,\ell_2} b_{\ell_1,\ell_2}(u,z),$$

where $\{b_{\ell_1,\ell_2}(u,z)\} = \{\phi_{\ell_1}(u)\varphi_{\ell_2}(z)\}$ are basis functions and $\{\beta^f_{P,n,i,a,k,c,\ell_1,\ell_2}\}$, $\{\beta^g_{P,n,j,b,k,c,\ell_1,\ell_2}\}$ are coefficients which we can estimate with OLS. The number of basis functions for time and the covariate values — denoted by $\tilde{c}_n$ and $\tilde{d}_n$, respectively — are chosen to increase with the sample size n at some rate. To fix ideas, we will use Legendre polynomials as the basis functions for both the theoretical analysis in this section and the numerical simulations in Section 5. Specifically, for each $\ell_1 \in [\tilde{c}_n]$ and $\ell_2 \in [\tilde{d}_n]$, let the basis functions for time $\{\phi_{\ell_1}(u)\}$ and the covariate values $\{\varphi_{\ell_2}(z)\}$ be *mapped* Legendre polynomials as in Example C.2 and Subsection 3.1.1 of Ding and Zhou [DZ21]. Next, we introduce the sieve estimators for the time-varying regression functions.

Remark 4.1 (Sequential sieve estimation)**.** *Our formulation of the sieve estimator from Ding and Zhou [DZ21] accommodates sequential estimation, in the sense that the predictors at time t are only constructed using the information up to time t. We emphasize that sequential estimation is not required for all settings, particularly when certain exogeneity conditions hold. The need for sequential estimation in some settings is due to the martingale difference sequence condition imposed on the error processes in Assumption 3.4 (c.f. Assumption 4.4), which becomes relevant when using our CI test for lag selection with nonlinear autoregressive processes, variable selection for forecasting, and dynamic causal inference for time-lagged effects. Crucially, we can still attain the desired convergence rates using sequential estimation, despite the nonstationarity, thanks to the smoothness of the time-varying regression functions from Assumption 4.3 and the "approximate stationarity" of the locally stationary processes over short periods of time from Assumption 4.6.*

Recall the following notation from Subsection 3.2. Let $\mathfrak{D}^{\hat{f}}_{t,n,i,a}$, $\mathfrak{D}^{\hat{g}}_{t,n,j,b}$ be the datasets used to form the "time-t estimators" $\hat{f}_{t,n,i,a}(t/n,\cdot)$, $\hat{g}_{t,n,j,b}(t/n,\cdot)$ of the time-varying regression functions at rescaled time $t/n \in \mathcal{U}_n$, let $\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}$, $\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}$ be the corresponding input sequences, and let $\mathcal{T}^{\hat{f}}_{t,n,i,a}$, $\mathcal{T}^{\hat{g}}_{t,n,j,b}$ be the corresponding sets of times with $T^{\hat{f}}_{t,n,i,a} = |\mathcal{T}^{\hat{f}}_{t,n,i,a}|$, $T^{\hat{g}}_{t,n,j,b} = |\mathcal{T}^{\hat{g}}_{t,n,j,b}|$. Note that each of the time-t estimators $\hat{f}_{t,n,i,a,k,c}(t/n,\cdot)$, $\hat{g}_{t,n,j,b,k,c}(t/n,\cdot)$ of the corresponding time-varying partial response functions at rescaled time $t/n \in \mathcal{U}_n$ may have a different number of basis functions (see Subsection 5.1). However, we write the number of basis functions as $\tilde{c}_n$ and $\tilde{d}_n$ instead of $\tilde{c}^{\hat{f}}_{t,n,i,a,k,c}$, $\tilde{c}^{\hat{g}}_{t,n,j,b,k,c}$ and $\tilde{d}^{\hat{f}}_{t,n,i,a,k,c}$, $\tilde{d}^{\hat{g}}_{t,n,j,b,k,c}$ to simplify the presentation.

For some fixed $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, and $(i,j,a,b) \in \mathcal{D}_n$, denote the design matrices by $\bar{\mathbf{Z}}_{t,n,i,a} \in \mathbb{R}^{T^{\hat{f}}_{t,n,i,a} \times \boldsymbol{d_Z} \tilde{c}_n \tilde{d}_n}$ and $\bar{\mathbf{Z}}_{t,n,j,b} \in \mathbb{R}^{T^{\hat{g}}_{t,n,j,b} \times \boldsymbol{d_Z} \tilde{c}_n \tilde{d}_n}$. The (s,p)-th entries of $\bar{\mathbf{Z}}_{t,n,i,a}$ and $\bar{\mathbf{Z}}_{t,n,j,b}$ are

$$\bar{\mathbf{Z}}^{(s,p)}_{t,n,i,a} = \phi_{\ell_{1,p}}(t_s/n)\varphi_{\ell_{2,p}}(Z_{t_s,n,k_p,c_p}),$$

$$\bar{\mathbf{Z}}^{(s,p)}_{t,n,j,b} = \phi_{\ell_{1,p}}(t_s/n)\varphi_{\ell_{2,p}}(Z_{t_s,n,k_p,c_p}),$$

where we use mappings for the rows $s \mapsto t_s \in \mathcal{T}^{\hat{f}}_{t,n,i,a}$ and $s \mapsto t_s \in \mathcal{T}^{\hat{g}}_{t,n,j,b}$ which maintain the sequential order of time (i.e. $t_{s_1} < t_{s_2}$ if $s_1 < s_2$), and some mappings for the columns $p \mapsto (k_p, c_p, \ell_{1,p}, \ell_{2,p})$ which determine orderings for the dimension/time-offset/basis-index combinations, where $k_p \in [d_Z]$, $c_p \in C_{k_p}$, $\ell_{1,p} \in [\tilde{c}_n]$, $\ell_{2,p} \in [\tilde{d}_n]$. That is, each row corresponds to one time and each column corresponds to one dimension/time-offset combination with a particular basis-index combination. For each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$ the (time-invariant) coefficient vectors

$$\boldsymbol{\beta}^f_{P,n,i,a} = (\beta^f_{P,n,i,a,k,c,\ell_1,\ell_2})^\top_{k,c,\ell_1,\ell_2} \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n},$$

$$\boldsymbol{\beta}^g_{P,n,j,b} = (\beta^g_{P,n,j,b,k,c,\ell_1,\ell_2})^\top_{k,c,\ell_1,\ell_2} \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n},$$

have the following time-t OLS estimators

$$\hat{\boldsymbol{\beta}}^f_{t,n,i,a} = (\bar{\mathbf{Z}}^\top_{t,n,i,a}\bar{\mathbf{Z}}_{t,n,i,a})^{-1}\bar{\mathbf{Z}}^\top_{t,n,i,a}\bar{\mathbf{X}}_{t,n,i,a} = (\hat{\beta}^f_{t,n,i,a,k,c,\ell_1,\ell_2})^\top_{k,c,\ell_1,\ell_2} \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n},$$
$$\hat{\boldsymbol{\beta}}^g_{t,n,j,b} = (\bar{\mathbf{Z}}^\top_{t,n,j,b}\bar{\mathbf{Z}}_{t,n,j,b})^{-1}\bar{\mathbf{Z}}^\top_{t,n,j,b}\bar{\mathbf{Y}}_{t,n,j,b} = (\hat{\beta}^g_{t,n,j,b,k,c,\ell_1,\ell_2})^\top_{k,c,\ell_1,\ell_2} \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n},$$

where

$$\bar{\mathbf{X}}_{t,n,i,a} = (X_{t,n,i,a})^\top_{t\in\mathcal{T}^{\hat{f}}_{t,n,i,a}} \in \mathbb{R}^{T^{\hat{f}}_{t,n,i,a}}, \quad \bar{\mathbf{Y}}_{t,n,j,b} = (Y_{t,n,j,b})^\top_{t\in\mathcal{T}^{\hat{g}}_{t,n,j,b}} \in \mathbb{R}^{T^{\hat{g}}_{t,n,j,b}}.$$

Finally, the time-t estimators of the time-varying regression functions $f_{P,n,i,a}(t/n,\cdot)$ and $g_{P,n,j,b}(t/n,\cdot)$ at rescaled time $t/n \in \mathcal{U}_n$ are given by

$$\hat{f}_{t,n,i,a}(t/n,\cdot) = \sum_{k=1}^{d_Z}\sum_{c=1}^{C_k}\hat{f}_{t,n,i,a,k,c}(t/n,\cdot),$$
$$\hat{g}_{t,n,j,b}(t/n,\cdot) = \sum_{k=1}^{d_Z}\sum_{c=1}^{C_k}\hat{g}_{t,n,j,b,k,c}(t/n,\cdot),$$

where the time-t estimators of the time-varying partial response functions $f_{P,n,i,a,k,c}(t/n,\cdot)$ and $g_{P,n,j,b,k,c}(t/n,\cdot)$ at rescaled time $t/n \in \mathcal{U}_n$ are given by

$$\hat{f}_{t,n,i,a,k,c}(t/n,\cdot) = \sum_{\ell_1=1}^{\tilde{c}_n}\sum_{\ell_2=1}^{\tilde{d}_n}\hat{\beta}^f_{t,n,i,a,k,c,\ell_1,\ell_2}b_{\ell_1,\ell_2}(t/n,\cdot),$$
$$\hat{g}_{t,n,j,b,k,c}(t/n,\cdot) = \sum_{\ell_1=1}^{\tilde{c}_n}\sum_{\ell_2=1}^{\tilde{d}_n}\hat{\beta}^g_{t,n,j,b,k,c,\ell_1,\ell_2}b_{\ell_1,\ell_2}(t/n,\cdot).$$

Although we only discuss the sieve estimator here, note that time-varying regression estimators based on kernel smoothing can be used as well [Vog12; ZW15; YN21; CSW22; DZ21]. In fact, one-sided kernels for time can be employed for sequential estimation when using kernel smoothing estimators, so that only data before time t is used to form the time-t predictor. Also, we point interested readers to results on asymptotically optimal *linear* forecasting for locally stationary time series [DZ23; KR24; CZ25]. We expect analogous results for asymptotically optimal *nonlinear* forecasting to be developed over the next few years.

4.4 Locally stationary error processes

We will now introduce the causal representations of the locally stationary error processes from Subsection 4.3. For each $a \in A$, $b \in B$, define the input sequences

$$\mathcal{H}^\varepsilon_{t,a} = (\eta^\varepsilon_{t,a}, \eta^\varepsilon_{t,a-1}, \ldots), \ \mathcal{H}^\xi_{t,b} = (\eta^\xi_{t,b}, \eta^\xi_{t,b-1}, \ldots),$$

where $(\eta^\varepsilon_{t,a}, \eta^\xi_{t,b})_{t\in\mathbb{Z}}$ is a sequence of iid random vectors. Denote the dimension of $\eta^\varepsilon_{t,a} = \eta^\varepsilon_{t,a,n}$ by $d^\eta_\varepsilon = d^\eta_{\varepsilon,n}$ and the dimension of $\eta^\xi_{t,b} = \eta^\xi_{t,b,n}$ by $d^\eta_\xi = d^\eta_{\xi,n}$, both of which can change with n. For the next assumption, let $\mathcal{H}^{\hat{\boldsymbol{f}}}_t = (\mathcal{H}^{\hat{f}}_{t,a})_{a\in A}$, $\mathcal{H}^{\hat{\boldsymbol{g}}}_t = (\mathcal{H}^{\hat{g}}_{t,b})_{b\in B}$ and $\mathcal{H}^{\hat{f}}_{t,a} = (\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}, \mathcal{H}^{\boldsymbol{Z}}_t)$, $\mathcal{H}^{\hat{g}}_{t,b} = (\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}, \mathcal{H}^{\boldsymbol{Z}}_t)$, where the input sequences $\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a}$, $\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b}$ were defined in Subsection 4.3 and $\mathcal{H}^{\boldsymbol{Z}}_t$ was defined in Subsection 4.2.

Assumption 4.4 (Causal representations of the error processes)**.** *Assume that for each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $(i,j,a,b) \in \mathcal{D}_n$, $t \in \mathcal{T}_n$, the error processes from Subsection 4.3 can be represented as*

$$\varepsilon_{P,t,n,i,a} = \tilde{G}^\varepsilon_{P,n,i,a}(t/n, \mathcal{H}^\varepsilon_{t,a}), \ \xi_{P,t,n,j,b} = \tilde{G}^\xi_{P,n,j,b}(t/n, \mathcal{H}^\xi_{t,b}),$$

with $\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{g}}}_t) = 0$ and $\mathbb{E}_P(\xi_{P,t,n,j,b}|\mathcal{H}^{\hat{\boldsymbol{f}}}_t) = 0$, where the input sequences $\mathcal{H}^{\hat{\boldsymbol{g}}}_t$, $\mathcal{H}^{\hat{\boldsymbol{f}}}_t$ were defined above. The causal representations

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

are defined at all $u \in \mathcal{U}_n$*, so that we have* $\varepsilon_{P,t,n,i,a} = \tilde{\varepsilon}_{P,t,n,i,a}(t/n)$*,* $\xi_{P,t,n,j,b} = \tilde{\xi}_{P,t,n,j,b}(t/n)$*.* $\tilde{G}^{\varepsilon}_{P,n,i,a}(u,\cdot)$ *and* $\tilde{G}^{\xi}_{P,n,j,b}(u,\cdot)$ *are measurable functions from* $(\mathbb{R}^{d^{\eta}_{\varepsilon}})^{\infty}$ *and* $(\mathbb{R}^{d^{\eta}_{\xi}})^{\infty}$*, respectively, to* $\mathbb{R}$ *— where we endow* $(\mathbb{R}^{d^{\eta}_{\varepsilon}})^{\infty}$ *and* $(\mathbb{R}^{d^{\eta}_{\xi}})^{\infty}$ *with the* σ*-algebra generated by all finite projections — so 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.*

As in Subsection 3.3, we have not defined the input sequences for the error processes separately for each dimension, because without loss of generality we may define the measurable functions $\tilde{G}^{\varepsilon}_{P,n,i,a}(u,\cdot)$, $\tilde{G}^{\xi}_{P,n,j,b}(u,\cdot)$ and inputs $\eta^{\varepsilon}_{t,a}, \eta^{\xi}_{t,b}$ so that each dimension of the error processes has idiosyncratic inputs.

Using the causal representations of the univariate error processes, we have the following causal representations of the vector-valued error processes

$$\tilde{\boldsymbol{\varepsilon}}_{P,t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{\varepsilon}}_{P,n}(u,\mathcal{H}^{\boldsymbol{\varepsilon}}_t) = (\tilde{G}^{\varepsilon}_{P,n,i,a}(u,\mathcal{H}^{\varepsilon}_{t,a}))_{i\in[d_X],a\in A_i},$$
$$\tilde{\boldsymbol{\xi}}_{P,t,n}(u) = \tilde{\boldsymbol{G}}^{\boldsymbol{\xi}}_{P,n}(u,\mathcal{H}^{\boldsymbol{\xi}}_t) = (\tilde{G}^{\xi}_{P,n,j,b}(u,\mathcal{H}^{\xi}_{t,b}))_{j\in[d_Y],b\in B_j},$$

so that we have $\boldsymbol{\varepsilon}_{P,t,n} = \tilde{\boldsymbol{\varepsilon}}_{P,t,n}(t/n)$, $\boldsymbol{\xi}_{P,t,n} = \tilde{\boldsymbol{\xi}}_{P,t,n}(t/n)$, where $\mathcal{H}^{\boldsymbol{\varepsilon}}_t = (\eta^{\boldsymbol{\varepsilon}}_t, \eta^{\boldsymbol{\varepsilon}}_{t-1}, \ldots)$, $\mathcal{H}^{\boldsymbol{\xi}}_t = (\eta^{\boldsymbol{\xi}}_t, \eta^{\boldsymbol{\xi}}_{t-1}, \ldots)$ with $\eta^{\boldsymbol{\varepsilon}}_t = (\eta^{\varepsilon}_{t,a})_{a\in A}$, $\eta^{\boldsymbol{\xi}}_t = (\eta^{\xi}_{t,b})_{b\in B}$ for each $t \in \mathbb{Z}$. Similarly, for each $m = (i,j,a,b) \in \mathcal{D}_n$ the error products can be represented as

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

so that we have $R_{P,t,n,m} = \tilde{R}_{P,t,n,m}(t/n)$, where $\mathcal{H}^{R}_{t,m} = (\eta^{R}_{t,m}, \eta^{R}_{t-1,m}, \ldots)$ with $\eta^{R}_{t,m} = (\eta^{\varepsilon}_{t,a}, \eta^{\varepsilon}_{t,b})^{\top}$ for each $t \in \mathbb{Z}$. Also, we have the following causal representation of the nonstationary $\mathbb{R}^{D_n}$-valued process of all the products of errors

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

so that we have $\boldsymbol{R}_{P,t,n} = \tilde{\boldsymbol{R}}_{P,n,t}(t/n)$, where $\mathcal{H}^{\boldsymbol{R}}_t = (\eta^{\boldsymbol{R}}_t, \eta^{\boldsymbol{R}}_{t-1}, \ldots)$ and $\eta^{\boldsymbol{R}}_t = (\eta^{\boldsymbol{\varepsilon}}_t, \eta^{\boldsymbol{\xi}}_t)^{\top}$ for each $t \in \mathbb{Z}$. We emphasize that 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 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 stationary ergodic $\mathbb{R}^{D_n}$-valued time series.

4.5 Assumptions on dependence and nonstationarity

In this subsection, we impose assumptions on the rate of decay in temporal dependence and the degree of nonstationarity of the observed processes and error processes. We emphasize that the assumptions here are strictly stronger than those in Subsection 3.4. We impose these stronger assumptions to guarantee that the sieve time-varying nonlinear regression estimator achieves the convergence rates required by Theorem 3.1. Note that the assumptions here require that the nonstationary processes evolve "smoothly" in time, which excludes nonstationary processes with abrupt changes. We do this mainly to simplify the presentation, and we discuss extensions to nonstationary processes with *both* smooth and abrupt changes in Subsection A.5.

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

$$\mathbb{W} = \{(X,i,a) : i \in [d_X], a \in A_i\} \cup \{(Y,j,b) : j \in [d_Y], b \in B_j\} \cup \{(Z,k,c) : k \in [d_Z], c \in C_k\},$$

so that we may conveniently refer to such well-defined combinations by $(W,l,d) \in \mathbb{W}$. Also, 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 via the functional dependence measure of Wu [Wu05]. Let $(\tilde{\eta}^X_t, \tilde{\eta}^Y_t, \tilde{\eta}^Z_t)_{t\in\mathbb{Z}}$ be an iid copy of $(\eta^X_t, \eta^Y_t, \eta^Z_t)_{t\in\mathbb{Z}}$. Going forward, the inputs with the tilde are from $(\tilde{\eta}^X_t, \tilde{\eta}^Y_t, \tilde{\eta}^Z_t)_{t\in\mathbb{Z}}$. For any tuple $(W,l,d) \in \mathbb{W}$ corresponding to a well-defined combination of an observed process, dimension, and time-offset, define

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

to be $\mathcal{H}^W_{t,d}$ with the input η^W_{t-h+d} replaced with the iid copy $\tilde{\eta}^W_{t-h+d}$. For example, for $i \in [d_X]$, $a \in A_i$, we have that $\tilde{\eta}^X_{t-h+a}$ is the copy of the input η^X_{t-h+a} in the input sequence $\mathcal{H}^X_{t,a}$ used in the causal representation of $X_{t,n,i,a}$. Analogously, for $\boldsymbol{W} \in \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ define $\tilde{\mathcal{H}}^{\boldsymbol{W}}_{t,h}$ as $\mathcal{H}^{\boldsymbol{W}}_t$ with the input $\eta^{\boldsymbol{W}}_{t-h}$ replaced with the iid copy $\tilde{\eta}^{\boldsymbol{W}}_{t-h}$ as in Subsection 4.2.

For any tuple $(e, l, d) \in \mathbb{E}$ corresponding to a well-defined combination of an error process, dimension, and time-offset, define

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

to be $\mathcal{H}^e_{t,d}$ with the input $\eta^e_{t-h,d}$ replaced with the iid copy $\tilde{\eta}^e_{t-h,d}$. Analogously, for $\boldsymbol{e} \in \{\boldsymbol{\varepsilon}, \boldsymbol{\xi}\}$ define $\tilde{\mathcal{H}}^{\boldsymbol{e}}_{t,h}$ as $\mathcal{H}^{\boldsymbol{e}}_t$ with the input $\eta^{\boldsymbol{e}}_{t-h}$ replaced with the iid copy $\tilde{\eta}^{\boldsymbol{e}}_{t-h}$ as in Subsection 4.4. Also, for the product of errors define $\tilde{\mathcal{H}}^R_{t,m,h}$ as $\mathcal{H}^R_{t,m}$ with the input $\eta^R_{t-h,m}$ replaced with the iid copy $\tilde{\eta}^R_{t-h,m}$ for $m = (i, j, a, b) \in \mathcal{D}_n$. Analogously, define $\tilde{\mathcal{H}}^{\boldsymbol{R}}_{t,h}$ as $\mathcal{H}^{\boldsymbol{R}}_t$ with the input $\eta^{\boldsymbol{R}}_{t-h}$ replaced with the iid copy $\tilde{\eta}^{\boldsymbol{R}}_{t-h}$ as in Subsection 4.4. Now, we define the functional dependence measures of the processes.

Definition 4.1 (Functional dependence measures)**.** *We define the following measures of temporal dependence for each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, and $t \in \mathcal{T}_n$. First, define the functional dependence measures of the observed processes $\tilde{G}^W_{n,l,d}(u, \mathcal{H}^W_{t,d})$ for each $(W, l, d) \in \mathbb{W}$ with $h \in \mathbb{N}_0$, $q > 2$ as*

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

and for the vector-valued process $\tilde{\boldsymbol{G}}^{\boldsymbol{W}}_n(u, \mathcal{H}^{\boldsymbol{W}}_t)$ for each $\boldsymbol{W} \in \{\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}\}$ with $h \in \mathbb{N}_0$, $q > 2$, $r \geq 1$ as

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

Second, define the L^∞ versions of the functional dependence measures of the error processes $\tilde{G}^e_{P,n,l,d}(u, \mathcal{H}^e_{t,d})$ for each $(e, l, d) \in \mathbb{E}$ with $h \in \mathbb{N}_0$ as

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

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

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

Third, define the functional dependence measures of the processes of error products $\tilde{G}^R_{P,n,m}(u, \mathcal{H}^R_{t,m})$ for each $m = (i, j, a, b) \in \mathcal{D}_n$ with $h \in \mathbb{N}_0$, $q > 4$ as

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

and for the vector-valued process $\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_t)$ with $h \in \mathbb{N}_0$, $q > 4$, $r \geq 2$ as

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

Next, 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. For some sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we make the following assumption.

Assumption 4.5 (Distribution-uniform decay of temporal dependence)**.** *Assume that there exist $\bar{\Theta} > 0$, $\bar{\beta} > 2$, $\bar{q} > 2$, such that for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, for the observed processes $(W, l, d) \in \mathbb{W}$, it holds that*

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

and for the error processes $(e, l, d) \in \mathbb{E}$, it holds that

$$\sup_{P\in\mathcal{P}_n} [\mathbb{E}_P(|\tilde{G}^e_{P,n,l,d}(u, \mathcal{H}^e_{0,d})|^{\bar{q}})]^{1/\bar{q}} \leq \bar{\Theta}.$$

Assume that there exist $\bar{\Theta}^{\infty} > 0$, $\bar{\beta}^{\infty} > 2$ *such that for all* $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, *error processes* $(e, l, d) \in \mathbb{E}$, *it holds that*

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

For additional control in terms of the product of errors alone, also assume that there exist $\bar{\Theta}^R > 0$, $\bar{\beta}^R > 3$, $\bar{q}^R > 4$, *such that for all* $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $m = (i, j, a, b) \in \mathcal{D}_n$, *it holds that*

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

In view of the previous assumption, we have the following bounds on the functional dependence measures of the corresponding vector-valued processes for each $n \in \mathbb{N}$, $u \in \mathcal{U}_n$ by Jensen's inequality. Recall the constants from Assumption 4.5. For the process of error products, we have

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_0)||_2^{\bar{q}^R})]^{1/\bar{q}^R} \leq D_n^{\frac{1}{2}} \bar{\Theta}^R, \quad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{R}}_{P,u,0,n}(h, \bar{q}^R, 2) \leq D_n^{\frac{1}{2}} \bar{\Theta}^R \cdot (h \vee 1)^{-\bar{\beta}^R}, \quad h \geq 0.$$

Also, for each observed process $\boldsymbol{W} \in (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, we have

$$\sup_{P \in \mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{W}}_n(u, \mathcal{H}^{\boldsymbol{W}}_0)||_2^{\bar{q}})]^{1/\bar{q}} \leq D_n^{\frac{1}{2}} \bar{\Theta}, \quad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{W}}_{P,u,0,n}(h, \bar{q}, 2) \leq D_n^{\frac{1}{2}} \bar{\Theta} \cdot (h \vee 1)^{-\bar{\beta}}, \quad h \geq 0,$$

Lastly, for each error process $\boldsymbol{e} \in (\boldsymbol{\varepsilon}, \boldsymbol{\xi})$, we have

$$\sup_{P \in \mathcal{P}_n} \left|\left|||\tilde{\boldsymbol{G}}^{\boldsymbol{e}}_{P,n}(u, \mathcal{H}^{\boldsymbol{e}}_0)||_2\right|\right|_{L^{\infty}(P)} \leq D_n^{\frac{1}{2}} \bar{\Theta}^{\infty}, \quad \sup_{P \in \mathcal{P}_n} \theta^{\boldsymbol{e},\infty}_{P,u,0,n}(h, 2) \leq D_n^{\frac{1}{2}} \bar{\Theta}^{\infty} \cdot (h \vee 1)^{-\bar{\beta}^{\infty}}, \quad h \geq 0.$$

Next, we discuss an additional regularity condition required by the sieve estimator that is analogous to Lemma 3.1 in Ding and Zhou [DZ21]. Recall the set of basis function $\{\varphi_{\ell_2}(z)\}$ from Subsection 4.3. For each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$ let

$$\begin{aligned}
\boldsymbol{w}^{\varphi(Z)}_{t,n} &= (\varphi_{\ell_2}(Z_{t,n,k,c}))_{k \in [d_Z], c \in C_k, 1 \leq \ell_2 \leq \tilde{d}_n}, \\
\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a} &= (\varphi_{\ell_2}(Z_{t,n,k,c}) \varepsilon_{P,t,n,i,a})_{k \in [d_Z], c \in C_k, 1 \leq \ell_2 \leq \tilde{d}_n}, \\
\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b} &= (\varphi_{\ell_2}(Z_{t,n,k,c}) \xi_{P,t,n,j,b})_{k \in [d_Z], c \in C_k, 1 \leq \ell_2 \leq \tilde{d}_n}.
\end{aligned}$$

As in Subsection 3.2 in Ding and Zhou [DZ21], the $\mathbb{R}^{\boldsymbol{d_Z} \tilde{d}_n}$-valued processes $\boldsymbol{w}^{\varphi(Z)}_{t,n}$, $\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a}$, and $\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b}$ all have causal representations

$$\begin{aligned}
\boldsymbol{w}^{\varphi(Z)}_{t,n} &= \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z)}}_n(t/n, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z)}}_t), \\
\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a} &= \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(t/n, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a}), \\
\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b} &= \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(t/n, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b}),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{H}^{\boldsymbol{w}^{\varphi(Z)}}_t &= (\eta^{\boldsymbol{w}^{\varphi(Z)}}_t, \eta^{\boldsymbol{w}^{\varphi(Z)}}_{t-1}, \ldots), \\
\mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a} &= (\eta^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a}, \eta^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t-1,a}, \ldots), \\
\mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b} &= (\eta^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b}, \eta^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t-1,b}, \ldots),
\end{aligned}$$

with $\eta^{\boldsymbol{w}^{\varphi(Z)}}_t = \eta^Z_{t+c_{\max}}$, $\eta^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a} = (\eta^Z_{t+c_{\max}}, \eta^X_{t+a})^{\top}$, and $\eta^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b} = (\eta^Z_{t+c_{\max}}, \eta^Y_{t+b})^{\top}$.

Define the functional dependence measures of the vector-valued processes $\boldsymbol{w}^{\varphi(Z)}_{t,n}$, $\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a}$, $\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b}$, by

$$\begin{aligned}
\theta^{\varphi(Z)}_{P,u,t,n}(h, q, 2) &= [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\varphi(Z)}_n(u, \mathcal{H}^{\varphi(Z)}_t) - \tilde{\boldsymbol{G}}^{\varphi(Z)}_n(u, \tilde{\mathcal{H}}^{\varphi(Z)}_{t,h})||_2^q)]^{1/q}, \\
\theta^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,u,t,n,i,a}(h, q, 2) &= [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a}) - \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \tilde{\mathcal{H}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a,h})||_2^q)]^{1/q}, \\
\theta^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,u,t,n,j,b}(h, q, 2) &= [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b}) - \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \tilde{\mathcal{H}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b,h})||_2^q)]^{1/q}.
\end{aligned}$$

Recall $\bar{\Theta} > 0$, $\bar{\Theta}^{\infty} > 0$, $\bar{\beta} > 2$, $\bar{\beta}^{\infty} > 2$, and $\bar{q} > 2$ from Assumption 4.5. Using the same arguments from Lemma 3.1 from Ding and Zhou [DZ21], for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, the vector-valued processes $\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a}$, $\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b}$ satisfy

$$\sup_{P\in\mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{t,a})||_2^{\tilde{q}})]^{1/\tilde{q}} \leq D_n^{\frac{1}{2}}\tilde{\Theta}, \quad \sup_{P\in\mathcal{P}_n} \theta^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,u,t,n,i,a}(h,\tilde{q},2) \leq D_n^{\frac{1}{2}}\tilde{\Theta}\cdot(h\vee 1)^{-\tilde{\beta}},$$
$$\sup_{P\in\mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{t,b})||_2^{\tilde{q}})]^{1/\tilde{q}} \leq D_n^{\frac{1}{2}}\tilde{\Theta}, \quad \sup_{P\in\mathcal{P}_n} \theta^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,u,t,n,j,b}(h,\tilde{q},2) \leq D_n^{\frac{1}{2}}\tilde{\Theta}\cdot(h\vee 1)^{-\tilde{\beta}},$$

for $h \geq 0$, where some $\tilde{q} = \bar{q} > 2$ with $\tilde{\beta} = \min(\bar{\beta}, \bar{\beta}^{\infty}) > 2$ and $\tilde{\Theta} = 2K_1(\max(\bar{\Theta}, \bar{\Theta}^{\infty}))^2 > 0$ where the constant factor $K_1 > 0$ is due to the basis functions. Similarly, for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, the vector-valued process $\boldsymbol{w}^{\varphi(Z)}_{t,n}$ satisfies

$$\sup_{P\in\mathcal{P}_n} [\mathbb{E}_P(||\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z)}}_{n}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z)}}_{t})||_2^{\bar{q}})]^{1/\bar{q}} \leq D_n^{\frac{1}{2}}\bar{\Theta}K_2, \quad \sup_{P\in\mathcal{P}_n} \theta^{\varphi(Z)}_{P,u,t,n}(h,\bar{q},2) \leq D_n^{\frac{1}{2}}\bar{\Theta}K_2\cdot(h\vee 1)^{-\bar{\beta}},$$

for $h \geq 0$, where the constant factor $K_2 > 0$ is due to the basis functions.

We impose the following regularity conditions to control the nonstationarity uniformly over a sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$. Note that for Theorem 3.1, we only require that the total variation of the causal mechanism of the process of error products can be bounded distribution-uniformly. However, the sieve regression estimator requires the stronger assumption that the causal mechanisms of the observed processes and error processes are all stochastic Lipschitz functions of rescaled time. It is easy to verify that Assumption 3.6 is satisfied under the following stronger assumption on the nonstationarity.

Assumption 4.6 (Distribution-uniform stochastic Lipschitz condition)**.** *For each $n \in \mathbb{N}$, $(W, l, d) \in \mathbb{W}$, $(e, l, d) \in \mathbb{E}$, and $t \in \mathbb{Z}$, we assume that $\tilde{G}^W_{n,l,d}(\cdot, \mathcal{H}^W_{t,d})$ and $\tilde{G}^e_{P,n,l,d}(\cdot, \mathcal{H}^e_{t,d})$ are stochastic Lipschitz functions of rescaled time $u \in \mathcal{U}_n$. In particular, assume that there exist $\bar{\Theta} > 0$, $\bar{q} > 4$ and $\bar{\Theta}^* > 0$, $\bar{q}^* > 4$, such that for all $n \in \mathbb{N}$, $u, v \in \mathcal{U}_n$, $(W, l, d) \in \mathbb{W}$, $(e, l, d) \in \mathbb{E}$, for some $\bar{L} > 0$, $\bar{L}^* > 0$ it holds that*

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

and

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

In view of the previous assumption, there exist $\bar{\Theta}^R > 0$, $\bar{L}^R > 0$ such that for all $n \in \mathbb{N}$, $u, v \in \mathcal{U}_n$, $m = (i, j, a, b) \in \mathcal{D}_n$ we have for $\bar{q}^R = \bar{q}^* > 4$ that

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

which follows from adding and subtracting cross-terms, the triangle inequality, the distributive property, Hölder's inequality, and applying the stochastic Lipschitz condition for the individual error processes from Assumption 4.6. Similarly, using the same arguments as Lemma 3.1 from Ding and Zhou [DZ21], the individual dimensions of the vector-valued processes $\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a}$, and $\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b}$ can be shown to satisfy this stochastic Lipschitz condition with $\tilde{q} = \min(\bar{q}, \bar{q}^*)/2 > 2$, $\tilde{L} = \max(\bar{L}, \bar{L}^*) > 0$, and $\tilde{\Theta} = 2K_1(\max(\bar{\Theta}, \bar{\Theta}^{\infty}))^2 > 0$, where the constant factor $K_1 > 0$ is due to the basis functions and $\bar{\Theta} > 0$, $\bar{q} > 4$, $\bar{L} > 0$ and $\bar{\Theta}^* > 0$, $\bar{q}^* > 4$, $\bar{L}^* > 0$ are from Assumption 4.6. Also, the individual dimensions of the vector-valued process $\boldsymbol{w}^{\varphi(Z)}_{t,n}$ can be shown to satisfy this stochastic Lipschitz condition with $\tilde{q} = \bar{q} > 4$, $\tilde{L} = \bar{L} > 0$, and $\tilde{\Theta} = K_2\bar{\Theta} > 0$, where the constant factor $K_2 > 0$ is due to the basis functions.

4.6 Assumptions on local long-run covariances

To ensure fast convergence rates by the sieve estimator, we require the following assumptions on the *local long-run* covariance matrices. Note that these assumptions are not made in Section 3.

Definition 4.2 (Local long-run covariance matrices of error products)**.** *For each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, define the local long-run covariance matrix $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{P,n}(u) \in \mathbb{R}^{D_n \times D_n}$ for the $\mathbb{R}^{D_n}$-valued stationary process $(\tilde{\boldsymbol{G}}^{\boldsymbol{R}}_{P,n}(u, \mathcal{H}^{\boldsymbol{R}}_t))_{t\in\mathbb{Z}}$ by*

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

By Lemma B.5, the local long-run covariance matrices of $\boldsymbol{w}^{\varphi(Z)}_{t,n}$, $\boldsymbol{w}^{\varphi(Z),\varepsilon}_{P,t,n,i,a}$, $\boldsymbol{w}^{\varphi(Z),\xi}_{P,t,n,j,b}$ are well-defined in view of the discussion following Assumption 4.5. Now, we will define the local long-run and *integrated* long-run covariance matrices of these processes as in Subsection 3.2 of Ding and Zhou [DZ21].

Definition 4.3 (Local long-run and integrated long-run covariance matrices)**.** *For each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, define the local long-run covariance matrices $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}(u)$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u)$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u) \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{d}_n \times \boldsymbol{d_Z}\tilde{d}_n}$ for the $\mathbb{R}^{\boldsymbol{d_Z}\tilde{d}_n}$-valued stationary processes $(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z)}}_n(u, \mathcal{H}^{\boldsymbol{R}}_t))_{t\in\mathbb{Z}}$, $(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \mathcal{H}^{\boldsymbol{R}}_t))_{t\in\mathbb{Z}}$, $(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \mathcal{H}^{\boldsymbol{R}}_t))_{t\in\mathbb{Z}}$, respectively, by*

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}(u) &= \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z)}}_n(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z)}}_0), \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z)}}_n(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z)}}_h)),\\
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u) &= \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{0,a}), \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{h,a})),\\
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u) &= \sum_{h\in\mathbb{Z}} \mathrm{Cov}_P(\tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{0,b}), \tilde{\boldsymbol{G}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u, \mathcal{H}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{h,b})).
\end{aligned}$$

Next, for each $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $u \in \mathcal{U}_n$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, define the corresponding integrated long-run covariance matrices $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b} \in \mathbb{R}^{\boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n \times \boldsymbol{d_Z}\tilde{c}_n\tilde{d}_n}$ by

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n} &= \int_{\mathcal{U}_n} \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}(u) \otimes (\boldsymbol{\phi}(u)\boldsymbol{\phi}^\top(u))du,\\
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a} &= \int_{\mathcal{U}_n} \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}(u) \otimes (\boldsymbol{\phi}(u)\boldsymbol{\phi}^\top(u))du,\\
\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b} &= \int_{\mathcal{U}_n} \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}(u) \otimes (\boldsymbol{\phi}(u)\boldsymbol{\phi}^\top(u))du,
\end{aligned}$$

where $\boldsymbol{\phi}(u) = (\phi_1(u), \ldots, \phi_{\tilde{c}_n}(u))^\top$.

We require the following regularity assumption due to the sieve estimator, which is analogous to Assumption 3.2 from Ding and Zhou [DZ21]. Specifically, for some sequence of collections of distributions $(\mathcal{P}_n)_{n\in\mathbb{N}}$, we impose a distribution-uniform lower bound on the eigenvalues of the integrated long-run covariance matrices.

Assumption 4.7 (Eigenvalue condition for integrated long-run covariance matrices)**.** *Recall $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}$, $\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b}$ from Definition 4.3. Assume that there exists a universal constant $\kappa > 0$ such that for all $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $i \in [d_X]$, $a \in A_i$, $j \in [d_Y]$, $b \in B_j$, we have*

$$\inf_{P\in\mathcal{P}_n} \min(\lambda_{\min}(\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z)}}_{P,n}), \lambda_{\min}(\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\varepsilon}}_{P,n,i,a}), \lambda_{\min}(\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{w}^{\varphi(Z),\xi}}_{P,n,j,b})) \geq \kappa,$$

where $\lambda_{\min}(\cdot)$ is the smallest eigenvalue of the given matrix.

Again, we emphasize that the locally stationary time series framework in this section fits into the more general triangular array framework from Section 3. Hence, we can use the same cumulative covariance estimator $\hat{Q}^{\boldsymbol{R}}_{t,n}$ from Subsection 3.5 for the cumulative covariance matrices $Q^{\boldsymbol{R}}_{P,t,n} = \sum_{s=\mathbb{T}^-_n}^{t} \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,s,n}$, where $\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,s,n} = \tilde{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{P,n}(s/n)$ denotes the local long-run covariance matrix at time $s \in \mathcal{T}_n$.

4.7 Theoretical result for Sieve-dGCM

The main result of this section is that (1) the previously stated assumptions imply the assumptions of Theorem 3.1, and (2) under these strictly stronger assumptions the sieve time-varying regression estimator will achieve the convergence rates required by Theorem 3.1. Hence, the Sieve-dGCM test (i.e. running Algorithm 1 with the predictions from the sieve time-varying regression estimator) will have uniformly asymptotic Type-I error control.

Theorem 4.1. *Suppose that Assumptions 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 all hold for the sequence of collections of distributions* $(\mathcal{P}^*_{0,n})_{n\in\mathbb{N}}$*, where* $\mathcal{P}^*_{0,n} \subset \mathcal{P}^{\text{CI}}_{0,n}$ *for each* $n \in \mathbb{N}$*. Further, suppose that we use the sieve time-varying regression estimator with the basis functions* $\{\phi_{\ell_1}(u)\}$*,* $\{\varphi_{\ell_2}(z)\}$ *chosen to be mapped Legendre polynomials, and that the number of basis functions are chosen to satisfy* $\tilde{c}_n = O(\log(n))$*,* $\tilde{d}_n = O(\log(n))$*. Then the assumptions of Theorem 3.1 hold for* $(\mathcal{P}^*_{0,n})_{n\in\mathbb{N}}$ *and the sieve estimator will satisfy the required convergence rates.*

5 Numerical Simulations

In this section, we conduct simulations with the Sieve-dGCM test from Section 4, which consists of running Algorithm 1 based on the predictions from the sieve time-varying nonlinear regression estimator. In Subsection 5.1, we discuss how to select the parameters of the sieve estimator and the lag-window parameter of the cumulative covariance estimator. In Subsection 5.2, we show the simulation results. In Subsection 5.3, we conclude by discussing the simulation results.

All rejection rates are based on running Algorithm 1 with the sieve estimator on 100 simulated time series. Due to computational constraints, we focus our simulations on the univariate dGCM test, as our primary applications of interest are variable selection and causal discovery. We consider Legendre polynomials as the basis functions for the sieve estimator, as in our theoretical analysis in Section 4. The number of basis functions for time and the covariate values were chosen using the subsampling cross-validation procedure from Subsection 5.1. The lag-window parameter for covariance estimation was selected according to the minimum volatility method from Subsection 5.1. We use $s = 5000$ Monte Carlo simulations to approximate the desired quantile.

5.1 Parameter selection via subsampling and minimum volatility

To begin, we introduce a novel cross-validation approach which can be used for selecting the parameters of the sieve estimator in the time-varying regression setting. Our approach complements the cross-validation procedure suggested in Subsection 5.1 of Ding and Zhou [DZ21], which only deals with parameter selection for the sieve estimator in the autoregressive forecasting setting. Note that in settings where we forecast with covariates, we can use now-standard approaches for time series cross-validation; see Subsection 5.10 of Hyndman [Hyn18]. Also, we note that Dahlhaus and Richter [DR19; DR23] theoretically investigated cross-validation for locally stationary time series in the context of selecting bandwidths for kernel smoothing estimators (i.e. a "local" estimation approach). In contrast, our cross-validation approach is for "global" estimators, such as the sieve estimator.

The main idea of our cross-validation scheme is to create several folds constructed by sampling the original time series at a *lower sampling frequency*. Specifically, for some buffer $\gamma \in \mathbb{N}_0$ and index $k = 1, \ldots, 2(\gamma+1)$, the k-th fold will consist of the subsampled time series

$$\mathcal{T}_n^{(k)} = \{\mathbb{T}_n^- + k - 1 + 2j(\gamma+1) : j = 0, 1, \ldots, \lfloor \frac{\mathbb{T}_n^+ - \mathbb{T}_n^- - k + 1}{2(\gamma+1)} \rfloor\}.$$

For instance, when the buffer $\gamma = 0$, we have $\mathcal{T}_n^{(1)} = \{\mathbb{T}_n^-, \mathbb{T}_n^- + 2, \ldots\}$ and $\mathcal{T}_n^{(2)} = \{\mathbb{T}_n^- + 1, \mathbb{T}_n^- + 3, \ldots\}$. Similarly, when the buffer $\gamma = 1$, we have $\mathcal{T}_n^{(1)} = \{\mathbb{T}_n^-, \mathbb{T}_n^- + 4, \ldots\}$, $\mathcal{T}_n^{(2)} = \{\mathbb{T}_n^- + 1, \mathbb{T}_n^- + 5, \ldots\}$, $\mathcal{T}_n^{(3)} = \{\mathbb{T}_n^- + 2, \mathbb{T}_n^- + 6, \ldots\}$, and $\mathcal{T}_n^{(4)} = \{\mathbb{T}_n^- + 3, \mathbb{T}_n^- + 7, \ldots\}$. The reason we refer to γ as a buffer will be made clear below.

We describe our cross-validation scheme in the context of a basic grid search procedure for pedagogical reasons. For each parameter combination, do the following. For each index $k = 1, \ldots, \gamma+1$, use the k-th fold $\mathcal{T}_n^{(k)}$ to estimate the *entire* time-varying regression function (i.e. on a suitably fine grid of rescaled times and covariate values) using the "global" estimator. Afterwards, calculate the residuals

based on the observations in the $(k+\gamma+1)$-th fold $\mathcal{T}_n^{(k+\gamma+1)}$. By construction, there are γ time points in between the observations in $\mathcal{T}_n^{(k)}$ and $\mathcal{T}_n^{(k+\gamma+1)}$. Next, reverse the roles of the folds. That is, for each index $k = 1, \ldots, \gamma+1$, estimate the *entire* time-varying regression function (i.e. on a suitably fine grid) using the $(k+\gamma+1)$-th fold $\mathcal{T}_n^{(k+\gamma+1)}$, and then calculate the corresponding residuals based on the observations in the k-th fold $\mathcal{T}_n^{(k)}$. Finally, for each $k = 1, \ldots, 2(\gamma+1)$, calculate the mean squared error $\mathrm{MSE}^{(k)}$ based on the residuals in fold $\mathcal{T}_n^{(k)}$. Select the parameter combination which yields the lowest average mean squared error

$$\overline{\mathrm{MSE}} = \frac{1}{2(\gamma+1)} \sum_{k=1}^{2(\gamma+1)} \mathrm{MSE}^{(k)}.$$

In practice, γ should be chosen large enough to account for the temporal dependence, but small enough so that there is enough data to estimate the time-varying regression functions. In our simulations with Sieve-dGCM, we use the buffer $\gamma = 1$ and the grid of parameters $\{1, 2, \ldots, 10\} \times \{1, 2, \ldots, 10\}$ corresponding to the number of sieve basis functions for time and the covariate values. Note that we allow for the regressions of X on $\boldsymbol{Z}$ and Y on $\boldsymbol{Z}$ to have different numbers of basis functions. In future work, we will study the statistical properties of this cross-validation procedure as the buffer $\gamma = \gamma_n$ grows with the sample size n using infill asymptotics. For now, this cross-validation approach serves as a practical technique for parameter selection for generic "global" estimators of time-varying regression functions, such as the sieve estimator.

Next, we discuss how to select the lag-window size parameter L_n for the covariance estimator with a version of the minimum volatility method suggested by Luo and Wu [LW23]. First, select $H \in \mathbb{N}$ candidate lag-window sizes $l_1 < l_2 < \ldots < l_H$. For each index $h = 1, \ldots, H$, let

$$\hat{\boldsymbol{\Sigma}}_{t,n,l_h} = \frac{1}{l_h} \left(\sum_{s=t-l_h+1}^{t} \hat{\boldsymbol{R}}_{s,n} \right)^{\otimes 2}$$

be the lag-window estimate of the local long-run covariance matrix at time t using the candidate lag-window size $l_h \in \mathbb{N}$. Second, calculate the minimum volatility criterion for each $j = 1, \ldots, H$,

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

where $\Delta \in \mathbb{N}$ is chosen heuristically to balance robustness and adaptivity, and

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

with $h_1 = 1 \vee (j - \Delta)$ and $h_2 = H \wedge (j + \Delta)$. Third, select the lag-window size L_n^* that corresponds to the index j^* which yields the smallest minimum volatility criterion

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

We use the following setup in our simulations. We consider $H = \lfloor n/2 \rfloor$ candidate lag-windows with sizes $l_1 = 1, l_2 = 2, \ldots, l_H = \lfloor n/2 \rfloor$. We use $\Delta = 12$ so that 25 consecutive lag-window sizes are typically used in the calculation of the minimum volatility criterion $\mathbf{MV}(j)$ for each $j = 1, \ldots, H$.

5.2 Analysis of level and power

To begin, we investigate the setting with $d_X = 1$, $d_Y = 1$, $d_Z = 1$ with no time-offsets, so $A = \{0\}$, $B = \{0\}$, $C = \{0\}$, and $\mathcal{T}_n = [n]$. In this case, we can simply refer to the process as

$$(X_{t,n}, Y_{t,n}, Z_{t,n})_{t \in [n]}.$$

We test for the null hypothesis

$$Y_{t,n} \perp\!\!\!\perp X_{t,n} \mid Z_{t,n} \text{ for all times } t \in [n],$$

versus the alternative hypothesis of

$$Y_{t,n} \not\perp\!\!\!\perp X_{t,n} \mid Z_{t,n} \text{ for all times } t \in [n],$$

because we assume that we can restrict the collection of distributions to be those with time-invariant conditional dependencies. We conduct simulations with the Sieve-dGCM test with $\alpha \in \{0.025, 0.05\}$ for the quantile $\hat{q}_{1-\alpha}^{\text{boot}}$ from Algorithm 1.

For this univariate setting, we use the test statistic based on the maximum absolute value achieved by the partial sum process of residual products. Let

$$Z_{t,n} = 0.4 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) Z_{t-1,n} + \mathcal{N}(0,1).$$

We use the following DGP for the size simulations with regression complexity parameter $K \in \{1, 2\}$:

$$\begin{aligned} Y_{t,n} &= 0.4 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \exp(-Z_{t,n}^2) \sin(K Z_{t,n}) + \xi_{t,n}, \\ X_{t,n} &= 0.4 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \exp(-Z_{t,n}^2) \sin(K Z_{t,n}) + \varepsilon_{t,n}. \end{aligned}$$

We use the following DGP for the power simulations with regression complexity parameter $K \in \{1, 2\}$ and effect size parameter $\beta \in \{0.3, 0.6, 0.9\}$:

$$\begin{aligned} Y_{t,n} &= 0.4 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \exp(-Z_{t,n}^2) \sin(K Z_{t,n}) + \beta X_{t,n} + \xi_{t,n}, \\ X_{t,n} &= 0.4 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \exp(-Z_{t,n}^2) \sin(K Z_{t,n}) + \varepsilon_{t,n}. \end{aligned}$$

We use the following error processes for the power and size simulations:

$$\begin{aligned} \varepsilon_{t,n} &= 0.6 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \varepsilon_{t-1,n} + 0.3\mathcal{N}(0,1), \\ \xi_{t,n} &= 0.6 \left(\frac{2 + \sin(2\pi t/n)}{2} \right) \xi_{t-1,n} + 0.3\mathcal{N}(0,1). \end{aligned}$$

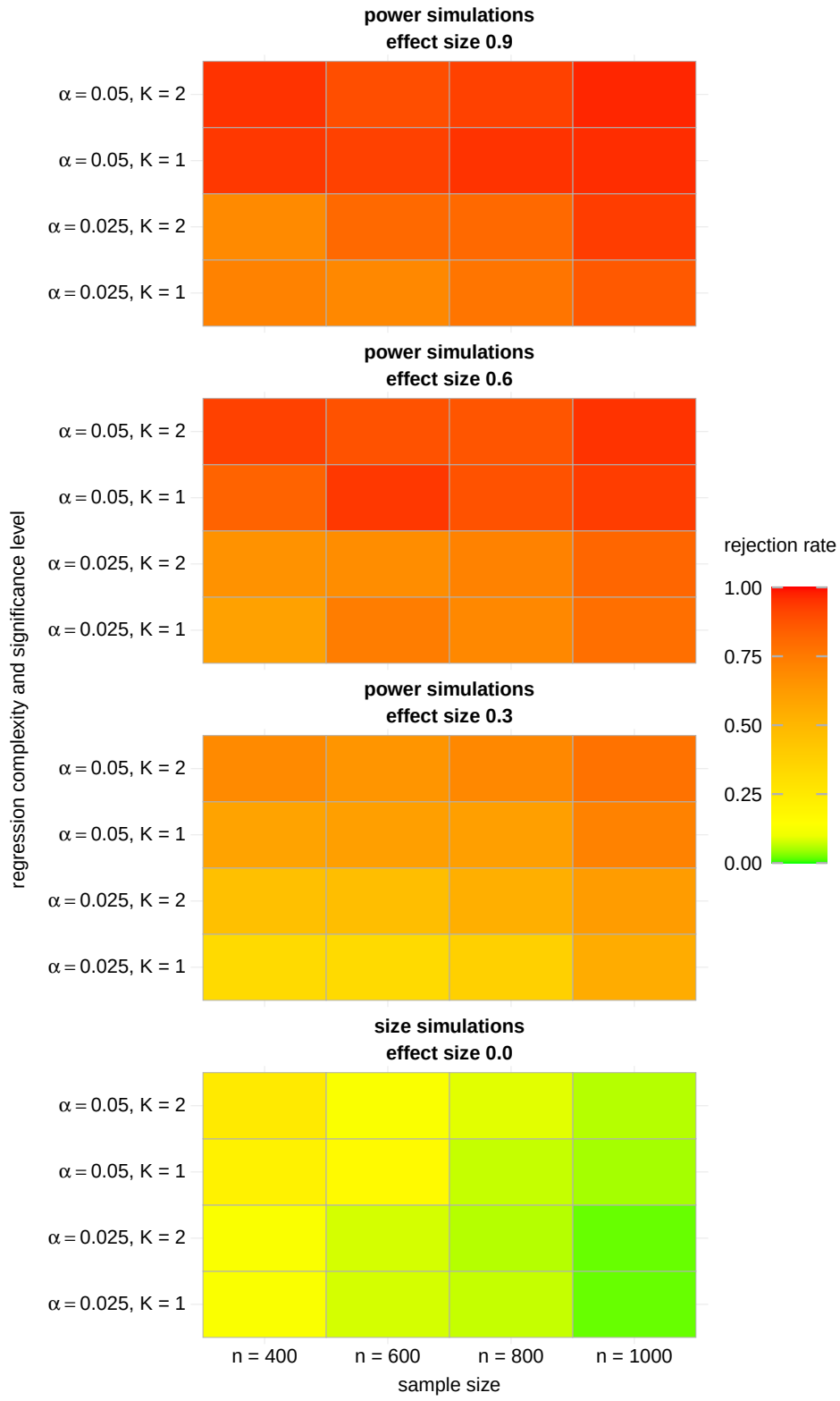

Figure 2: Empirical rejection rates for the dGCM test with the sieve time-varying regression estimator in the $d_X = 1$, $d_Y = 1$, $d_Z = 1$ setting.

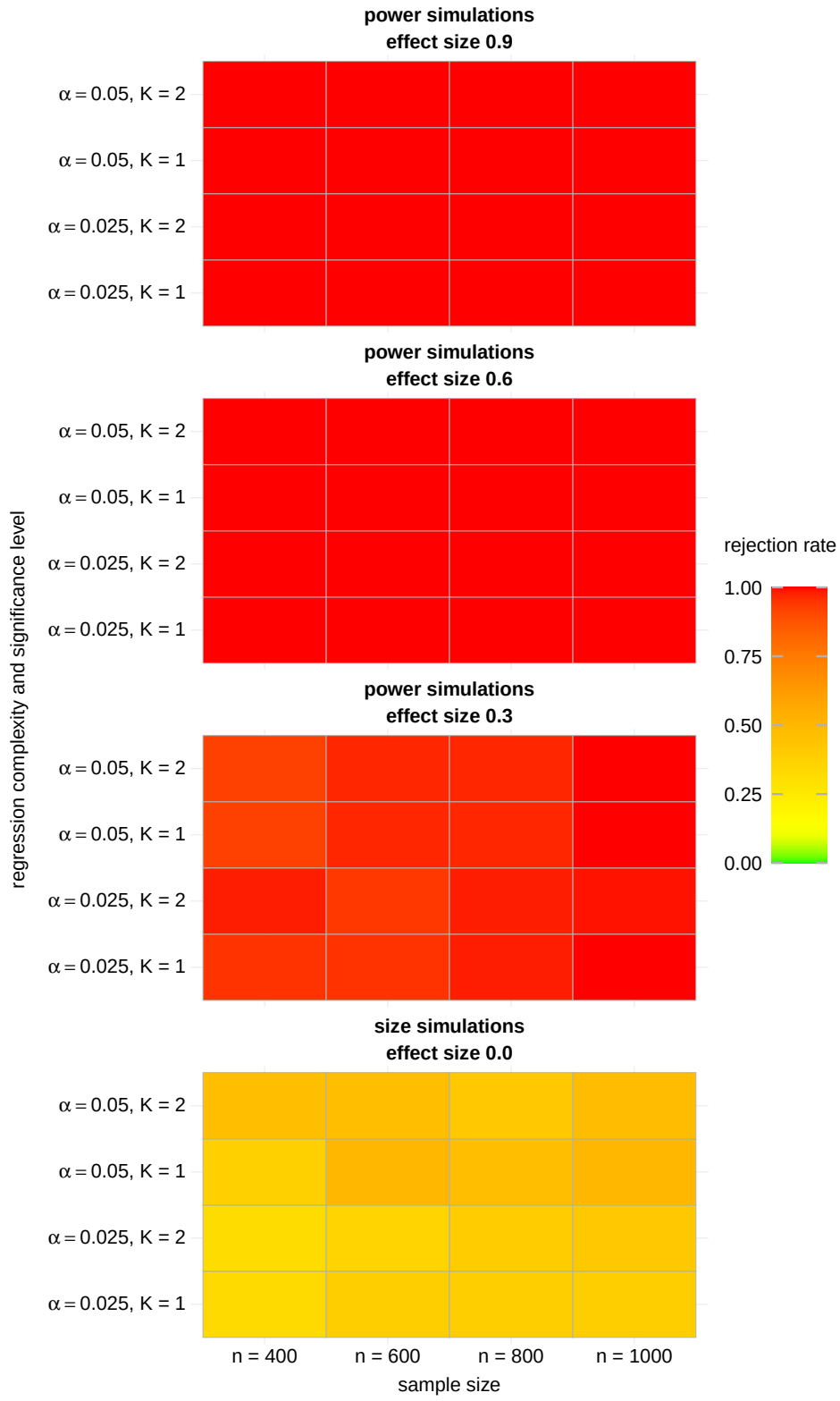

Figure 3: Empirical rejection rates for the original GCM test with a generalized additive model in the $d_X = 1$, $d_Y = 1$, $d_Z = 1$ setting.

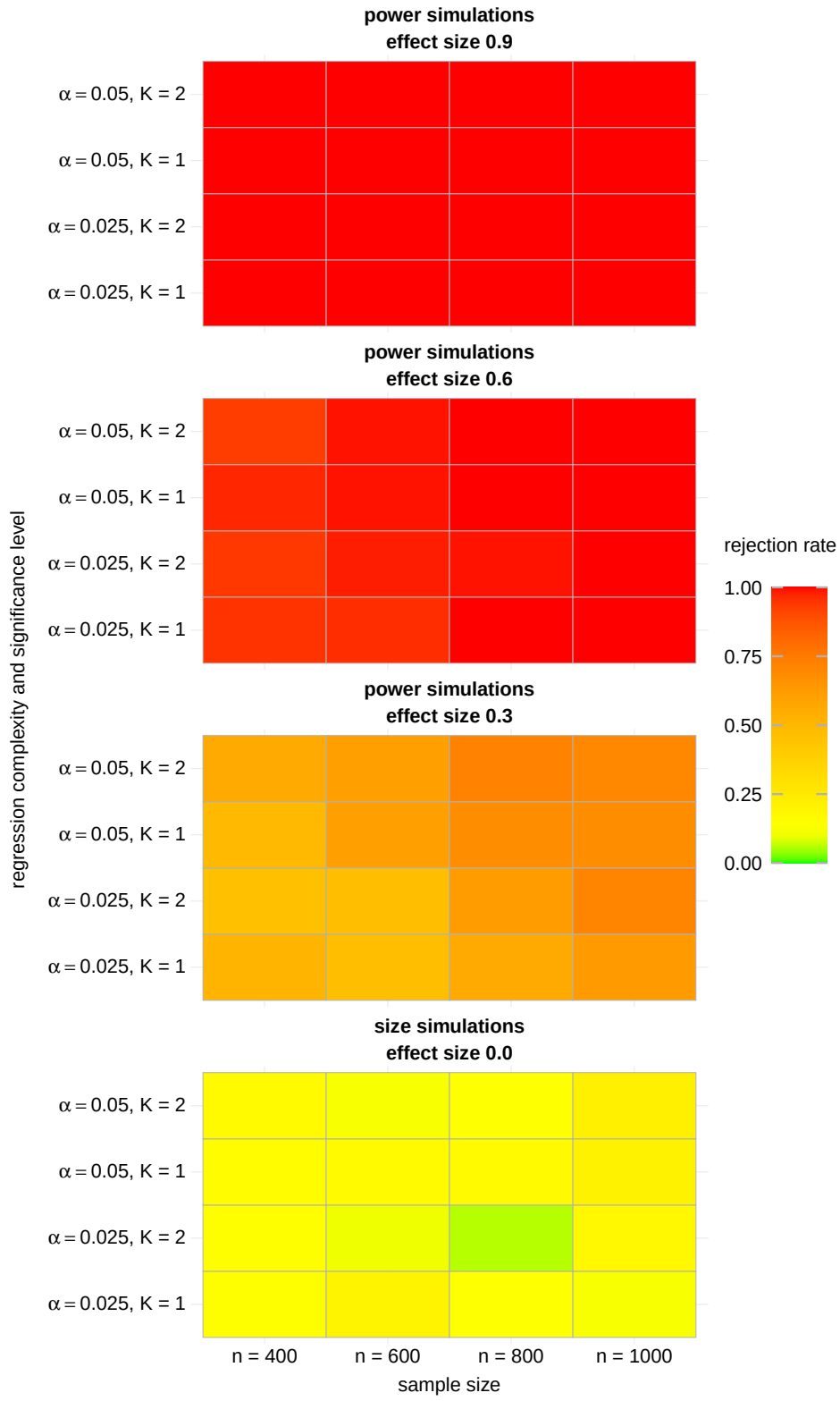

Figure 4: Empirical rejection rates for the residual prediction test with the Nyström method and a random forest model in the $d_X = 1$, $d_Y = 1$, $d_Z = 1$ setting.

Next, we consider the setting with $d_X = 1$, $d_Y = 1$, $d_Z = 2$ with no time-offsets, so $A = \{0\}$, $B = \{0\}$, $C = \{0\}$, and $\mathcal{T}_n = [n]$. In this case, we refer to the process as

$$(X_{t,n}, Y_{t,n}, Z_{t,n,1}, Z_{t,n,2})_{t\in[n]}$$

because the covariate process has two dimensions. Similarly to the previous setting, we test for the null hypothesis of

$$X_{t,n} \perp\!\!\!\perp Y_{t,n} \mid (Z_{t,n,1}, Z_{t,n,2}) \text{ for all times } t \in [n],$$

versus the alternative hypothesis of

$$X_{t,n} \not\perp\!\!\!\perp Y_{t,n} \mid (Z_{t,n,1}, Z_{t,n,2}) \text{ for all times } t \in [n],$$

because we assume that we can restrict the collection of distributions to be those with time-invariant conditional dependencies. We conduct simulations with the Sieve-dGCM test with $\alpha \in \{0.025, 0.05\}$ for the quantile $\hat{q}^{\text{boot}}_{1-\alpha}$ from Algorithm 1.

For this univariate setting, we use the test statistic based on the maximum absolute value achieved by the partial sum process of residual products. Let

$$Z_{t,n,k} = 0.4\left(\frac{2+\sin(2\pi t/n)}{2}\right) Z_{t-1,n,k} + \mathcal{N}(0,1).$$

We use the following DGP for the size simulations with regression complexity parameter $K \in \{1, 2\}$:

$$Y_{t,n} = 0.4\left(\frac{2+\sin(2\pi t/n)}{2}\right) \sum_{k=1}^{2} \exp(-Z_{t,n,k}^2)\sin(KZ_{t,n,k}) + \xi_{t,n},$$
$$X_{t,n} = 0.4\left(\frac{2+\sin(2\pi t/n)}{2}\right) \sum_{k=1}^{2} \exp(-Z_{t,n,k}^2)\sin(KZ_{t,n,k}) + \varepsilon_{t,n}.$$

We use the following DGP for the power simulations with regression complexity parameter $K \in \{1, 2\}$ and effect size parameter $\beta \in \{0.3, 0.6, 0.9\}$:

$$Y_{t,n} = 0.4\left(\frac{2+\sin(2\pi t/n)}{2}\right) \sum_{k=1}^{2} \exp(-Z_{t,n,k}^2)\sin(KZ_{t,n,k}) + \beta X_{t,n} + \xi_{t,n},$$
$$X_{t,n} = 0.4\left(\frac{2+\sin(2\pi t/n)}{2}\right) \sum_{k=1}^{2} \exp(-Z_{t,n,k}^2)\sin(KZ_{t,n,k}) + \varepsilon_{t,n}.$$

We use the following error processes for the power and size simulations:

$$\varepsilon_{t,n} = 0.6\left(\frac{2+\sin(2\pi t/n)}{2}\right) \varepsilon_{t-1,n} + 0.3\mathcal{N}(0,1),$$
$$\xi_{t,n} = 0.6\left(\frac{2+\sin(2\pi t/n)}{2}\right) \xi_{t-1,n} + 0.3\mathcal{N}(0,1).$$

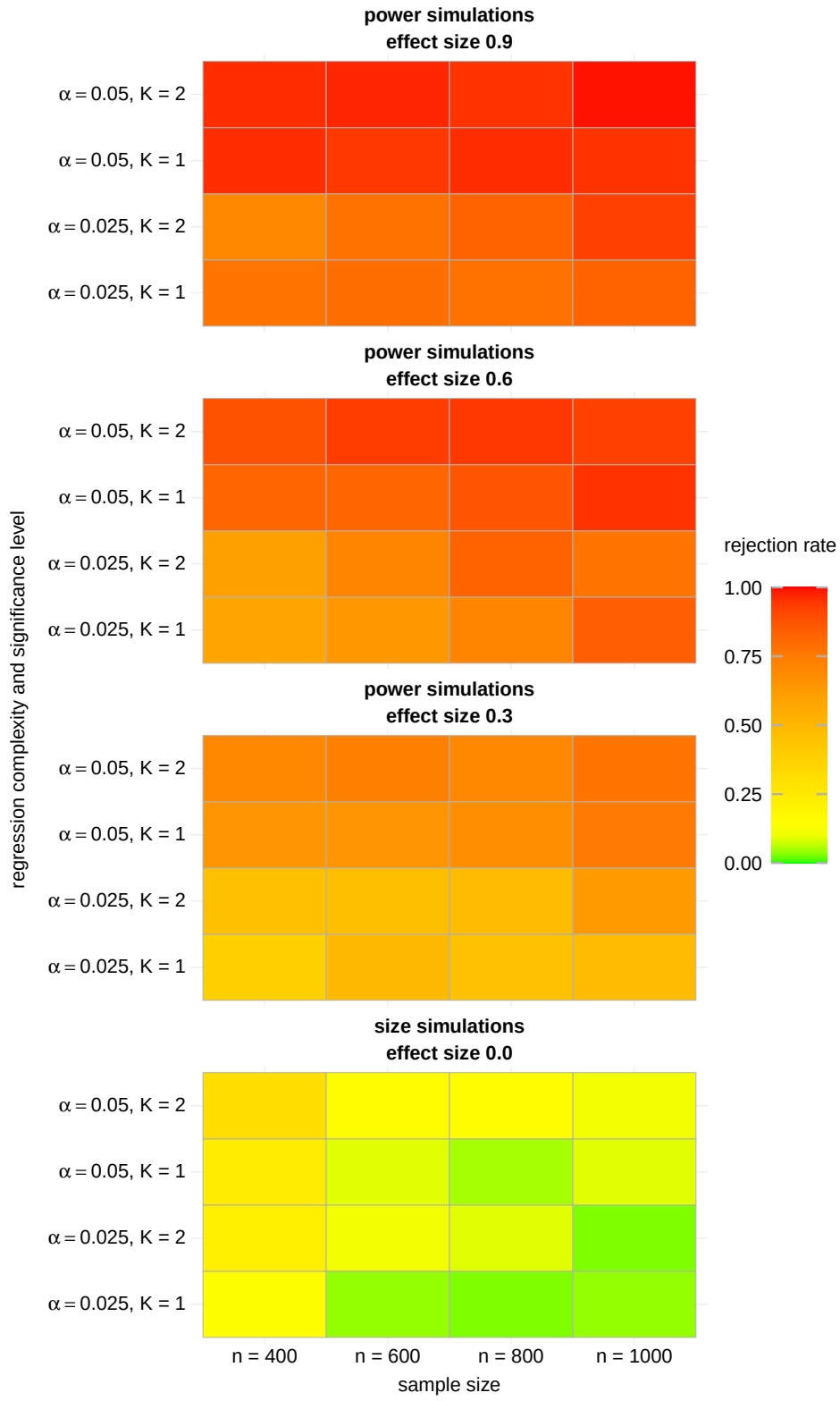

Figure 5: Empirical rejection rates for the dGCM test with the sieve time-varying regression estimator in the $d_X = 1$, $d_Y = 1$, $d_Z = 2$ setting.

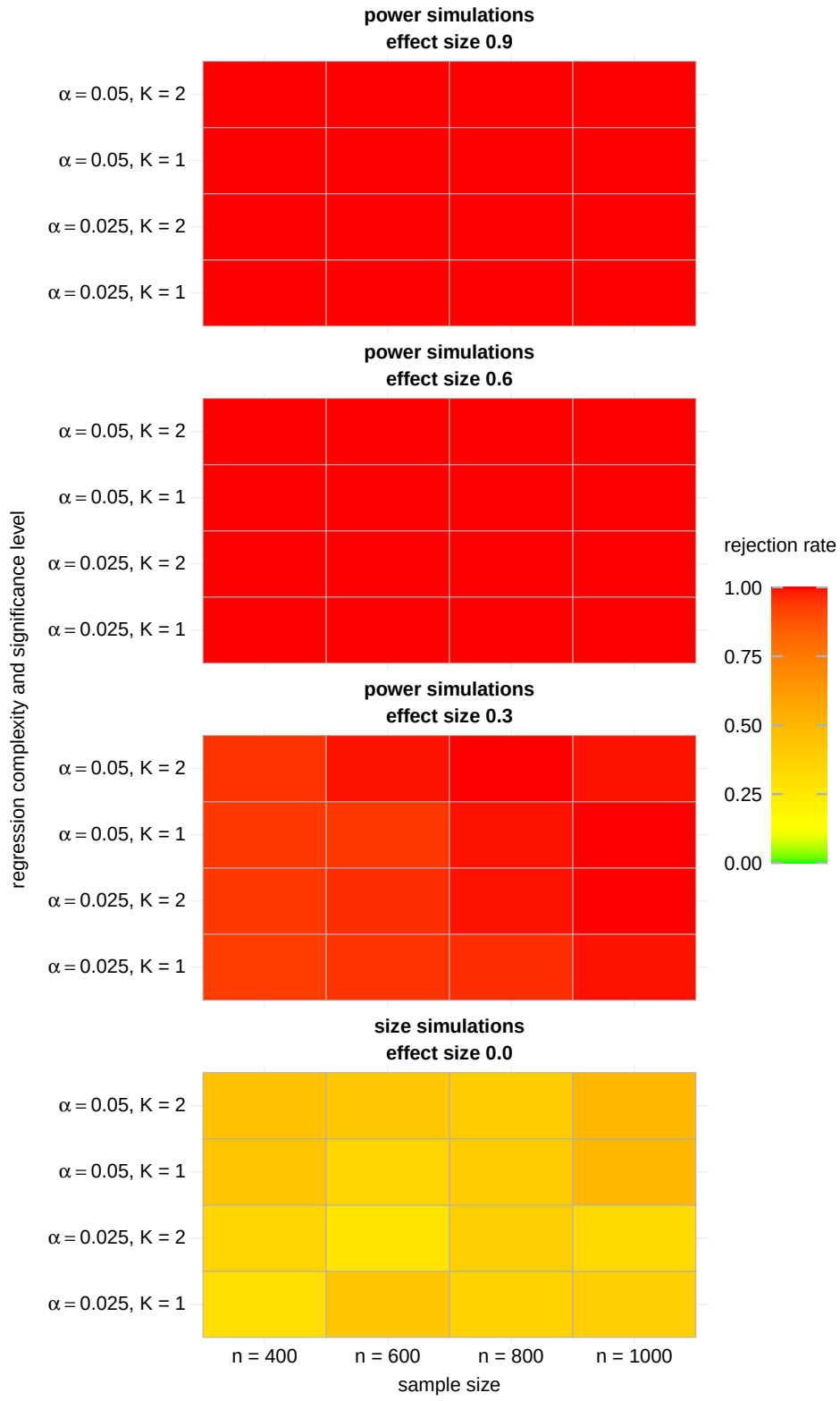

Figure 6: Empirical rejection rates for the original GCM test with a generalized additive model in the $d_X = 1$, $d_Y = 1$, $d_Z = 2$ setting.

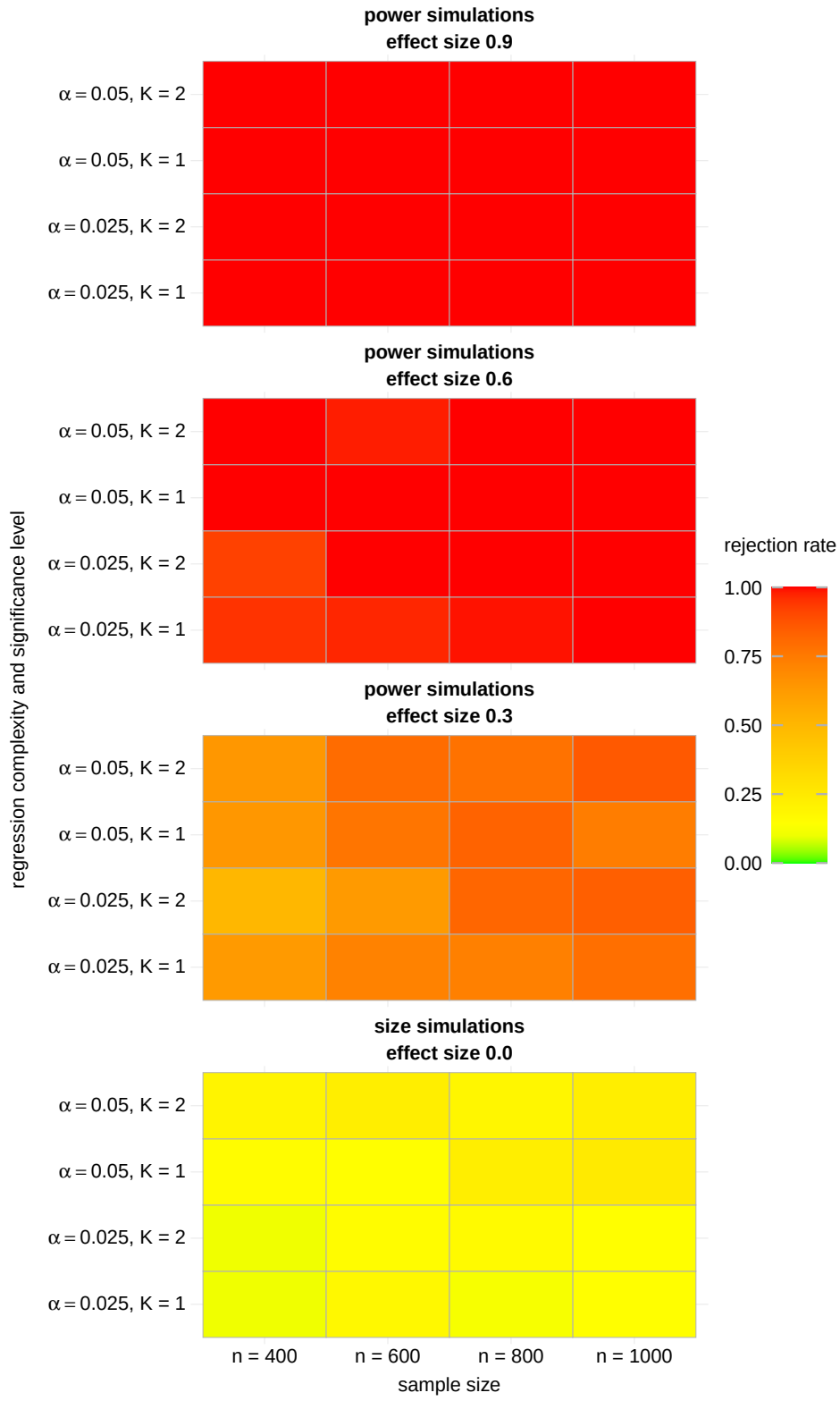

Figure 7: Empirical rejection rates for the residual prediction test with the Nyström method and a random forest model in the $d_X = 1$, $d_Y = 1$, $d_Z = 2$ setting.

5.3 Discussion of simulation results

CI testing for general discrete-time stochastic processes is a hard statistical problem [BP24]. This is not surprising in view of the original no-free-lunch result from Shah and Peters [SP20], which already shows the hardness of CI testing in the idealized iid setting. Following Shah and Peters [SP20], we provide an empirical demonstration of the hardness of CI testing with our simulations by increasing the complexity of the time-varying regression functions while holding the sample size fixed.

We observe a degradation in Type-I error control because the time-varying regression functions cannot be reliably estimated when the regression complexity parametrized by K is large relative to the sample size. This phenomena can also be observed in the simulation experiments with the GCM test; see the discussion in Section 5 of Shah and Peters [SP20]. In light of the no-free-lunch results, we understand that this is inevitable: no matter how large the sample size, it is impossible to ensure the correct significance level for every null distribution. Hence, there will always be some null distribution in which the Type-I error rate exceeds the prespecified significance level.

Crucially, the uniform level guarantee for our test only applies when the assumptions on the *sequence* of collections of distributions are satisfied. In the context of our numerical simulations, this means that any sequence of distributions parametrized by a sequence $(K_n)_{n\in\mathbb{N}}$ in which the regression complexity parameter K_n grows with n at some rate will violate Assumption 4.6. Therefore, the uniform asymptotic Type-I error control guarantee we provide is not applicable. On the other hand, if we increase the sample size while holding the degree of complexity of the time-varying regression functions fixed (e.g. setting $K_n = K = 1$ or $K_n = K = 2$ for all sample sizes n), then the uniform asymptotic Type-I error control guarantee is applicable. Although subtle, this point is crucial, as it explains why CI tests may fail to control Type-I error and highlights the transparency and "honesty" of uniform level guarantees [Li89; Kas18; Tib+18; RWG19; KBW23].

6 Future Work

In this paper, we introduced the first CI test that can be used with a single realization of a nonstationary nonlinear time series. Our test is flexible enough to be used in a broad range of downstream statistical applications in complicated settings. For instance, in our companion paper [WHR25], we apply our test to nonstationary nonlinear time series in the domain of epidemiology. Next, we discuss several promising avenues for future work.

First, we plan to develop statistical techniques for nonstationary nonlinear time series which utilize our CI test as a core component. For example, a Markov blanket-based variable selection procedure for forecasting and a causal discovery algorithm for nonstationary nonlinear time series. We note that it may be possible to use the theoretical tools for nonlinear locally stationary time series and the functional dependence measure to develop a unified causal inference framework for time series [Sag+20; RGR22; Run+23; Run+19a; Run18a]. However, this line of inquiry requires a solution to the problem of post-selection inference [KKK22], as there does not yet exist a general solution analogous to sample splitting for the iid setting.

Second, we will explore additional topics in time-varying nonlinear regression estimation. This line of work is of broad interest to any domain that deals with nonstationary nonlinear time series, but in particular it compliments our regression-based CI testing methodology. We will theoretically investigate our subsampling cross-validation procedure from Subsection 5.1. Additionally, we will develop a computationally efficient online estimation procedure for the sieve estimator from Ding and Zhou [DZ21] by taking inspiration from the Sieve Stochastic Gradient Descent (Sieve-SGD) estimator from Zhang and Simon [ZS22]. Along the way, we will study the sieve estimator in the high-dimensional setting so that the dimensions of $\boldsymbol{Z}_{t,n}$ can grow with n. We discuss additional avenues for future work related to estimating time-varying regression functions in Subsection A.2.

Additionally, it would be of interest to develop guarantees for time-varying nonlinear regression estimators in the context of processes with bounded variation-type nonstationarity as in Assumption 3.6. For example, similar blocking techniques to those in Mies and Steland [MS23] could potentially be used to study the convergence rates of the K-Nearest Neighbor fused lasso estimator from Padilla et al. [Pad+20], analogous to Padilla, Padilla, and Wang [PPW23] who analyze this estimator in the stationary β-mixing spatiotemporal setting. It may also be possible to investigate the convergence rates for deep neural network regression estimators as in Kurisu, Fukami, and Koike [KFK25], but in

the context of nonstationary nonlinear time series with the functional dependence measure.

Third, there are several possible future research directions for CI testing in this setting. While our test is based on the expected conditional covariance functional, our framework can easily be adapted to use any other functional equal to zero under the null of CI. In particular, using higher-order functionals may be of interest in more complicated settings because the expected conditional covariance functional lacks sensitivity to nonlinear relationships and interactions; see Zhang and Janson [ZJ20] for more discussion. Specifically, it would be valuable to develop such tests without compromising on practicality, which is one of the key advantages of our regression-based approach. In Subsection A.4, we discuss various "local" CI tests designed for the locally stationary time series setting. However, those test statistics utilize kernel smoothing, so the resulting tests can be very sensitive to the choice of the bandwidth parameters. Nevertheless, it still may be of interest to develop these "local" CI tests.

References

[AD12] Alekh Agarwal and John C. Duchi. "The generalization ability of online algorithms for dependent data". *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". *Dependence Modeling* 1 (2013), pp. 65–93.

[AM12] Pierre-Olivier Amblard and Olivier J. J. Michel. "The relation between Granger causality and directed information theory: a review". *Entropy* 15.1 (2012), pp. 113–143.

[AB12] Donald W.K. Andrews and Panle Jia Barwick. "Inference for parameters defined by moment inequalities: a recommended moment selection procedure". *Econometrica: Journal of the Econometric Society* 80.6 (2012), pp. 2805–2826.

[AG09] Donald W.K. Andrews and Patrik Guggenberger. "Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities". *Econometric Theory* 25.3 (2009), pp. 669–709.

[AS10] Donald W.K. Andrews and Gustavo Soares. "Inference for parameters defined by moment inequalities using generalized moment selection". *Econometrica: Journal of the Econometric Society* 78.1 (2010), pp. 119–157.

[BW23] Lujia Bai and Weichi Wu. "Time-varying correlation network analysis of non-stationary multivariate time series with complex trends". arXiv preprint arXiv:2302.05158. 2023.

[BCS20] Rina Foygel Barber, Emmanuel J. Candés, and Richard J. Samworth. "Robust inference with knockoffs". *The Annals of Statistics* 48.3 (2020), pp. 1409–1431.

[BJ22] Rina Foygel Barber and Lucas Janson. "Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling". *The Annals of Statistics* 50.5 (2022), pp. 2514–2544.

[BM15] Sumanta Basu and George Michailidis. "Regularized estimation in sparse high-dimensional time series models". *Annals of Statistics* 43.4 (2015), pp. 1535–1567.

[BR23] Sumanta Basu and Suhasini Subba Rao. "Graphical models for nonstationary time series". *The Annals of Statistics* 51.4 (2023), pp. 1453–1483.

[BSM15] Sumanta Basu, Ali Shojaie, and George Michailidis. "Network Granger causality with inherent grouping structure". *The Journal of Machine Learning Research* 16.1 (2015), pp. 417–453.

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

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 57.1 (1995), pp. 289–300.

[BY01] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". *Annals of Statistics* 29.4 (2001), pp. 1165–1188.

[Ben58] W.R. Bennett. "Statistics of regenerative digital transmission". *Bell System Technical Journal* 37.6 (1958), pp. 1501–1542.

[BHL94] Peter Bloomfield, Harry L. Hurd, and Robert B. Lund. "Periodic correlation in stratospheric ozone data". *Journal of Time Series Analysis* 15.2 (1994), pp. 127–150.

[BP24] Juraj Bodik and Olivier C. Pasche. "Granger causality in extremes". arXiv preprint arXiv:2407.09632. 2024.

[BKW24] Soham Bonnerjee, Sayar Karmakar, and Wei Biao Wu. "Gaussian approximation for nonstationary time series with optimal rate and explicit construction". 2024.

[BRT12] Taoufik Bouezmarni, Jeroen V.K. Rombouts, and Abderrahim Taamouti. "Nonparametric copula-based test for conditional independence with applications to Granger causality". *Journal of Business and Economic Statistics* 30.2 (2012), pp. 275–287.

[Bre01] Leo Breiman. "Random forests". *Machine Learning* 45 (2001), pp. 5–32.

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

[Can+18] Emmanuel J. Candés et al. "Panning for gold: model-X knockoffs for high dimensional controlled variable selection". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.

[CZK24] Abhinav Chakraborty, Jeffrey Zhang, and Eugene Katsevich. "Doubly robust and computationally efficient high-dimensional variable selection" (2024). arXiv preprint arXiv:2409.09512.

[CCW24] Jinyuan Chang, Xiaohui Chen, and Mingcong Wu. "Central limit theorems for high dimensional dependent data". *Bernoulli* 30.1 (2024), pp. 712–742.

[CSW22] Likai Chen, Ekaterina Smetanina, and Wei Biao Wu. "Estimation of nonstationary nonparametric regression model with multiplicative structure". *The Econometrics Journal* 25.1 (2022), pp. 176–214.

[CG16] Tianqi Chen and Carlos Guestrin. "Xgboost: a scalable tree boosting system". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

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

[CCK19] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. "Inference on causal and structural parameters using many moment inequalities". *The Review of Economic Studies* 86.5 (2019), pp. 1867–1900.

[CCK16] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. "Testing many moment inequalities". 2016.

[CPH22] Alexander Mangulad Christgau, Lasse Petersen, and Niels Richard Hansen. "Nonparametric conditional local independence testing". *The Annals of Statistics* 51.5 (2022), pp. 2116–2144.

[CZ25] Yan Cui and Zhou Zhou. "Optimal Short-Term Forecast for Locally Stationary Functional Time Series". *IEEE Transactions on Information Theory* (2025).

[Dah97] Rainer Dahlhaus. "Fitting time series models to nonstationary processes". *The Annals of Statistics* 25.1 (1997), pp. 1–37.

[Dah12] Rainer Dahlhaus. "Locally stationary processes". *Handbook of statistics* 30 (2012), pp. 351–413.

[DR23] Rainer Dahlhaus and Stefan Richter. "Adaptation for nonparametric estimators of locally stationary processes". *Econometric Theory* 39.6 (2023), pp. 1123–1153.

[DR19] Rainer Dahlhaus and Stefan Richter. "Cross validation for locally stationary processes". *Annals of Statistics* 47.4 (2019), pp. 2145–2173.

[DRW19] Rainer Dahlhaus, Stefan Richter, and Wei Biao Wu. "Towards a general theory for nonlinear locally stationary processes". *Bernoulli* 25.2 (2019), pp. 1013–1044.

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

[DN20] Richard A. Davis and Mikkel S. Nielsen. "Modeling of time series using random forests: theoretical developments". *Electronic Journal of Statistics* (2020), pp. 3644–3671.

[DT20] Philip A. Dawid and Ambuj Tewari. "On learnability under general stochastic processes". arXiv preprint arXiv:2005.07605. 2020.

[DWZ19] Holger Dette, Weichi Wu, and Zhou Zhou. "Change point analysis of correlation in nonstationary time series". *Statistica Sinica* 29.2 (2019), pp. 611–643.

[DZ23] Xiucai Ding and Zhou Zhou. "Autoregressive approximations to nonstationary time series with inference and applications". *The Annals of Statistics* 51.3 (2023), pp. 1207–1231.

[DZ21] Xiucai Ding and Zhou Zhou. "Simultaneous sieve inference for time-inhomogeneous nonlinear time series regression". arXiv preprint arXiv:2112.08545. 2021.

[Don+23] Xinshuai Dong et al. "On the three demons in causality in finance: time resolution, nonstationarity, and latent factors". arXiv preprint arXiv:2401.05414. 2023.

[Dor+14] Gary Doran et al. "A permutation-based kernel conditional independence test". *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence* (2014), pp. 132–141.

[Eic12] Michael Eichler. "Graphical modelling of multivariate time series". *Probability Theory and Related Fields* 153 (2012), pp. 233–268.

[FFX20] Jianqing Fan, Yang Feng, and Lucy Xia. "A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models". *Journal of Econometrics* 218.1 (2020), pp. 119–139.

[FHG23] Muhammad Hasan Ferdous, Uzma Hasan, and Md Osman Gani. "Cdans: temporal causal discovery from autocorrelated and non-stationary time series data". *Proceedings of Machine Learning Research* 219 (2023), pp. 186–207.

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

[FM82] Jean-Pierre Florens and Michel Mouchart. "A note on noncausality". *Econometrica: Journal of the Econometric Society* (1982), pp. 583–591.

[Fuk+07] Kenji Fukumizu et al. "Kernel measures of conditional dependence". *Advances in Neural Information Processing Systems* 20 (2007), pp. 489–496.

[Gar94] William A. Gardner. *Cyclostationarity in communications and signal processing*. IEEE Press, 1994.

[GNP06] William A. Gardner, Antonio Napolitano, and Luigi Paura. "Cyclostationarity: half a century of research". *Signal Processing* 86.4 (2006), pp. 639–697.

[Goe20] Benjamin Goehry. "Random forests for time-dependent processes". *ESAIM: Probability and Statistics* 24 (2020), pp. 801–826.

[Gra69] Clive WJ Granger. "Investigating causal relations by econometric models and cross-spectral methods". *Econometrica: Journal of the Econometric Society* 37.3 (1969), pp. 424–438.

[Gra80] Clive WJ Granger. "Testing for causality: a personal viewpoint". *Journal of Economic Dynamics and Control* 2 (1980), pp. 329–352.

[Gre+07] Arthur Gretton et al. "A kernel statistical test of independence". *Advances in Neural Information Processing Systems* 20 (2007).

[GHL24] Peter Grünwald, Alexander Henzi, and Tyron Lardy. "Anytime-valid tests of conditional independence under model-X". *Journal of the American Statistical Association* 119.546 (2024), pp. 1554–1565.

[Gup12] Shuva Gupta. "A note on the asymptotic distribution of LASSO estimator for correlated data". *Sankhya A* 74.1 (2012), pp. 10–28.

[HS14] Hanyuan Hang and Ingo Steinwart. "Fast learning from alpha-mixing observations". *Journal of Multivariate Analysis* 127 (2014), pp. 184–199.

[Han21] Steve Hanneke. "Learning whenever learning is possible: universal learning under general stochastic processes". *The Journal of Machine Learning Research* 22.130 (2021), pp. 1–116.

[HY19] Steve Hanneke and Liu Yang. "Statistical learning under nonstationary mixing processes". *International Conference on Artificial Intelligence and Statistics* 22.1 (2019), pp. 1678–1686.

[Hin+22] Oliver Hines et al. "Demystifying statistical learning based on efficient influence functions". *The American Statistician* 76.3 (2022), pp. 292–304.

[Hua+20] Biwei Huang et al. "Causal discovery from heterogeneous/nonstationary data". *Journal of Machine Learning Research* 21.89 (2020), pp. 1–53.

[HJ20] Dongming Huang and Lucas Janson. "Relaxing the assumptions of knockoffs by conditioning". *The Annals of Statistics* 48.5 (2020), pp. 3021–3042.

[Hua10] Tzee-Ming Huang. "Testing conditional independence using maximal nonlinear conditional correlation". *The Annals of Statistics* 38.4 (2010), pp. 2047–2091.

[Hyn18] R. J. Hyndman. *Forecasting: principles and practice*. OTexts, 2018.

[IM04] Guido W. Imbens and Charles F. Manski. "Confidence intervals for partially identified parameters". *Econometrica: Journal of the Econometric Society* 72.6 (2004), pp. 1845–1857.

[KKR24] Iden Kalemaj, Shiva Kasiviswanathan, and Aaditya Ramdas. "Differentially private conditional independence testing". *International Conference on Artificial Intelligence and Statistics* 238 (2024), pp. 3700–3708.

[Kal21] Olav Kallenberg. *Foundations of modern probability*. Third edition. Springer, 2021.

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

[KW20] Sayar Karmakar and Wei Biao Wu. "Optimal Gaussian approximation for multiple time series". 30.3 (2020), pp. 1399–1417.

[Kas18] Maximilian Kasy. "Uniformity and the delta method". *Journal of Econometric Methods* 8.1 (2018).

[Ke+17] Guolin Ke et al. "Lightgbm: a highly efficient gradient boosting decision tree". *Advances in Neural Information Processing Systems* 30 (2017).

[Ken24] Edward H. Kennedy. "Semiparametric doubly robust targeted double machine learning: a review". *Handbook of Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, 2024, pp. 207–236.

[Kim+22] Ilmun Kim et al. "Local permutation tests for conditional independence". *The Annals of Statistics* 50.6 (2022), pp. 3388–3414.

[KR24] Jonas Krampe and Suhasini Subba Rao. "Inverse covariance operators of multivariate nonstationary time series". *Bernoulli* 30.2 (2024), pp. 1177–1196.

[KKK22] Arun K. Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. "Post-selection inference". *Annual Review of Statistics and Its Application* 9.1 (2022), pp. 505–527.

[KBW23] Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. "Median regularity and honest inference". *Biometrika* 110.3 (2023), pp. 831–838.

[KFK25] Daisuke Kurisu, Riku Fukami, and Yuta Koike. "Adaptive deep learning for nonlinear time series models". *Bernoulli* 31.1 (2025), pp. 240–270.

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

[KM14] Vitaly Kuznetsov and Mehryar Mohri. "Generalization bounds for time series prediction with non-stationary processes". *Algorithmic Learning Theory* 25 (2014), pp. 260–274.

[KM15] Vitaly Kuznetsov and Mehryar Mohri. "Learning theory and algorithms for forecasting non-stationary time series". *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". *The Annals of Statistics* 17.3 (1989), pp. 1001–1008.

[Li+11] Lingling Li et al. "Higher order inference on a treatment effect under low regularity conditions". *Statistics and Probability Letters* 81.7 (2011), pp. 821–828.

[Liu+22] Molei Liu et al. "Fast and powerful conditional randomization testing via distillation". *Biometrika* 109.2 (2022), pp. 277–293.

[LXW13] Weidong Liu, Han Xiao, and Wei Biao Wu. "Probability and moment inequalities under dependence". *Statistica Sinica* 23.3 (2013), pp. 1257–1272.

[Liu+23] Zhaolu Liu et al. "Kernel-based joint independence tests for multivariate stationary and non-stationary time series". *Royal Society Open Science* 10.11 (2023).

[Lob01] Ignacio N. Lobato. "Testing that a dependent process is uncorrelated". *Journal of the American Statistical Association* 96.455 (2001), pp. 1066–1076.

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

[Lun+22] Anton Rask Lundborg et al. "The projected covariance measure for assumption-lean variable significance testing". arXiv preprint arXiv:2211.02039. 2022.

[LW23] Tianpai Luo and Weichi Wu. "Simultaneous inference for monotone and smoothly time varying functions under complex temporal dynamics". arXiv preprint arXiv:2310.02177. 2023.

[Lüt05] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springers Science and Business Media, 2005.

[MS19] Daniel Malinsky and Peter Spirtes. "Learning the structure of a nonstationary vector autoregression". *International Conference on Artificial Intelligence and Statistics* 89 (2019), pp. 2986–2994.

[Man+24] Georg Manten et al. "Signature kernel conditional independence tests in causal discovery for stochastic processes". arXiv preprint arXiv:2402.18477. 2024.

[Mar05] Dimitris Margaritis. "Distribution-free learning of Bayesian network structure in continuous domains". *AAAI* 5 (2005), pp. 825–830.

[Mie23] Fabian Mies. "Functional estimation and change detection for nonstationary time series". *Journal of the American Statistical Association* 118.542 (2023), pp. 1011–1022.

[Mie24] Fabian Mies. "Strong Gaussian approximations with random multipliers". arXiv preprint arXiv:2412.14346. 2024.

[MS23] Fabian Mies and Ansgar Steland. "Sequential Gaussian approximation for nonstationary time series in high dimensions". *Bernoulli* 29.4 (2023), pp. 3114–3140.

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

[Nap16a] Antonio Napolitano. "Cyclostationarity: limits and generalizations". *Signal Processing* 120 (2016), pp. 323–347.

[Nap16b] Antonio Napolitano. "Cyclostationarity: new trends and applications". *Signal Processing* 120 (2016), pp. 385–408.

[NBW21] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. "Minimax optimal conditional independence testing". *The Annals of Statistics* 49.4 (2021), pp. 2151–2177.

[Niu+24] Ziang Niu et al. "Reconciling model-X and doubly robust approaches to conditional independence testing". *The Annals of Statistics* 52.3 (2024), pp. 895–921.

[PPW23] Carlos Misael Madrid Padilla, Oscar Hernan Madrid Padilla, and Daren Wang. "Temporal-spatial model via trend filtering". arXiv preprint arXiv:2308.16172. 2023.

[Pad+20] Oscar Hernan Madrid Padilla et al. "Adaptive nonparametric regression with the k-nearest neighbour fused lasso". *Biometrika* 107.2 (2020), pp. 293–310.

[PP79] Emanuel Parzen and Marcello Pagano. "An approach to modeling seasonally stationary time series". *Journal of Econometrics* 9.1-2 (1979), pp. 137–153.

[Pat+09] Hoyer Patrik et al. "Nonlinear causal discovery with additive noise models". *Advances in Neural Information Processing Systems* 21 (2009).

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

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[PZZ23] Ling Peng, Yan Zhu, and Wenxuan Zhong. "Lasso regression in sparse linear model with phi-mixing errors". *Metrika* 86.1 (2023), pp. 1–26.

[Pet+14] Jonas Peters et al. "Causal discovery with continuous additive noise models". *Journal of Machine Learning Research* 15.58 (2014), pp. 2009–2053.

[Pis16] Gilles Pisier. *Martingales in Banach spaces*. Vol. 155. Cambridge University Press, 2016.

[QKC15] Christopher J. Quinn, Negar Kiyavash, and Todd P. Coleman. "Directed information graphs". *IEEE Transactions on Information Theory* 61.12 (2015), pp. 6887–6909.

[Ram14] Joseph D Ramsey. "A scalable conditional independence test for nonlinear, non-gaussian data". arXiv preprint arXiv:1401.5031. 2014.

[Reb80] Rolando Rebolledo. "Central limit theorems for local martingales". *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 51.3 (1980), pp. 269–286.

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

[RS13] Yeonwoo Rho and Xiaofeng Shao. "Improving the bandwidth-free inference methods by prewhitening". *Journal of Statistical Planning and Inference* 143.11 (2013), pp. 1912–1922.

[RWG19] Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. "Bootstrapping and sample splitting for high-dimensional, assumption-lean inference". *The Annals of Statistics* 47.6 (2019), pp. 3438–3469.

[Rob+08] James Robins et al. "Higher order influence functions and minimax estimation of nonlinear functionals". *Probability and Statistics: Essays in Honor of David A. Freedman* 2 (2008), pp. 335–422.

[Rob+09] James Robins et al. "Semiparametric minimax rates". *Electronic Journal of Statistics* 3 (2009), pp. 1305–1321.

[Rob+17] James M. Robins et al. "Minimax estimation of a functional on a structured high-dimensional model". *Annals of Statistics* 45.5 (2017), pp. 1951–1987.

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

[RS08] Joseph P. Romano and Azeem M. Shaikh. "Inference for identifiable parameters in partially identified econometric models". *Journal of Statistical Planning and Inference* 138.9 (2008), pp. 2786–2807.

[RSW14] Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. "A practical two-step method for testing moment inequalities". *Econometrica: Journal of the Econometric Society* 82.5 (2014), pp. 1979–2002.

[Ros61] Murray Rosenblatt. "Independence and dependence". *Proc. 4th Berkeley Symp. on Math. Statist. and Prob.* 2 (1961), pp. 431–443.

[RK16] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*. John Wiley and Sons, 2016.

[Run18a] Jakob Runge. "Causal network reconstruction from time series: from theoretical assumptions to practical estimation". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.7 (2018).

[Run18b] Jakob Runge. "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information". *International Conference on Artificial Intelligence and Statistics* 84 (2018), pp. 938–947.

[Run+23] Jakob Runge et al. "Causal inference for time series". *Nature Reviews Earth and Environment* 4.7 (2023), pp. 487–505.

[Run+19a] Jakob Runge et al. "Detecting and quantifying causal associations in large nonlinear time series datasets". *Science Advances* 5.11 (2019).

[Run+19b] Jakob Runge et al. "Inferring causation from time series in Earth system sciences". *Nature communications* 10.1 (2019).

[SGF24] Agathe Sadeghi, Achintya Gopal, and Mohammad Fesanghary. "Causal discovery from nonstationary time series". *International Journal of Data Science and Analytics* 29.2 (2024), pp. 1–27.

[Sag+20] Elena Saggioro et al. "Reconstructing regime-dependent causal relationships from observational time series". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30.11 (2020).

[SHB22] Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. "The weighted generalised covariance measure". *Journal of Machine Learning Research* 23.273 (2022), pp. 1–68.

[Sen+17] Rajat Sen et al. "Model-powered conditional independence test". *Advances in Neural Information Processing Systems* 30 (2017).

[SP11] Sohan Seth and Jose C. Principe. "Assessing Granger non-causality using nonparametric measure of conditional independence". *IEEE Transactions on Neural Networks and Learning Systems* 1 (2011), pp. 47–59.

[SMR23] Shalev Shaer, Gal Maman, and Yaniv Romano. "Model-X sequential testing for conditional independence via testing by betting". *International Conference on Artificial Intelligence and Statistics* (2023), pp. 2054–2086.

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

[Sha10] Xiaofeng Shao. "A self-normalized approach to confidence interval construction in time series". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.3 (2010), pp. 343–366.

[Sha15] Xiaofeng Shao. "Self-normalization for time series: a review of recent developments". *Journal of the American Statistical Association* 110.512 (2015), pp. 1797–1817.

[SF22] Ali Shojaie and Emily B. Fox. "Granger causality: a review and recent advances". *Annual Review of Statistics and Its Application* 9.1 (2022), pp. 289–319.

[SGS01] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2001.

[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". *Annals of Statistics* 37.2 (2009), pp. 841–875.

[SHS09] Ingo Steinwart, Don Hush, and Clint Scovel. "Learning from dependent observations". *Journal of Multivariate Analysis* 100.1 (2009), pp. 175–194.

[SW07] Liangjun Su and Halbert White. "A consistent characteristic function-based test for conditional independence". *Journal of Econometrics* 141.2 (2007), pp. 807–834.

[SW08] Liangjun Su and Halbert White. "A nonparametric Hellinger metric test for conditional independence". *Econometric Theory* 24.4 (2008), pp. 829–864.

[SW14] Liangjun Su and Halbert White. "Testing conditional independence via empirical likelihood". *Journal of Econometrics* 182.1 (2014), pp. 27–44.

[Tan+21] Alex Tank et al. "Neural Granger causality". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8 (2021), pp. 4267–4279.

[Tib+18] Ryan J Tibshirani et al. "Uniform asymptotic inference and the bootstrap after model selection". *The Annals of Statistics* 46.3 (2018), pp. 1255–1287.

[Vog12] Michael Vogt. "Nonparametric regression for locally stationary time series". *Annals of Statistics* 50.5 (2012), pp. 2601–2633.

[WNR24] Jonas Wahl, Urmi Ninad, and Jakob Runge. "Foundations of causal discovery on groups of variables". *Journal of Causal Inference* 12.1 (2024).

[WT23] Di Wang and Ruey S. Tsay. "Rate-optimal robust estimation of high-dimensional vector autoregressive models". *The Annals of Statistics* 51.2 (2023), pp. 846–877.

[WR23] Ian Waudby-Smith and Aaditya Ramdas. "Distribution-uniform anytime-valid inference". arXiv preprint arXiv:2311.03343. 2023.

[Wau+24] Ian Waudby-Smith et al. "Time-uniform central limit theory and asymptotic confidence sequences". *Annals of Statistics* (2024).

[WHR25] Michael Wieck-Sosa, Michel F. C. Haddad, and Aaditya Ramdas. "Lightweight methods for detecting leading indicators in unstable environments and measuring importance over time". 2025.

[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". *The Annals of Statistics* 48.2 (2020), pp. 1124–1142.

[Wu11] Wei Biao Wu. "Asymptotic theory for stationary processes". *Statistics and its Interface* 4.2 (2011), pp. 207–226.

[Wu05] Wei Biao Wu. "Nonlinear system theory: another look at dependence". *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". *Handbook of Statistics*. Vol. 30. Elsevier, 2012, pp. 187–209.

[WW16] Wei-Biao Wu and Ying Nian Wu. "Performance bounds for parameter estimates of high-dimensional linear models with correlated errors". *Electronic Journal of Statistics* 10.1 (2016), pp. 352–379.

[WZ19] Weichi Wu and Zhou Zhou. "Multiscale jump testing and estimation under complex temporal dynamics". arXiv preprint arXiv:1909.06307. 2019.

[XCG24] Jiaqi Xia, Yu Chen, and Xiao Guo. "Inference for high-dimensional linear models with locally stationary error processes". *Journal of Time Series Analysis* 45.1 (2024), pp. 78–102.

[XX18] Fang Xie and Zhijie Xiao. "Square-root LASSO for high-dimensional sparse linear systems with weakly dependent errors". *Journal of Time Series Analysis* 39.2 (2018), pp. 212–238.

[XXY17] Fang Xie, Lihu Xu, and Youcai Yang. "Lasso for sparse linear regression with exponentially beta-mixing errors". *Statistics and Probability Letters* 125 (2017), pp. 64–70.

[YN21] Kashif Yousuf and Serena Ng. "Boosting high dimensional predictive regressions with time varying parameters". *Journal of Econometrics* 224.1 (2021), pp. 60–87.

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

[ZW17] Danna Zhang and Wei Biao Wu. "Gaussian approximation for high dimensional time series". *Annals of Statistics* 45.5 (2017), pp. 1895–1919.

[ZZG17] Hao Zhang, Shuigeng Zhou, and Jihong Guan. "Causal discovery using regression-based conditional independence tests". *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (2017), pp. 1250–1256.

[Zha+19] Hao Zhang et al. "Measuring conditional independence by independent residuals for causal discovery". *ACM Transactions on Intelligent Systems and Technology* 10.5 (2019).

[Zha+11] Kun Zhang et al. "Kernel-based conditional independence test and application in causal discovery". *Conference on Uncertainty in Artificial Intelligence* 27 (2011), pp. 804–813.

[ZJ20] Lu Zhang and Lucas Janson. "Floodgate: inference for model-free variable importance". arXiv preprint arXiv:2007.01283. 2020.

[ZS22] Tianyu Zhang and Noah Simon. "A sieve stochastic gradient descent estimator for online nonparametric regression in Sobolev ellipsoids". *The Annals of Statistics* 50.5 (2022), pp. 2848–2871.

[ZW15] Ting Zhang and Wei Biao Wu. "Time-varying nonlinear regression models: nonparametric estimation and model selection". *Annals of Statistics* 43.2 (2015), pp. 741–768.

[ZC14] Xianyang Zhang and Guang Cheng. "Bootstrapping high dimensional time series". arXiv preprint arXiv:1406.1037. 2014.

[ZS24] Yi Zhang and Xiaofeng Shao. "Another look at bandwidth-free inference: a sample splitting approach". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 86.1 (2024), pp. 246–272.

[Zho13] Zhou Zhou. "Heteroscedasticity and autocorrelation robust structural change detection". *Journal of the American Statistical Association* 108.502 (2013), pp. 726–740.

[ZW09] Zhou Zhou and Wei Biao Wu. "Local linear quantile estimation for nonstationary time series". *The Annals of Statistics* 37.5B (2009), pp. 2696–2729.

A Extensions

A.1 Alternative test statistics

Consider the test statistic

$$S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n) = \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \in \mathcal{T}_{n,L}} \hat{\boldsymbol{R}}_{t,n} \right|\right|_p,$$

based on the ℓ_p-norm ($p \geq 2$) of the full sum of residual products. For example, we can use the test statistics

$$S^{\star}_{n,\infty}(\hat{\boldsymbol{R}}_n) = \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \in \mathcal{T}_{n,L}} \hat{\boldsymbol{R}}_{t,n} \right|\right|_{\infty}, \; S^{\star}_{n,2}(\hat{\boldsymbol{R}}_n) = \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \in \mathcal{T}_{n,L}} \hat{\boldsymbol{R}}_{t,n} \right|\right|_{2}.$$

Crucially, the full sum test statistic $S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n)$ will not have power against alternatives in which the time-averages of the time-varying expected conditional covariances are close to zero (e.g. positive during the first half of times, and negative during the second half). On the other hand, the maximum partial sum test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ from (8) *does* have power against these alternatives. If the time-varying expected conditional covariances are suspected to consistently maintain the same sign (whether positive or negative), then users might be able to gain *some* power by using $S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n)$, although we emphasize that $S_{n,p}(\hat{\boldsymbol{R}}_n)$ will also have power against these alternatives. However, in settings where we have little prior knowledge about the time-varying expected conditional covariances between the nonstationary processes under alternatives, then the maximum partial sum test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ should be used because it has power against a wider range of alternatives. For similar reasons, we recommend using $S_{n,p}(\hat{\boldsymbol{R}}_n)$ when conducting automated multiple CI testing (e.g. for screening out irrelevant time series in a large database of possible forecasting signals).

It is perhaps most intuitive to frame the problem in the following way. Consider the time-varying partially linear model

$$\mathbb{E}_P(Y_{t,n,j,b}|X_{t,n,i,a}, \boldsymbol{Z}_{t,n}) = \beta_{P,t,n,m} X_{t,n,i,a} + h_{P,t,n,j,b}(\boldsymbol{Z}_{t,n}),$$

for some function $h_{P,t,n,j,b}(\cdot)$. When the time-varying conditional expectation $\mathbb{E}_P(Y_{t,n,j,b}|X_{t,n,i,a}, \boldsymbol{Z}_{t,n})$ is assumed to have this time-varying partially linear form, the time-varying coefficient $\beta_{P,t,n,m}$ is

equal to the expected conditional covariance of $X_{t,n,i,a}$ and $Y_{t,n,j,b}$ given $\boldsymbol{Z}_{t,n}$ divided by the expected conditional variance of $X_{t,n,i,a}$ given $\boldsymbol{Z}_{t,n}$; see Robins et al. [Rob+09] and Hines et al. [Hin+22] for more discussion. If domain knowledge suggests that the time-varying coefficients $(\beta_{P,t,n,m})_{m=(i,j,a,b)\in\mathcal{D}_n}$ consistently maintain the same sign over time $t \in \mathcal{T}_n$, then the full sum test statistic $S^{\star}_{n,p}(\hat{\boldsymbol{R}}_n)$ can be used to gain some power. Otherwise, if we cannot make this assumption, then use the maximum partial sum test statistic $S_{n,p}(\hat{\boldsymbol{R}}_n)$ which has power against a broader spectrum of alternatives.

A.2 Other approaches for estimating time-varying regression functions

The theoretical justification of the dGCM test requires that we improve our estimates of the time-varying regression functions as n grows; see Theorem 3.1 for the details. In this subsection, we discuss additional asymptotic frameworks that are compatible with our general triangular array framework for nonstationary processes. We discuss the simpler cases of iid sequences and stationary processes in Subsection A.3, where conventional long-run asymptotic arguments can be used.

First, we note that it is also possible to use infill asymptotics (as in Section 4) with the bounded variation-type nonstationarity from Assumption 3.6 as opposed to the Lipschitz-type nonstationarity for locally stationary time series from Assumption 4.6. For example, see Section 2 of Mies [Mie23]. However, there has been no work studying the theoretical guarantees for time-varying nonlinear regression estimators for nonstationary processes with bounded variation-type nonstationarity. We leave this for future work.

Second, our triangular array framework also allows for nonstationary processes that exhibit some form of repetition over time, such as periodic stationarity or cyclostationarity [Ben58; PP79; BHL94; Gar94; GNP06; Nap16b]. See Subsection 2.5.1 of Bonnerjee, Karmakar, and Wu [BKW24] for a discussion of how strong Gaussian approximations for nonstationary nonlinear processes with causal representations as in Subsections 3.1, 3.3, 3.4 can be used with cyclostationary processes. Also, see the discussion preceding Remark 2.1 in Chen, Smetanina, and Wu [CSW22] for more discussion about time-varying regression with periodic stationary or cyclostationary processes, which are not necessarily locally stationary. However, we would ideally like to be able to handle even more complex forms of nonstationarity than cyclostationarity; see Gardner, Napolitano, and Paura [GNP06] and Napolitano [Nap16a] for generalizations of this concept.

Third, simulation-and-regression approaches can be used with our CI testing framework. Suppose we have access to a black-box simulator which we can use to generate realistic paths of (Y, Z) or (X, Z). The main idea of this approach is to simulate s paths of the process, then fit separate regression models, such as XGBoost [CG16], LightGBM [Ke+17], or random forests [Bre01], for each time t by using data across the s iid simulated paths. Afterwards, the fitted regression models for each time t can be applied to the observed process. The asymptotic arguments can be based on letting the number of simulations s grow with n, where n can be linked to the sample size (e.g. sampling frequency and/or duration of time) and number of dimensions. We can also allow n to be linked to the quality of the simulations, so that as n grows we can generate more realistic simulations — perhaps at a higher computational cost — and the simulator can be seen as converging in some sense to the true data generating process. Note that CI tests which require multiple paths of a nonstationary process, such as [Man+24; LSP22; Liu+23], can be used if the simulator can generate paths for (X, Y, Z).

A.3 Simplifications under stationarity

Throughout this paper, we have completely avoided the assumption of stationarity. However, it is worth explaining how things would simplify if we are willing to assume that the time series are stationary. Overall, the takeaway is that the original GCM test from Shah and Peters [SP20] would require minimal modifications.

To begin, suppose we have $n \in \mathbb{N}$ observations of a stationary mixing time series, so that the regression functions are time-invariant. Further, suppose that the errors are iid. The statistical guarantees of many machine learning algorithms and statistical models, such as support vector machines [SA09; SHS09; HS14], random forests [Goe20; DN20], lasso [WLT20], and high-dimensional vector autoregressive models [WT23], have been studied in the context of stationary mixing time series with iid errors. Over the last decade, the literature on statistical learning theory for time series 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; HY19; MK20]. This literature has recently considered new notions of learnability for general non-iid stochastic processes [DT20] and conditions under which learning from general non-iid stochastic processes is possible [Han21].

Next, suppose that the regression functions are time-varying but have iid errors; this setting is studied by Zhang and Wu [ZW15]. Since the errors are iid, a multiplier bootstrap testing procedure can be justified by the Gaussian approximation from Chernozhukov, Chetverikov, and Kato [CCK13] which was used by Shah and Peters [SP20]. Hence, the resulting test would be very similar to the original GCM test for the iid setting from Shah and Peters [SP20]. The main difference is that there can be time-lagged conditional dependencies in the stationary time series setting.

Next, suppose that the observed processes are temporally dependent (e.g. some form of mixing) and stationary so that the regression functions are time-invariant as before, but the errors are also temporally dependent. The guarantees of the lasso and vector autoregressive models are fairly well-studied in this setting. Basu and Michailidis [BM15] investigated high-dimensional vector autoregressive models with serially correlated errors. Gupta [Gup12] and Xie and Xiao [XX18] studied the lasso with errors satisfying various weak dependence conditions. Peng, Zhu, and Zhong [PZZ23] and Xie, Xu, and Yang [XXY17] studied the lasso with ϕ-mixing and β-mixing errors, respectively. Wu and Wu [WW16] studied the guarantees of the lasso in the setting with temporally dependent errors by using the functional dependence measure of Wu [Wu05].

In the serially correlated error setting, the key difference with the GCM test from Shah and Peters [SP20] is that one must use a suitable Gaussian approximation result to justify a multiplier bootstrap-type testing procedure. See Chang, Chen, and Wu [CCW24] for a comprehensive overview of Gaussian approximations for dependent data. Chernozhukov, Chetverikov, and Kato [CCK16; CCK19] investigated a block multiplier bootstrap under a β-mixing assumption, and Zhang and Cheng [ZC14] explored a wild multiplier bootstrap under the functional dependence measure of Wu [Wu05]. Also, Zhang and Wu [ZW17] discuss estimators for the long-run covariance matrix so that their Gaussian approximation for high-dimensional time series can be applied in practice. See Wu and Xiao [WX12] and Wu [Wu11] for more discussion about long-run covariance matrix estimation for stationary time series.

One could also have time-invariant regression functions with errors that are nonstationary and temporally dependent. For instance, Xia, Chen, and Guo [XCG24] studies the lasso with locally stationary errors. However, the statistical guarantees of other machine learning algorithms and statistical models have not been studied in this setting. If the process of error products is mean-nonstationary (i.e. time-varying expected conditional covariance) under alternatives, then the same test statistics from Subsection 2.4 can be used. Otherwise, if domain knowledge suggests that the time-varying expected conditional covariances usually maintain the same sign, then the test statistics from Subsection A.1 can be used.

To recap, in this subsection we considered how the assumption of stationarity would vastly simplify the problem. As a result, CI testing for stationary time series would only require very minor modifications of the original GCM test for the iid setting from Shah and Peters [SP20]. In contrast, we consider the much more complicated setting in which the observed processes are nonstationary and temporally dependent, the regression functions vary over time, and the errors are nonstationary and serially correlated.

A.4 Local CI testing for locally stationary time series

In this subsection, we discuss three CI tests for locally stationary time series that we did not pursue in this paper. Crucially, the test statistics below require local long-run covariance estimation. Most local long-run covariance estimators use kernel smoothing, and therefore require selecting bandwidths. Unfortunately, test statistics that use kernel smoothing can be very sensitive to the choice of the bandwidths, which can be hard to select in practice. Inspired by the success of bandwidth-free approaches in other areas of time series analysis [Lob01; Sha10; RS13; Sha15; ZS24], we designed our dGCM test so that it does not require local long-run covariance estimation and therefore avoids kernel smoothing.

To begin, let us translate the "weak" conditional independence criterion of Daudin [Dau80] into the locally stationary time series setting as follows. For some $n \in \mathbb{N}$, $u \in \mathcal{U}_n$, $(i, j, a, b) \in \mathcal{D}_n$, $t \in \mathcal{T}_n$, if

$$\tilde{X}_{\lfloor un \rfloor, n, i, a}(u) \perp\!\!\!\perp \tilde{Y}_{\lfloor un \rfloor, n, j, b}(u) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor, n}(u),$$

then

$$\mathbb{E}_P[\phi(\tilde{X}_{\lfloor un \rfloor,n,i,a}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u))\varphi(\tilde{Y}_{\lfloor un \rfloor,n,j,b}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u))] = 0,$$

for all functions

$$\phi \in L^2_{\tilde{X}_{\lfloor un \rfloor,n,i,a}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)}, \quad \varphi \in L^2_{\tilde{Y}_{\lfloor un \rfloor,n,j,b}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)},$$

such that

$$\mathbb{E}_P[\phi(\tilde{X}_{\lfloor un \rfloor,n,i,a}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)] = 0,$$
$$\mathbb{E}_P[\varphi(\tilde{Y}_{\lfloor un \rfloor,n,j,b}(u), \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u)] = 0.$$

Hence, the corresponding *local* expected conditional covariance

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

is equal to zero for $m = (i, j, a, b) \in \mathcal{D}_n$.

First, consider the global null hypothesis of conditional independence

$$\tilde{X}_{\lfloor un \rfloor,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{\lfloor un \rfloor,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u) \text{ for all } u \in \mathcal{U}_n, \text{ for all } (i, j, a, b) \in \mathcal{D}_n. \tag{15}$$

In the univariate setting, $\mathcal{D}_n$ simply consists of one dimension/time-offset tuple as in Subsection 2.2. Also, note that this hypothesis can be extended to the group of time series setting as discussed in Subsection 2.2. Note that the null hypothesis (15) implies the null hypothesis (14), so 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 $(\boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$ will still have mean zero as in Section 4. To test for the null hypothesis (15), we could, for example, use the test statistic

$$\sup_{u\in\mathcal{U}_n} \left|\left| \frac{1}{\sqrt{T_n}} \sum_{t=\mathbb{T}_n}^{\lfloor un \rfloor} (\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n})^{-1/2} \hat{\boldsymbol{R}}_{t,n} \right|\right|_p,$$

based on some ℓ_p norm ($p \geq 2$) of the studentized partial sum process, where $\hat{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{t,n}$ is an estimate of the local long-run covariance matrix at time t. The theoretical guarantees for the test based on this test statistic would utilize the recent results from Mies [Mie24] about strong Gaussian approximations with random multipliers.

Second, it is possible to develop a test for the *local* null hypothesis of conditional independence

$$\tilde{X}_{\lfloor un \rfloor,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{\lfloor un \rfloor,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u) \text{ for all } (i, j, a, b) \in \mathcal{D}_n, \tag{16}$$

for a *particular* rescaled time $u \in \mathcal{U}_n$ (i.e. instead of for all $\mathcal{U}_n$ as in (15)) by using, for example, the test statistic

$$\max_{m=(i,j,a,b)\in\mathcal{D}_n} \left| \frac{1}{\sqrt{T_n h_{n,m}}} \sum_{t\in\mathcal{T}_n} K\left(\frac{t/n - u}{h_n}\right) \hat{R}_{t,n,m} \right| / \hat{\sigma}^R_{n,m}(u),$$

for some bandwidths $h_{n,m} \to 0$ and local long-run variance estimates $(\hat{\sigma}^R_{n,m}(u))^2$. The main idea for this local CI test is that since $\mathbb{E}_P(\tilde{R}_{P,\lfloor un \rfloor,n,m}(u)) = 0$ under the null and the process of error products is "approximately stationary" over short periods of time, we can expect that the means of $R_{P,t,n,m} = \tilde{R}_{P,t,n,m}(t/n)$ for rescaled times t/n near u are also close to zero under the null. We can make this mathematically precise by using the technical tools developed for nonlinear locally stationary time series from Dahlhaus, Richter, and Wu [DRW19].

Third, it is also possible to simultaneously test whether conditional independence

$$\tilde{X}_{\lfloor un \rfloor,n,i,a}(u) \perp\!\!\!\perp \tilde{Y}_{\lfloor un \rfloor,n,j,b}(u) \mid \tilde{\boldsymbol{Z}}_{\lfloor un \rfloor,n}(u) \text{ for all } (i, j, a, b) \in \mathcal{D}_n, \tag{17}$$

holds at *each* rescaled time $u \in \mathcal{U}_n$ (i.e. instead of for a particular $u \in \mathcal{U}_n$ as in (16)). This can be done by creating *simultaneous confidence bands* (i.e. over time) for expected conditional covariance curves $(\rho_{P,n,m}(u))_{u\in\mathcal{U}_n}$ for each $m = (i, j, a, b) \in \mathcal{D}_n$. Depending on whether or not estimates of the local long-run variances $(\hat{\sigma}^R_{n,m}(u))^2$ are used, these simultaneous confidence bands will have time-varying or time-invariant widths, respectively. The main idea is that the local null hypothesis of conditional

independence at rescaled time $u \in \mathcal{U}_n$ for some dimension/time-offset tuple $m = (i, j, a, b) \in \mathcal{D}_n$ can be rejected if zero is not included in the corresponding confidence interval for the local expected conditional covariance $\rho_{P,n,m}(u)$. 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 stronger assumptions (e.g. Donsker-type), data decomposition techniques (e.g. splitting, fission, or thinning) for nonstationary time series, or two independent realizations of the same nonstationary process — rarely possible outside of experimental settings.

An approach for inferring expected conditional covariance curves would have a range of applications outside of testing for conditional independence, since this functional frequently appears in the causal inference literature [Ken24; Rob+08; Rob+09; Li+11; Rob+17; Rob18]. We suspect that similar approaches can be used to infer curves based on other functionals of interest in causal inference. Hence, this line of work would be of significant interest to the emerging field of time series causal inference [Sag+20; RGR22; Run+23; Run+19a; Run18a]. Lastly, we note that it may be possible to extend the tests discussed in this subsection to the piecewise locally stationary setting (see Subsection A.5), however we leave the details for future work.

A.5 Piecewise locally stationary time series

We briefly describe how to extend the Sieve-dGCM test from Section 4 from *locally stationary time series* [Dah97; ZW09; Dah12; DRW19] to a more general class of nonstationary time series known as *piecewise locally stationary time series* (PLS) introduced in Zhou [Zho13]. Specifically, the class of PLS time series generalizes the stochastic Lipschitz condition for the nonstationarity from Assumption 4.6 by allowing for finitely many breakpoints; see Zhou [Zho13], Wu and Zhou [WZ19], and Dette, Wu, and Zhou [DWZ19]. We emphasize that PLS time series are included in the even more general class of time series from Section 3 with the "bounded variation nonstationarity" condition from Assumption 3.6.

The main idea is to identify the breakpoints, fit a separate sieve model on each locally stationary segment, and run Algorithm 1 on *all* the residuals. If the breakpoints are known exactly, then the same arguments can be used to show that the sieve estimator achieves the required convergence rates (i.e. within each locally stationary segment). If the breakpoints must be identified, then our arguments must be extended to account for this. As far as we know, Wu and Zhou [WZ19] is the most relevant work on identifying breakpoints for PLS time series. We leave the full details of this extension for exciting future work.

A.6 Weakening the temporal dependence assumptions for error processes

In Assumption 3.5, we assumed that there is an upper bound on the L^∞ version of the functional dependence measures for the error processes. We use this assumption to show (23) in the proof of Theorem 3.1. Afterwards, we use the time-uniform convergence rates for the time-varying regression estimators to show Step 1.2 in the proof of Theorem 3.1. It is possible to weaken the dependence assumptions imposed on the errors by making more complicated assumptions about the term $\mathbb{E}_P \, ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,t,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,t,n,h-1}||_2^2$ from (23), or stronger assumptions on the time-varying regression estimators. Instead, we opt for assumptions on the errors and prediction errors which are simpler and more transparent. Lastly, we note that the Sieve-dGCM test performs well in practice (see Section 5) even when the error processes violate this temporal dependence assumption used to prove Theorem 3.1.

B Distribution-Uniform Theory

In this section, we state distribution-uniform versions of the results from Mies and Steland [MS23]. All of the results in this section can be applied to general triangular array frameworks for high-dimensional nonstationary nonlinear time series, such as locally stationary time series.

B.1 Literature review of distribution-uniform inference

We briefly review prior work on distribution-uniform inference. First, we discuss the CI testing literature. Recently, there has been a great deal of work on distribution-uniform CI testing frameworks

due to the hardness result and CI testing framework from 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 in which n realizations of a point process are observed. Christgau, Petersen, and Hansen [CPH22] also introduce a distribution-uniform version 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. The work in Waudby-Smith and Ramdas [WR23] is motivated by prior work on asymptotic anytime-valid inference from Waudby-Smith et al. [Wau+24], in which the authors defined the concept of an "asymptotic confidence sequence". In particular, Waudby-Smith et al. [Wau+24] introduced asymptotic confidence sequences for iid random variables and a Lindeberg-type asymptotic confidence sequence which can capture time-varying means under martingale dependence.

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 [Li89; Kas18; Tib+18; RWG19; KBW23]. Also, there is a plethora of literature on distribution-uniform moment inequality testing [IM04; RS08; AG09; AS10; AB12; 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. In the supplemental appendix, Li, Liao, and Zhou [LLZ22] introduce a distribution-uniform strong Gaussian approximation for the full sum of a high-dimensional mixingale.

B.2 Distribution-uniform strong Gaussian approximation

To begin, 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 high-dimensional 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 $P \in \mathcal{P}_n$. Here $\mathcal{H}_t = (\eta_t, \eta_{t-1}, \ldots)$, where $(\eta_t)_{t \in \mathbb{Z}}$ is a sequence of iid random vectors with dimension $d^\eta = d_n^\eta$, and $G_{t,n} : (\mathbb{R}^{d^\eta})^\infty \to \mathbb{R}^{d_n}$ is a measurable function — where we endow $(\mathbb{R}^{d^\eta})^\infty$ with the σ-algebra generated by all finite projections. For each $t \in [n]$, $G_{t,n}(\mathcal{H}_s)$ 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.

For each $n \in \mathbb{N}$, write the $\mathbb{R}^{d_n}$-valued process of interest as $(W_{t,n})_{t \in [n]}$. We assume that for each $n \in \mathbb{N}$ and $t \in [n]$, the random vector $W_{t,n}$ has a causal representation; that is, it can be represented as a measurable function of these iid random vectors

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

Similarly, for the causal representations in the locally stationary time series setting, the measurable function $\tilde{G}_n(u, \cdot) : (\mathbb{R}^{d^\eta})^\infty \to \mathbb{R}^{d_n}$ is defined for each rescaled time $u \in [0, 1]$, and we assume that

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

We can use the results in this section for locally stationary time series by writing

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

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 $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P \in \mathcal{P}_n}$. 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 $P \in \mathcal{P}_n$.

Note that the causal representations in this paper use sequences of iid random vectors, whereas the causal representations in Mies and Steland [MS23] use sequences of iid Unif$[0,1]$ random variables. The same arguments used in Mies and Steland [MS23] can be applied when using our formulation of the causal representations with iid random vectors. The only reason we write the causal representations in this way is for the sake of clarity.

In fact, standard results in probability theory imply that the causal representations based on measurable functions of sequences of iid Unif$[0,1]$ random variables as in Mies and Steland [MS23] are already sufficiently general. For example, see Kallenberg [Kal21] Lemma 4.21, Lemma 4.22, and the surrounding discussion. More specifically, the causal representations with sequences of iid Unif$[0,1]$ random variables can express the causal representations with sequences of random vectors by including compositions with additional measurable functions for (1) replicating each of the iid Unif$[0,1]$ random variables, and (2) inverse sampling via products of conditional distributions; see Section 2.5 of Rubinstein and Kroese [RK16] on random vector generation.

Next, we define our measure of temporal dependence for the process. For the following definition, let $(\tilde{\eta}_t)_{t \in \mathbb{Z}}$ be an iid copy of $(\eta_t)_{t \in \mathbb{Z}}$ and denote

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

to be $\mathcal{H}_t$ with the j-th input in the past η_{t-j} replaced with the iid copy $\tilde{\eta}_{t-j}$.

Definition B.1 (Functional dependence measure)**.** *For $n \in \mathbb{N}$, $P \in \mathcal{P}_n$, $t \in \mathcal{T}_n$, define the functional dependence measure of $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ as*

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

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

We state the following distribution-uniform assumptions about the temporal dependence and nonstationarity of the process for some collections of distributions $\mathcal{P}_n$ for some $n \in \mathbb{N}$. Note that we write the time of the input sequence as 0 when it does not matter due to stationarity.

Assumption B.1 (Distribution-uniform decay of temporal dependence)**.** *We assume that there exist $\beta > 0$, $q \geq 2$ and a constant $\Theta_n > 0$, 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 B.2 (Distribution-uniform total variation condition for nonstationarity)**.** *Recall Θ_n from Assumption B.1. Assume that there exists some $\Gamma_n \geq 1$ such that*

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

Note that the assumptions regarding the temporal dependence and nonstationarity of the process of error products, as stated in Subsection 3.4, ensure that both Assumptions B.1 and B.2 hold for each $n \in \mathbb{N}$. Furthermore, since the assumptions in Subsection 4.5 are strictly stronger than those in Subsection 3.4, the results from this section can be applied to the process of error products in both Sections 3 and 4.

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. In general, the Gaussian approximation allows the dimensions to grow as $d_n = O(n^{\frac{1-\delta}{1+\frac{1}{2\xi(q,\beta)}}})$ for some $\delta > 0$. In the limiting case when $\beta \geq 3$ and $q \to \infty$, this corresponds to $d_n = O(n^{\frac{1}{4}-\delta'})$ for some $\delta' > 0$.

Let us briefly recall some notation used in the main text. Recall $\bar{\beta}^R > 3$, $\bar{q}^R > 4$ from Assumption 3.5, as well as the number of times T_n and the number of dimension/time-offset combinations D_n from Subsection 2.1. For the dGCM test, we allow $D_n = O(T_n^{r(\bar{q}^R, \bar{\beta}^R)})$ where

$$r(\bar{q}^R, \bar{\beta}^R) = \min\left(\frac{1-\delta}{1+\frac{1}{2\xi(\bar{q}^R,\bar{\beta}^R)}}, \frac{1}{6}\right), \tag{18}$$

for some $\delta > 0$. The limiting factor that leads to this requirement is *not* from the strong Gaussian approximation, but rather to ensure that the convergence rate requirements can be achieved by the time-varying nonparametric regression estimators.

The following result is a distribution-uniform version of the strong Gaussian approximation from Theorem 3.1 in Mies and Steland [MS23].

Lemma B.1. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$ for the stochastic system $(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}}$, suppose that Assumption B.1 is satisfied for $\mathcal{P}_n$ with some $q > 2$, $\beta > 1$ and constant $\Theta_n > 0$. Let the $\mathbb{R}^{d_n}$-valued process $(W_{t,n})_{t\in[n]}$ be defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ with $\mathbb{E}_P(W_{t,n}) = 0$ for each time $t \in [n]$ and distribution $P \in \mathcal{P}_n$. Also, suppose the dimension $d_n < cn$ for some constant $c > 0$. Then, on a potentially enriched collection of probability spaces $(\Omega', \mathcal{B}', \mathbb{P}'_P)_{P\in\mathcal{P}_n}$, there exist random vectors $(W'_{t,n})_{t\in[n]}$ with the same distribution as $(W_{t,n})_{t\in[n]}$ for each $P \in \mathcal{P}_n$, and independent Gaussian random vectors $(V'_{t,n})_{t\in[n]}$ with $\mathbb{E}_P(V'_{t,n}) = 0$ for each $t \in [n]$, $P \in \mathcal{P}_n$, such that*

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

for some universal constant K 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 $t \in [n]$, $P \in \mathcal{P}_n$ by Lemma B.5. If Assumption B.2 is also satisfied for $\mathcal{P}_n$, then on $(\Omega', \mathcal{B}', \mathbb{P}'_P)_{P\in\mathcal{P}_n}$ there exist random vectors $(W'_{t,n})_{t\in[n]}$ which have the same distribution as $(W_{t,n})_{t\in[n]}$ for each $P \in \mathcal{P}_n$, and independent Gaussian random vectors $(V^*_{t,n})_{t\in[n]}$ where $V^*_{t,n} \sim \mathcal{N}(0, \Sigma_{P,t,n})$ for each $t \in [n]$, $P \in \mathcal{P}_n$, such that

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

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

Proof of Lemma B.1: Assumptions B.1 and B.2 are distribution-uniform versions of conditions (G.1) and (G.2) from Mies and Steland [MS23]. Hence, under the assumptions of the Lemma related to Assumptions B.1 and B.2, the distribution-pointwise inequalities from Theorem 3.1 in Mies and Steland [MS23] hold for each $P \in \mathcal{P}_n$. Since the suprema over all distributions in the collection $\mathcal{P}_n$ of the upper bounds are finite, the distribution-uniform inequalities from the Lemma hold for $\mathcal{P}_n$ by basic properties of the supremum. □

Recently, Bonnerjee, Karmakar, and Wu [BKW24] introduced univariate strong Gaussian approximation results with optimal rates and explicit constructions, building on prior work by Karmakar and Wu [KW20]. We emphasize that the distribution-uniform strong Gaussian approximation for high-dimensional nonstationary nonlinear time series from Lemma B.1 does *not* achieve this optimal rate. However, the convergence rates for the prediction errors from the estimation of the time-varying regression functions dominate the strong Gaussian approximation rates. Therefore, we do not "lose anything" by using Lemma B.1 in our regression-based CI test instead of a distribution-uniform version of the strong Gaussian approximation from Bonnerjee, Karmakar, and Wu [BKW24], as our main results would not change in any meaningful way.

The following result is a distribution-uniform version of Theorem 3.2 from Mies and Steland [MS23].

Lemma B.2. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$ for the stochastic system $(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}}$, let the $\mathbb{R}^{d_n}$-valued process $(W_{t,n})_{t\in[n]}$ be defined on the collection of probability spaces $(\Omega,\mathcal{B},\mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ with $W_{t,n} \in L_q(P)$ and $\theta_{P,t,n}(j,q,r)$ as in Definition B.1 for each $P \in \mathcal{P}_n$ and some $2 \leq r \leq q < \infty$. There exists a universal constant $K = K(q,r)$ such that*

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

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

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

Proof of Lemma B.2: Under the assumptions of the Lemma, the distribution-pointwise inequalities from Theorem 3.2 in Mies and Steland [MS23] hold for each $P \in \mathcal{P}_n$. Since the suprema over all distributions in the collection $\mathcal{P}_n$ 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. □

B.3 Distribution-uniform feasible Gaussian approximation

In this subsection, we introduce distribution-uniform versions of Theorem 4.1 and Proposition 4.2 from Mies and Steland [MS23] so that the distribution-uniform strong Gaussian approximation from Subsection B.2 can be used for statistical inference. The key is a distribution-uniform cumulative covariance estimator $\hat{Q}_{k,n}$ of the cumulative covariance matrices $Q_{P,k,n} = \sum_{t=1}^{k}\Sigma_{P,t,n}$ 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 [MS23], namely

$$\hat{Q}_{k,n} = \sum_{r=L_n}^{k}\frac{1}{L_n}\left(\sum_{s=r-L_n+1}^{r}W_{s,n}\right)^{\otimes 2}$$

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

The following result is a distribution-uniform version of Theorem 4.1 from Mies and Steland [MS23].

Lemma B.3. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$ for the stochastic system $(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}}$, let the $\mathbb{R}^{d_n}$-valued process $(W_{t,n})_{t\in[n]}$ be defined on the collection of probability spaces $(\Omega,\mathcal{B},\mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that $W_{t,n} = G_{t,n}(\mathcal{H}_t)$ and Assumptions B.1 and B.2 are satisfied for $\mathcal{P}_n$ with $q \geq 4$ and $\beta > 2$. Then*

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

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

Proof of Lemma B.3: Assumptions B.1 and B.2 are distribution-uniform versions of conditions (G.1) and (G.2) from Mies and Steland [MS23]. Hence, under the assumptions of the Lemma related to Assumptions B.1 and B.2, the distribution-pointwise inequalities from Theorem 4.1 in Mies and Steland [MS23] hold for each $P \in \mathcal{P}_n$. Since the supremum over all distributions in the collection $\mathcal{P}_n$

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 next result is a distribution-uniform version of Proposition 4.2 from Mies and Steland [MS23].

Lemma B.4. *For some sample size $n \in \mathbb{N}$, let $\mathcal{P}_n$ be a collection of distributions for the stochastic system $(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}}$. Let $\Sigma_{P,t,n}, \Sigma'_{P,t,n} \in \mathbb{R}^{d_n\times d_n}$ be symmetric, positive definite matrices for each $t \in [n]$, $P \in \mathcal{P}_n$, and let $(V_{t,n})_{t\in[n]}$ be independent random vectors defined on the collection of probability spaces $(\Omega, \mathcal{B}, \mathbb{P}_P)_{P\in\mathcal{P}_n}$ so that $V_{t,n} \sim \mathcal{N}(0, \Sigma_{P,t,n})$ for each $t \in [n]$, $P \in \mathcal{P}_n$. Then, on a potentially enriched collection of probability spaces $(\Omega', \mathcal{B}', \mathbb{P}'_P)_{P\in\mathcal{P}_n}$, there exist independent random vectors $(V'_{t,n})_{t\in[n]}$ with $V'_{t,n} \sim \mathcal{N}(0, \Sigma'_{P,t,n})$ for each $t \in [n]$, $P \in \mathcal{P}_n$ such that*

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

where

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

Proof of Lemma B.4: The distribution-pointwise inequalities from Proposition 4.2 in Mies and Steland [MS23] hold for each $P \in \mathcal{P}_n$. Since the supremum over all distributions in the collection $\mathcal{P}_n$ 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. □

B.4 Auxiliary Lemmas

The following result is a distribution-uniform version of Proposition 5.4 from Mies and Steland [MS23].

Lemma B.5. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$ for the stochastic system $(G_{t,n}(\mathcal{H}_s))_{t\in[n],s\in\mathbb{Z}}$, let Assumption B.1 be satisfied for $\mathcal{P}_n$ with some $q \geq 2$, $\beta > 0$, and constant $\Theta_n > 0$. Denote*

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

Then for all $t \in [n]$, $h \in \mathbb{Z}$, we have

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

where $||\cdot||_{\text{tr}}$ 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 $t \in [n]$, $P \in \mathcal{P}_n$.

Proof of Lemma B.5: The distribution-pointwise inequality from Proposition 5.4 in Mies and Steland [MS23] holds for each $P \in \mathcal{P}_n$. Since the supremum over all distributions in the collection $\mathcal{P}_n$ 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 result is a distribution-uniform version of the Rosenthal inequality from the first part of Theorem 5.6 from Mies and Steland [MS23].

Lemma B.6. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$, let $(M_{t,n})_{t\in[n]}$ be a $\mathbb{R}^{d_n}$-valued martingale-difference sequence with distribution determined by $P \in \mathcal{P}_n$. For each $2 \leq r \leq q < \infty$, there exists a finite factor $C_{q,r}$ such that for any $n, d_n \in \mathbb{N}$, we have*

$$\sup_{P\in\mathcal{P}_n}\left(\mathbb{E}_P \max_{k\leq n}\left|\left|\sum_{t=1}^{k} M_{t,n}\right|\right|_r^q\right)^{\frac{1}{q}} \leq C_{q,r}n^{\frac{1}{2}-\frac{1}{q}}\sup_{P\in\mathcal{P}_n}\left(\sum_{t=1}^{n}\mathbb{E}_P||M_{t,n}||_r^q\right)^{\frac{1}{q}}$$
$$\leq C_{q,r}n^{\frac{1}{2}}\sup_{P\in\mathcal{P}_n}\left(\max_{t\leq n}\left(\mathbb{E}_P||M_{t,n}||_r^q\right)^{\frac{1}{q}}\right).$$

Proof of Lemma B.6: The distribution-pointwise inequalities from the first part of Theorem 5.6 in Mies and Steland [MS23] hold for each $P \in \mathcal{P}_n$. Since the suprema over all distributions in the collection $\mathcal{P}_n$ of the upper bounds are always finite, the distribution-uniform inequality from the Lemma holds for $\mathcal{P}_n$ by basic properties of the supremum. □

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

Lemma B.7. *For some sample size $n \in \mathbb{N}$ and collection of distributions $\mathcal{P}_n$, let X_n be a generic 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 $K > 0$, and suppose that $|X_n| \leq K$ for all $n \in \mathbb{N}$ and $X_n = o_{\mathcal{P}}(1)$. Then we have*

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_P(|X_n|) = o(1).$$

Proof of Lemma B.7: For any given $\epsilon > 0$,

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

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

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

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

C Proofs

We will denote the three bias terms by

$$\begin{aligned}
\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,t,n} &= (\hat{w}^{f,g}_{P,t,n,m})_{m \in \mathcal{D}_n} = (\hat{w}^{f}_{P,t,n,i,a} \hat{w}^{g}_{P,t,n,j,b})_{m \in \mathcal{D}_n}, \\
\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} &= (\hat{w}^{g,\varepsilon}_{P,t,n,m})_{m \in \mathcal{D}_n} = (\hat{w}^{g}_{P,t,n,j,b} \varepsilon_{P,t,n,i,a})_{m \in \mathcal{D}_n}, \\
\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,t,n} &= (\hat{w}^{f,\xi}_{P,t,n,m})_{m \in \mathcal{D}_n} = (\hat{w}^{f}_{P,t,n,i,a} \xi_{P,t,n,j,b})_{m \in \mathcal{D}_n},
\end{aligned}$$

where $m = (i, j, a, b) \in \mathcal{D}_n$. Also, denote

$$\begin{aligned}
\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,n} &= (\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,t,n})_{t \in \mathcal{T}_{n,L}}, \\
\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,n} &= (\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n})_{t \in \mathcal{T}_{n,L}}, \\
\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,n} &= (\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,t,n})_{t \in \mathcal{T}_{n,L}}.
\end{aligned}$$

Note that in this proof when we write $o_{\mathcal{P}}(\cdot)$ and $O_{\mathcal{P}}(\cdot)$, we will always be doing so with reference to the collection of distributions $\mathcal{P}^*_{0,n}$ defined in the statement of the theorem.

C.1 Proof of Theorem 3.1

Step 1 (Bias Terms): First, we decompose the products of residuals into the products of errors and the three bias terms, then apply the triangle inequality and subadditivity, yielding

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\hat{\boldsymbol{R}}_n) > \hat{q}_{1-\alpha+\nu_n} + \tau_n) \\
&\leq \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}_{P,n}) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{2}) \\
&+ \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,n}) > \frac{\tau_n}{6}) \\
&+ \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,n}) > \frac{\tau_n}{6}) \\
&+ \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,n}) > \frac{\tau_n}{6}).
\end{aligned}$$

We will handle each of the three bias terms separately.

Step 1.1: Observe that for any $\delta > 0$, we have

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(\tau_n^{-1} S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,n}) > \delta) \\
&\overset{(1)}{\leq} \delta^{-1}\tau_n^{-1} T_{n,L}^{-\frac{1}{2}} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{E}_P \left(\max_{s \in \mathcal{T}_{n,L}} \left\| \sum_{t \leq s} \hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,t,n} \right\|_2 \right) \\
&\overset{(2)}{\leq} \delta^{-1}\tau_n^{-1} T_{n,L}^{-\frac{1}{2}} D_n \sup_{P \in \mathcal{P}^*_{0,n}} \max_{(i,j,a,b) \in \mathcal{D}_n} \mathbb{E}_P \left(\sum_{t \in \mathcal{T}_{n,L}} |\hat{w}^f_{P,t,n,i,a}||\hat{w}^g_{P,t,n,j,b}| \right) \\
&\overset{(3)}{\leq} \delta^{-1}\tau_n^{-1} T_{n,L}^{\frac{1}{2}} D_n \sup_{P \in \mathcal{P}^*_{0,n}} \max_{(i,j,a,b) \in \mathcal{D}_n} \mathbb{E}_P \left(|\hat{w}^f_{P,t,n,i,a}|^2 \right)^{\frac{1}{2}} \mathbb{E}_P \left(|\hat{w}^g_{P,t,n,j,b}|^2 \right)^{\frac{1}{2}} \\
&\overset{(4)}{=} o(1),
\end{aligned}$$

where the previous lines follow by (1) Markov's inequality and ℓ_p-norm inequalities, (2) the triangle inequality, ℓ_p-norm inequalities, linearity of expectation, (3) linearity of expectation and the Cauchy-Schwarz inequality, (4) the rate requirements for the time-varying regression estimators.

Step 1.2:
Observe that for any $\delta > 0$, we have

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(\tau_n^{-1} S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,n}) > \delta) \\
&\overset{(1)}{\leq} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(\tau_n^{-2} \max_{s \in \mathcal{T}_{n,L}} \left\| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \leq s} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} \right\|_2^2 \geq \delta^2 \right) \\
&\overset{(2)}{\leq} \delta^{-2}\tau_n^{-2} T_{n,L}^{-1} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{E}_P \left(\max_{s \in \mathcal{T}_{n,L}} \left\| \sum_{t \leq s} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} \right\|_2^2 \right) \\
&\overset{(3)}{\leq} \delta^{-2}\tau_n^{-2} T_{n,L}^{-1} (\bar{K} T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} \sup_{P \in \mathcal{P}^*_{0,n}} \max_{t \in \mathcal{T}_{n,L}} \max_{j \in [d_Y]} \max_{b \in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)^{\frac{1}{2}})^2 \\
&\overset{(4)}{\leq} \delta^{-2}\tau_n^{-2} \bar{K}^2 D_n \sup_{P \in \mathcal{P}^*_{0,n}} \max_{t \in \mathcal{T}_{n,L}} \max_{j \in [d_Y]} \max_{b \in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2) \\
&\overset{(5)}{=} o(1),
\end{aligned}$$

where the previous lines follow by (1) the assumption about the form of the test statistic and squaring, (2) Markov's inequality and linearity of expectation, (3) for some constant $\bar{K} > 0$ by the arguments below, (4) simplifying the expression, and (5) the rate requirements for the time-varying regression estimator.

The following arguments are to show (3). These arguments are based on the constructions used in the proof of Theorem 3.2 in Mies and Steland [MS23], which build on the proof techniques from Theorem 1 in Liu, Xiao, and Wu [LXW13]. For each $t \in \mathcal{T}_{n,L}$ and $h \in \mathbb{N}_0$, let

$$\mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t,h} = \sigma(\eta^{\boldsymbol{\varepsilon}}_t, \eta^{\boldsymbol{\varepsilon}}_{t-1} \ldots, \eta^{\boldsymbol{\varepsilon}}_{t-h}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t),$$

where the input $\eta^{\boldsymbol{\varepsilon}}_t$ is from (11) and the input sequence $\mathcal{H}^{\hat{\boldsymbol{g}}}_t$ is from (10).

For each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $t \in \mathcal{T}_{n,L}$, and $h \in \mathbb{N}_0$ let

$$\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} = \sum_{k \leq t} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,k,n,h},$$
$$\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} = \mathbb{E}_P\left(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} | \mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t,h}\right),$$
$$\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,-1} = \mathbb{E}_P\left(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} | \mathcal{H}^{\hat{\boldsymbol{g}}}_{t}\right) = \boldsymbol{0},$$

almost surely, because for each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $(i,j,a,b) \in \mathcal{D}_n$, $t \in \mathcal{T}_{n,L}$ we have

$$\mathbb{E}_P(\hat{w}^g_{P,t,n,j,b}\varepsilon_{P,t,n,i,a} | \mathcal{H}^{\hat{\boldsymbol{g}}}_t) = \hat{w}^g_{P,t,n,j,b}\mathbb{E}_P(\varepsilon_{P,t,n,i,a} | \mathcal{H}^{\hat{\boldsymbol{g}}}_t) = 0, \tag{19}$$

almost surely, by Assumptions 3.3 and 3.4.

For each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $t \in \mathcal{T}_{n,L}$, and $h \in \mathbb{N}_0$ we have that

$$\mathbb{E}_P(||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h}||_2^2) < \infty, \tag{20}$$

by linearity of expectation, the contraction property of conditional expectation, and Assumption 3.3. For each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $t \in \mathcal{T}_{n,L}$, and $h \in \mathbb{N}_0$, by the tower property of conditional expectation, we have

$$\mathbb{E}_P(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h+1} | \mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t,h}) = \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h},$$

almost surely. Hence, for each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, and $t \in \mathcal{T}_{n,L}$, $(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h})_{h=0}^{\infty}$ is a martingale with respect to the filtration $(\mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{t,h})_{h=0}^{\infty}$. The martingale convergence theorem (see e.g. Theorem 1.5 of [Pis16]) ensures that for each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $t \in \mathcal{T}_{n,L}$ there exists some random vector $\tilde{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n}$ such that $\mathbb{E}_P||\tilde{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h}||_2^2 \to 0$ as $h \to \infty$. The measurability of $\boldsymbol{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{P,t,n}$ with respect to the projection σ-algebra, in view of Assumptions 3.1, 3.2, 3.3, 3.4, ensures that $\tilde{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} = \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n}$. Thus, for each $t \in \mathcal{T}_{n,L}$ we have

$$\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n} = \sum_{k \leq t} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,k,n} = \sum_{h=0}^{\infty} (\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}), \tag{21}$$

by telescoping.

For each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, and $h \in \mathbb{N}_0$,

$$(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n,h-1})_{k=0}^{\mathbb{T}^+_n - \mathbb{T}^-_n - L_n},$$

are martingale differences with respect to the filtration $(\mathcal{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n,k,h})_{k=0}^{\mathbb{T}^+_n - \mathbb{T}^-_n - L_n}$, where

$$\mathcal{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n,k,h} = \sigma(\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n - k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n - k - h}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n - k - h + 1} \ldots),$$

because for any $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $h \in \mathbb{N}_0$ and $k = 0, 1, \ldots,$ we have

$$\begin{aligned}
&\mathbb{E}_P(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n,h-1} | \mathcal{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n,k-1,h}) \\
&= \mathbb{E}_P\left(\mathbb{E}_P\left(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n} | \mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n - k,h}\right) | \mathcal{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n,k-1,h}\right) \\
&- \mathbb{E}_P\left(\mathbb{E}_P\left(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n - k,n} | \mathcal{F}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n - k,h-1}\right) | \mathcal{G}^{\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}}_{\mathbb{T}^+_n,k-1,h}\right) \\
&= \boldsymbol{0},
\end{aligned}$$

almost surely, because for each $n \in \mathbb{N}$, $P \in \mathcal{P}^*_{0,n}$, $(i,j,a,b) \in \mathcal{D}_n$, $h \in \mathbb{N}_0$ and $k = 0, 1, \ldots,$ we have

$$
\begin{aligned}
&\mathbb{E}_P(\mathbb{E}_P\left(\hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots)\\
&- \mathbb{E}_P(\mathbb{E}_P\left(\hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots)\\
&\overset{(1)}{=} \hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\mathbb{E}_P(\mathbb{E}_P\left(\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k})\\
&- \hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\mathbb{E}_P(\mathbb{E}_P\left(\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k})\\
&\overset{(2)}{=} \hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\mathbb{E}_P\left(\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&- \hat{w}^g_{P,\mathbb{T}^+_n-k,n,j,b}\mathbb{E}_P\left(\varepsilon_{P,\mathbb{T}^+_n-k,n,i,a}|\mathcal{H}^{\hat{\boldsymbol{w}}^{\boldsymbol{g}}}_{\mathbb{T}^+_n-k}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+1}, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k-h+2}, \ldots, \eta^{\boldsymbol{\varepsilon}}_{\mathbb{T}^+_n-k}\right)\\
&\overset{(3)}{=} 0,
\end{aligned}
$$

almost surely, by (1) Assumption 3.3, (2) the tower property and measurability, and (3) subtraction. Also, $\mathbb{E}_P(||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n-k,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n-k,n,h-1}||^2_2) < \infty$ by the triangle inequality, squaring, linearity of expectation, the Cauchy-Schwarz inequality, and the same arguments as (20) (i.e. linearity of expectation, the contraction property of conditional expectation, and Assumption 3.3).

For each $n \in \mathbb{N}$ and $h \in \mathbb{N}_0$, we have

$$
\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P\max_{t\in\mathcal{T}_{n,L}}||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||^2_2\right)^{\frac{1}{2}} \qquad (22)\\
&\overset{(1)}{\leq} \sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h-1}||^2_2\right)^{\frac{1}{2}}\\
&+ \sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P\max_{t\in\mathcal{T}_{n,L}}||(\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h-1}) - (\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1})||^2_2\right)^{\frac{1}{2}}\\
&\overset{(2)}{\leq} \sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h-1}||^2_2\right)^{\frac{1}{2}}\\
&+ \sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P\max_{\ell=0,\ldots,\mathbb{T}^+_n-\mathbb{T}^-_n-L_n}||\sum_{k=0}^{\ell}(\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n-k,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n-k,n,h-1})||^2_2\right)^{\frac{1}{2}}\\
&\overset{(3)}{\leq} 3\sup_{P\in\mathcal{P}^*_{0,n}}\left(\mathbb{E}_P||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h-1}||^2_2\right)^{\frac{1}{2}}\\
&\overset{(4)}{\leq} K\sup_{P\in\mathcal{P}^*_{0,n}}\left(\sum_{t\in\mathcal{T}_{n,L}}\mathbb{E}_P||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||^2_2\right)^{\frac{1}{2}},
\end{aligned}
$$

by (1) adding and subtracting $\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,\mathbb{T}^+_n,n,h-1}$ and the triangle inequality, (2) including the "last" term in this reversed partial sum and rewriting as the corresponding martingale, (3) Doob's maximal inequality (see e.g. Theorem 1.9 of [Pis16]), and (4) upper bounding by max of partial sums and applying Lemma B.6 with the finite constant $K/3 > 0$.

Next, observe that for $h = 1, 2, \ldots$ we have

$$
\begin{aligned}
&\mathbb{E}_P \; ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||_2^2 \\
&\overset{(1)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\mathbb{E}_P(\hat{w}^g_{P,t,n,j,b}\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \ldots, \eta^{\boldsymbol{\varepsilon}}_{t-h}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t) \\
&- \mathbb{E}_P(\hat{w}^g_{P,t,n,j,b}\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \ldots, \eta^{\boldsymbol{\varepsilon}}_{t-h+1}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t)|^2) \\
&\overset{(2)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}(\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \ldots, \eta^{\boldsymbol{\varepsilon}}_{t-h}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t) \\
&- \mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \ldots, \eta^{\boldsymbol{\varepsilon}}_{t-h+1}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t))|^2) \\
&\overset{(3)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}\mathbb{E}_P[(\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t) \\
&- \mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h+1,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t))|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t]|^2) \\
&\overset{(4)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}\mathbb{E}_P[(\mathbb{E}_P(G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{t,a})|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t) \\
&- \mathbb{E}_P(G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{t,a})|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h+1,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t))|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t]|^2) \\
&\overset{(5)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}\mathbb{E}_P[(\mathbb{E}_P(G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{t,a})|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t) \\
&- \mathbb{E}_P(G^{\varepsilon}_{P,t,n,i,a}(\tilde{\mathcal{H}}^{\varepsilon}_{t,a,h})|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t))|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t]|^2) \\
&\overset{(6)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}\mathbb{E}_P[G^{\varepsilon}_{P,t,n,i,a}(\mathcal{H}^{\varepsilon}_{t,a}) \\
&- G^{\varepsilon}_{P,t,n,i,a}(\tilde{\mathcal{H}}^{\varepsilon}_{t,a,h})|\eta^{\varepsilon}_{t,a}, \ldots, \eta^{\varepsilon}_{t-h,a}, \mathcal{H}^{\hat{\boldsymbol{g}}}_t]|^2) \\
&\overset{(7)}{\leq} D_n \max_{(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\theta^{\varepsilon,\infty}_{P,t,n,i,a}(h))^2 \\
&\overset{(8)}{\leq} D_n \max_{(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\bar{\Theta}^{\infty}(h \vee 1)^{-\bar{\beta}^{\infty}})^2,
\end{aligned}
$$

by (1) rewriting the expression, (2) Assumption 3.3, (3) measurability of the conditional expectations and the linearity property of conditional expectation, (4) the causal representation from Assumption 3.4, (5) replacing $\eta^{\varepsilon}_{t-h,a}$ with the iid copy $\tilde{\eta}^{\varepsilon}_{t-h,a}$, (6) measurability and linearity of the conditional expectations, and (7) Hölder's inequality, contraction property of conditional expectation, rewriting as the functional dependence measure from Definition 3.1, and upper bounding by the sum by D_n times the maximum over the dimension/time-offset combinations in $\mathcal{D}_n$, and (8) the upper bound on the functional dependence measure from Assumption 3.5.

Similarly, for $h = 0$, we have

$$
\begin{aligned}
&\mathbb{E}_P \; ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,0} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,-1}||_2^2 \\
&\overset{(1)}{=} \mathbb{E}_P \; ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,0}||_2^2 \\
&\overset{(2)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\mathbb{E}_P(\hat{w}^g_{P,t,n,j,b}\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \mathcal{H}^{\hat{\boldsymbol{g}}}_t)|^2) \\
&\overset{(3)}{=} \sum_{m=(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}\mathbb{E}_P(\varepsilon_{P,t,n,i,a}|\eta^{\boldsymbol{\varepsilon}}_t, \mathcal{H}^{\hat{\boldsymbol{g}}}_t)|^2) \\
&\overset{(4)}{\leq} D_n \max_{(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\bar{\Theta}^{\infty})^2,
\end{aligned}
$$

because (1) $\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,-1} = 0$, (2) rewriting the expression, (3) Assumption 3.3, and (4) Hölder's inequality, contraction property of conditional expectation, applying the upper bound on the L^{∞} norm from Assumption 3.5, and upper bounding by the sum by D_n times the maximum over the dimension/time-

offset combinations in $\mathcal{D}_n$. Hence, for all $h \in \mathbb{N}_0$ we have

$$\mathbb{E}_P \, ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||_2^2 \leq D_n \max_{(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\bar{\Theta}^\infty(h \vee 1)^{-\bar{\beta}^\infty})^2. \tag{23}$$

Summing over $h \in \mathbb{N}_0$, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}} \left(\mathbb{E}_P \max_{t\in\mathcal{T}_{n,L}} ||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n}||_2^2\right)^{\frac{1}{2}} \\
&\overset{(1)}{\leq} \sum_{h=0}^{\infty} \sup_{P\in\mathcal{P}^*_{0,n}} \left(\mathbb{E}_P \max_{t\in\mathcal{T}_{n,L}} ||\hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{S}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||_2^2\right)^{\frac{1}{2}} \\
&\overset{(2)}{\leq} \sum_{h=0}^{\infty} K \sup_{P\in\mathcal{P}^*_{0,n}} \left(\sum_{t\in\mathcal{T}_{n,L}} \mathbb{E}_P \, ||\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h} - \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\varepsilon}}_{P,t,n,h-1}||_2^2\right)^{\frac{1}{2}} \\
&\overset{(3)}{\leq} \sum_{h=0}^{\infty} K \sup_{P\in\mathcal{P}^*_{0,n}} \left(\sum_{t\in\mathcal{T}_{n,L}} D_n \max_{j\in[d_Y]} \max_{b\in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\bar{\Theta}^\infty(h \vee 1)^{-\bar{\beta}^\infty})^2\right)^{\frac{1}{2}} \\
&\overset{(4)}{\leq} \sum_{h=0}^{\infty} K \sup_{P\in\mathcal{P}^*_{0,n}} \left(T_{n,L} D_n \max_{t\in\mathcal{T}_{n,L}} \max_{j\in[d_Y]} \max_{b\in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)(\bar{\Theta}^\infty(h \vee 1)^{-\bar{\beta}^\infty})^2\right)^{\frac{1}{2}} \\
&\overset{(5)}{\leq} \bar{\Theta}^\infty K T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} \sup_{P\in\mathcal{P}^*_{0,n}} \max_{t\in\mathcal{T}_{n,L}} \max_{j\in[d_Y]} \max_{b\in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)^{\frac{1}{2}} \sum_{h=0}^{\infty} (h \vee 1)^{-\bar{\beta}^\infty} \\
&\overset{(6)}{\leq} \bar{\Theta}^\infty \bar{K}^\infty K T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} \sup_{P\in\mathcal{P}^*_{0,n}} \max_{t\in\mathcal{T}_{n,L}} \max_{j\in[d_Y]} \max_{b\in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)^{\frac{1}{2}} \\
&\overset{(7)}{\leq} \bar{K} T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} \sup_{P\in\mathcal{P}^*_{0,n}} \max_{t\in\mathcal{T}_{n,L}} \max_{j\in[d_Y]} \max_{b\in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2)^{\frac{1}{2}},
\end{aligned}$$

by (1) telescoping from (21) and the triangle inequality, (2) inequality (22), (3) inequality (23), (4) upper bounding each term by the maximum over time t, (5) simplifying the expression, (6) writing $\bar{K}^\infty = \sum_{h=0}^{\infty}(h \vee 1)^{-\bar{\beta}^\infty} < \infty$ since $\bar{\beta}^\infty > 1$ by upper Assumption 3.5, and (7) grouping together the positive constants into $\bar{K} = \bar{\Theta}^\infty \bar{K}^\infty K < \infty$.

Step 1.3: The same arguments as **Step 1.2** (i.e. exchanging g, ε with f, ξ) can be used to show that for $n \in \mathbb{N}$ and $\delta > 0$ we have

$$\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(\tau_n^{-1} S_{n,p}(\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,n}) > \delta) = o(1).$$

Next, we turn to the products of errors $(\boldsymbol{R}_{P,t,n})_{t\in\mathcal{T}_{n,L}}$.

Step 2 (Strong Gaussian Approximation): Denote the Gaussian random vectors associated with the strong Gaussian approximation of the product of errors by $\boldsymbol{R}^\dagger_{t,n} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n})$ for $t \in \mathcal{T}_{n,L}$.

Observe that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}_{P,n}) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{2}) \\
&\overset{(1)}{\leq} \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&+ \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P\left(\max_{s\in\mathcal{T}_{n,L}} \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} (\boldsymbol{R}_{P,t,n} - \boldsymbol{R}^\dagger_{t,n}) \right|\right|_2 > \frac{\tau_n}{4}\right) \\
&\overset{(2)}{\leq} \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&+ 4\tau_n^{-1} \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P\left(\max_{s\in\mathcal{T}_{n,L}} \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t\leq s} (\boldsymbol{R}_{P,t,n} - \boldsymbol{R}^\dagger_{t,n}) \right|\right|_2\right) \\
&\overset{(3)}{\leq} \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&+ 4\tau_n^{-1} K D_n^{\frac{1}{2}} \bar{\Theta}^R (\bar{\Gamma}^R_n)^{\frac{1}{2}\frac{\bar{\beta}^R-2}{\bar{\beta}^R-1}} \sqrt{\log(T_{n,L})} \left(\frac{D_n}{T_{n,L}}\right)^{\xi(\bar{q}^R,\bar{\beta}^R)},
\end{aligned}$$

where (1) follows from the triangle inequality, subadditivity, and the assumption about the form of the test statistic, (2) follows by Markov's inequality, and (3) follows by the distribution-uniform strong Gaussian approximation for high-dimensional nonstationary time series from Lemma B.1.

By subadditivity and monotonicity, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&\leq \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > q_{1-\alpha}) \\
&+ \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(q_{1-\alpha} > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&= \alpha + \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(q_{1-\alpha} > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}).
\end{aligned}$$

Step 3 (Covariance Approximation): Now, we focus on upper bounding

$$\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P(q_{1-\alpha} > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}).$$

Step 3.1: Let us reflect on the implications of Proposition 4.2 of Mies and Steland [MS23], which is the distribution-pointwise version of our distribution-uniform result in Lemma B.4 (i.e. just consider the collection of distributions consisting of a single distribution). Proposition 4.2 states that for each $n \in \mathbb{N}$ and $P \in \mathcal{P}^*_{0,n}$, for some cumulative covariance $\bar{Q}^{\boldsymbol{R}}_{P,n}$, there exist *independent* Gaussian random vectors $\bar{\boldsymbol{R}}_{t,n} \sim \mathcal{N}(0, \bar{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{P,t,n})$ for $t \in \mathcal{T}_{n,L}$ with $\bar{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{P,t,n} = \bar{Q}^{\boldsymbol{R}}_{P,t,n} - \bar{Q}^{\boldsymbol{R}}_{P,t-1,n}$ that are coupled with the Gaussian random vectors from the strong Gaussian approximation of the product of errors $\boldsymbol{R}^\dagger_{t,n} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n})$ for $t \in \mathcal{T}_{n,L}$, such that

$$\mathbb{E}_P \max_{k\in\mathcal{T}_{n,L}} \left|\left| \sum_{t\leq k} \boldsymbol{R}^\dagger_{t,n} - \sum_{t\leq k} \bar{\boldsymbol{R}}_{t,n} \right|\right|_2^2 \leq K \log(T_{n,L}) \left[\sqrt{T_{n,L}\bar{\delta}_{P,n}\rho_{P,n}} + \rho_{P,n}\right] = \bar{\Delta}_{P,n},$$

where

$$\bar{\delta}_{P,n} = \max_{k\in\mathcal{T}_{n,L}} \left|\left| \sum_{t\leq k} \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n} - \sum_{t\leq k} \bar{\boldsymbol{\Sigma}}^{\boldsymbol{R}}_{P,t,n} \right|\right|_{\mathrm{tr}}$$

and

$$\rho_{P,n} = \max_{t \in \mathcal{T}_{n,L}} ||\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n}||_{\text{tr}}.$$

Let $\bar{\boldsymbol{R}}_n = (\bar{\boldsymbol{R}}_{t,n})_{t \in \mathcal{T}_{n,L}}$ and denote the $(1-\alpha)$ quantile of $S_{n,p}(\bar{\boldsymbol{R}}_n)$ by $\bar{q}_{1-\alpha}$. For each $n \in \mathbb{N}$ and $P \in \mathcal{P}^*_{0,n}$, we have

$$\begin{aligned}
&\mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \bar{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&\overset{(1)}{\leq} \mathbb{P}_P(S_{n,p}(\bar{\boldsymbol{R}}_n) > \bar{q}_{1-\alpha+\nu_n}) \\
&+ \mathbb{P}_P\left(\max_{s \in \mathcal{T}_{n,L}} \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \leq s} (\boldsymbol{R}^\dagger_{t,n} - \bar{\boldsymbol{R}}_{t,n}) \right|\right|_2 > \frac{\tau_n}{4}\right) \\
&\overset{(2)}{=} \mathbb{P}_P(S_{n,p}(\bar{\boldsymbol{R}}_n) > \bar{q}_{1-\alpha+\nu_n}) \\
&+ \mathbb{P}_P\left(\max_{s \in \mathcal{T}_{n,L}} \left|\left| \frac{1}{\sqrt{T_{n,L}}} \sum_{t \leq s} (\boldsymbol{R}^\dagger_{t,n} - \bar{\boldsymbol{R}}_{t,n}) \right|\right|^2_2 > \frac{\tau_n^2}{16}\right) \\
&\overset{(3)}{\leq} \mathbb{P}_P(S_{n,p}(\bar{\boldsymbol{R}}_n) > \bar{q}_{1-\alpha+\nu_n}) \\
&+ 16\tau_n^{-2} T_{n,L}^{-1} \mathbb{E}_P\left(\max_{s \in \mathcal{T}_{n,L}} \left|\left| \sum_{t \leq s} (\boldsymbol{R}^\dagger_{t,n} - \bar{\boldsymbol{R}}_{t,n}) \right|\right|^2_2\right) \\
&\overset{(4)}{\leq} (\alpha - \nu_n) + 16\tau_n^{-2} \bar{\Delta}_{P,n} T_{n,L}^{-1} \overset{(5)}{=} \alpha + \left[16\tau_n^{-2} \bar{\Delta}_{P,n} T_{n,L}^{-1} - \nu_n\right],
\end{aligned}$$

where the previous lines follow by (1) the triangle inequality, subadditivity, the assumption about the form of the test statistic, (2) squaring, (3) Markov's inequality, (4) Proposition 4.2 from Mies and Steland [MS23], and (5) rearranging terms.

We see that if $\left[16\tau_n^{-2} \bar{\Delta}_{P,n} T_{n,L}^{-1} - \nu_n\right] < 0$ then

$$\mathbb{P}_P(S_{n,p}(\boldsymbol{R}^\dagger_n) > \bar{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) < \alpha,$$

which implies that $\bar{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}$ is greater than the $(1-\alpha)$ quantile of $S_{n,p}(\boldsymbol{R}^\dagger_n)$, which we denote by $q^\dagger_{1-\alpha}$. Hence, if $q^\dagger_{1-\alpha} \geq \bar{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}$ then $\left[16\tau_n^{-2} \bar{\Delta}_{P,n} T_{n,L}^{-1} - \nu_n\right] \geq 0$, or equivalently $\bar{\Delta}_{P,n} \geq \frac{1}{16} T_{n,L} \nu_n \tau_n^2$.

Step 3.2: Now, we apply this idea with the cumulative covariance of the residual products $\hat{Q}^{\boldsymbol{R}}_n$. By the implication stated at the end of **Step 3.1** and monotonicity, we have

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(q_{1-\alpha} > \hat{q}_{1-\alpha+\nu_n} + \frac{\tau_n}{4}) \\
&\leq \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P(\hat{\Delta}_{P,n} \geq \frac{1}{16} T_{n,L} \nu_n \tau_n^2),
\end{aligned}$$

where we have replaced $\bar{\Delta}_{P,n}$, $\bar{\delta}_{P,n}$ with $\hat{\Delta}_{P,n}$, $\hat{\delta}_{P,n}$ which are defined by

$$\hat{\Delta}_{P,n} = K \log(T_{n,L}) \left[\sqrt{T_{n,L} \hat{\delta}_{P,n} \rho_{P,n}} + \rho_{P,n}\right],$$

$$\hat{\delta}_{P,n} = \max_{k \in \mathcal{T}_{n,L}} \left|\left| \sum_{t \leq k} \boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n} - \hat{Q}^{\boldsymbol{R}}_{k,n} \right|\right|_{\text{tr}},$$

and $\rho_{P,n}$ is defined in the same way as

$$\rho_{P,n} = \max_{t \in \mathcal{T}_{n,L}} ||\boldsymbol{\Sigma}^{\boldsymbol{R}}_{P,t,n}||_{\text{tr}}.$$

Thus, if we can find φ_n such that $\hat{\Delta}_{P,n} = O_{\mathcal{P}}(\varphi_n)$ and if we select the offsets so that $\nu_n \tau_n^2 \gg T_{n,L}^{-1}\varphi_n$, or equivalently $\nu_n \gg \tau_n^{-2} T_{n,L}^{-1}\varphi_n$, then we will have

$$\sup_{P \in \mathcal{P}_{0,n}^*} \mathbb{P}_P(\hat{\Delta}_{P,n} \geq \frac{1}{16} T_{n,L}\nu_n\tau_n^2) \to 0.$$

By Lemma B.5 and Assumption 3.5, we have

$$\sup_{P \in \mathcal{P}_{0,n}^*} \rho_{P,n} \leq K_\rho D_n (\bar{\Theta}^R)^2,$$

for some constant $K_\rho > 0$, so we obtain $\hat{\Delta}_{P,n} = O_{\mathcal{P}}(\varphi_n)$ with

$$\varphi_n = \log(T_{n,L}) D_n \left[T_{n,L}^{\frac{1}{2}} D_n^{-\frac{1}{2}} \hat{\delta}_{P,n}^{\frac{1}{2}} + 1 \right].$$

Plugging φ_n into the offset condition $\nu_n \gg \tau_n^{-2} T_{n,L}^{-1}\varphi_n$ that we wish to satisfy, if we have

$$\nu_n \gg \log(T_{n,L}) D_n \left(\tau_n^{-2} \left(T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} \hat{\delta}_{P,n}^{\frac{1}{2}} + T_{n,L}^{-1} \right) \right),$$

then

$$\sup_{P \in \mathcal{P}_{0,n}^*} \mathbb{P}_P(\hat{\Delta}_{P,n} \geq \frac{1}{16} T_{n,L}\nu_n\tau_n^2) \to 0.$$

It remains to analyze $\hat{\delta}_{P,n}$.

Step 3.3: By the triangle inequality, we have

$$\begin{aligned}
\hat{\delta}_{P,n} &= \max_{k \in \mathcal{T}_{n,L}} \left\| \sum_{t \leq k} \boldsymbol{\Sigma}_{P,t,n}^{\boldsymbol{R}} - \hat{Q}_{k,n}^{\boldsymbol{R}} \right\|_{\text{tr}} \\
&\leq \max_{k \in \mathcal{T}_{n,L}} \left\| \sum_{t \leq k} \boldsymbol{\Sigma}_{P,t,n}^{\boldsymbol{R}} - Q_{P,k,n}^{\boldsymbol{R}} \right\|_{\text{tr}} \\
&+ \max_{k \in \mathcal{T}_{n,L}} \left\| \hat{Q}_{k,n}^{\boldsymbol{R}} - Q_{P,k,n}^{\boldsymbol{R}} \right\|_{\text{tr}}.
\end{aligned}$$

By Lemma B.3, Assumption 3.5, and Assumption 3.6, the covariance estimation error can be bounded as

$$\begin{aligned}
&\sup_{P \in \mathcal{P}_{0,n}^*} \mathbb{E}_P \left(\max_{k \in \mathcal{T}_{n,L}} \left\| \sum_{t \leq k} \boldsymbol{\Sigma}_{P,t,n}^{\boldsymbol{R}} - Q_{P,k,n}^{\boldsymbol{R}} \right\|_{\text{tr}} \right) \\
&\leq K(\bar{\Theta}^R)^2 D_n \left(\bar{\Gamma}_n^R L_n^{\frac{1}{2}} + T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} L_n^{\frac{1}{2}} + T_{n,L} L_n^{-1} + T_{n,L} L_n^{2-\bar{\beta}^R} \right) \\
&= O(r_{n,1}^\delta),
\end{aligned}$$

where

$$r_{n,1}^\delta = D_n \left(\bar{\Gamma}_n^R L_n^{\frac{1}{2}} + T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} L_n^{\frac{1}{2}} + T_{n,L} L_n^{-1} + T_{n,L} L_n^{2-\bar{\beta}^R} \right).$$

Next, we must handle the prediction error due using the products of residuals. For $\epsilon > 0$, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P\left(\max_{k\in\mathcal{T}_{n,L}} \left|\left|\hat{Q}^{\boldsymbol{R}}_{k,n} - Q^{\boldsymbol{R}}_{P,k,n}\right|\right|_{\text{tr}} \wedge \epsilon\right) \\
&\overset{(1)}{=} \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P\left(\max_{k\in\mathcal{T}_{n,L}} \left|\left|\frac{1}{L_n}\sum_{r=L_n+\mathbb{T}_n^- -1}^{k}\left[\left(\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{R}}_{s,n}\right)^{\otimes 2} - \left(\sum_{s=r-L_n+1}^{r}\boldsymbol{R}_{P,s,n}\right)^{\otimes 2}\right]\right|\right|_{\text{tr}} \wedge \epsilon\right) \\
&\overset{(2)}{\leq} \frac{1}{L_n}\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P\left([\sum_{r\in\mathcal{T}_{n,L}}\left|\left|\left(\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{R}}_{s,n}\right)^{\otimes 2} - \left(\sum_{s=r-L_n+1}^{r}\boldsymbol{R}_{P,s,n}\right)^{\otimes 2}\right|\right|_{\text{tr}}] \wedge \epsilon\right) \\
&\overset{(3)}{\leq} \frac{2}{L_n}\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P([\sum_{r\in\mathcal{T}_{n,L}}(\left|\left|\sum_{s=r-L_n+1}^{r}\left(\hat{\boldsymbol{R}}_{s,n} - \boldsymbol{R}_{P,s,n}\right)\right|\right|_2 \left|\left|\sum_{s=r-L_n+1}^{r}\boldsymbol{R}_{P,s,n}\right|\right|_2 \\
&+ \left|\left|\sum_{s=r-L_n+1}^{r}\left(\hat{\boldsymbol{R}}_{s,n} - \boldsymbol{R}_{P,s,n}\right)\right|\right|_2^2)] \wedge \epsilon)
\end{aligned}$$

where (1) is from the definitions of $Q^{\boldsymbol{R}}_{P,k,n}$, $\hat{Q}^{\boldsymbol{R}}_{k,n}$, (2) is from the triangle inequality, and (3) is from the following outer product inequality for vectors $\hat{v}, v \in \mathbb{R}^d$

$$\begin{aligned}
&||\hat{v}\hat{v}^\top - vv^\top||_{\text{tr}} \\
&\overset{(i)}{=} ||(\hat{v}-v)v^\top + v(\hat{v}-v)^\top + (\hat{v}-v)(\hat{v}-v)^\top||_{\text{tr}} \\
&\overset{(ii)}{\leq} 2||(\hat{v}-v)v^\top||_{\text{tr}} + ||(\hat{v}-v)(\hat{v}-v)^\top||_{\text{tr}} \\
&\overset{(iii)}{=} 2||\hat{v}-v||_2||v||_2 + ||\hat{v}-v||_2^2,
\end{aligned}$$

where (i) follows from adding and subtracting terms, (ii) follows from the triangle inequality, and (iii) follows by the properties of outer products and the definition of the trace norm.

For each $r \in \mathcal{T}_{n,L}$, we have the following decomposition into the three bias terms from Step 1 by the triangle inequality

$$\begin{aligned}
&\left|\left|\sum_{s=r-L_n+1}^{r}\left(\hat{\boldsymbol{R}}_{s,n} - \boldsymbol{R}_{P,s,n}\right)\right|\right|_2 \\
&\leq \left|\left|\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,s,n}(\boldsymbol{Z}_{s,n})\right|\right|_2 + \left|\left|\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,s,n}(\boldsymbol{Z}_{s,n})\right|\right|_2 + \left|\left|\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,s,n}(\boldsymbol{Z}_{s,n})\right|\right|_2.
\end{aligned}$$

Observe that for any $\delta > 0$ and any $r \in \mathcal{T}_{n,L}$, we have

$$\begin{aligned}
&\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{P}_P\left(\left|\left|\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,s,n}(\boldsymbol{Z}_{s,n})\right|\right|_2 > \delta L_n^{\frac{1}{2}}\tau_n^7 D_n^{-2}\right) \\
&\leq \delta^{-1}L_n^{-\frac{1}{2}}\tau_n^{-7}D_n^2 \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P\left(\left|\left|\sum_{s=r-L_n+1}^{r}\hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{g}}_{P,s,n}(\boldsymbol{Z}_{s,n})\right|\right|_2\right) \\
&\delta^{-1}L_n^{\frac{1}{2}}\tau_n^{-7}D_n^3 \sup_{P\in\mathcal{P}^*_{0,n}} \max_{(i,j,a,b)\in\mathcal{D}_n} \mathbb{E}_P\left(|\hat{w}^f_{P,t,n,i,a}|^2\right)^{\frac{1}{2}} \mathbb{E}_P\left(|\hat{w}^g_{P,t,n,j,b}|^2\right)^{\frac{1}{2}} \\
&= o(1),
\end{aligned}$$

using the same arguments as Step 1.1 replacing $T_{n,L}$ with L_n, and noting that D_n grows as $D_n = O(T_n^{\frac{1}{6}})$ which corresponds to a lag-window size of $L_n = O(T_n^{\frac{1}{3-\delta_L}})$ for any $\delta_L > 0$.

Next, for any $\delta > 0$ and any $r \in \mathcal{T}_{n,L}$, we have

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,s,n}(\boldsymbol{Z}_{s,n}) \right|\right|_2 > \delta L_n^{\frac{1}{2}} D_n^{-2} \tau_n^7 \right) \\
&= \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,s,n}(\boldsymbol{Z}_{s,n}) \right|\right|_2^2 > \delta^2 L_n D_n^{-4} \tau_n^{14} \right) \\
&\leq \delta^{-2} L_n^{-1} D_n^4 \tau_n^{-14} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{E}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{w}}^{\boldsymbol{g},\boldsymbol{\epsilon}}_{P,s,n}(\boldsymbol{Z}_{s,n}) \right|\right|_2^2 \right) \\
&\leq \delta^{-2} D_n^5 \tau_n^{-14} \bar{K}^2 \sup_{P \in \mathcal{P}^*_{0,n}} \max_{t \in \mathcal{T}_{n,L}} \max_{j \in [d_Y]} \max_{b \in B_j} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^2) \\
&= o(1)
\end{aligned}$$

for some $\bar{K} > 0$ using the same arguments as Step 1.2 replacing $T_{n,L}$ with L_n.

The same arguments can be used to show that

$$\left|\left| \sum_{s=r-L_n+1}^{r} \hat{\boldsymbol{w}}^{\boldsymbol{f},\boldsymbol{\xi}}_{P,s,n}(\boldsymbol{Z}_{s,n}) \right|\right|_2 = o_{\mathcal{P}}(L_n^{\frac{1}{2}} D_n^{-2} \tau_n^7).$$

Putting it all together, for any $r \in \mathcal{T}_{n,L}$ we have

$$\left|\left| \sum_{s=r-L_n+1}^{r} \left(\hat{\boldsymbol{R}}_{s,n} - \boldsymbol{R}_{P,s,n} \right) \right|\right|_2 = o_{\mathcal{P}}(L_n^{\frac{1}{2}} D_n^{-2} \tau_n^7).$$

By Lemma B.2 and Assumption 3.5, we have for all $r \in \mathcal{T}_{n,L}$ that

$$\begin{aligned}
&\sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \boldsymbol{R}_{P,s,n} \right|\right|_2 > L_n^{\frac{1}{2}} D_n^{\frac{1}{2}} \epsilon \right) \\
&= \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{P}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \boldsymbol{R}_{P,s,n} \right|\right|_2^2 > L_n D_n \epsilon^2 \right) \\
&\leq L_n^{-1} D_n^{-1} \epsilon^{-2} \sup_{P \in \mathcal{P}^*_{0,n}} \mathbb{E}_P \left(\left|\left| \sum_{s=r-L_n+1}^{r} \boldsymbol{R}_{P,s,n} \right|\right|_2^2 \right) \\
&\leq L_n^{-1} D_n^{-1} \epsilon^{-2} (2 L_n^{\frac{1}{2}} D_n^{\frac{1}{2}} \bar{\Theta}^R K \sum_{h=1}^{\infty} h^{-\bar{\beta}^R})^2 \\
&= O(1)
\end{aligned}$$

where $\sum_{h=1}^{\infty} h^{-\bar{\beta}^R} < \infty$ since $\bar{\beta}^R > 1$ by Assumption 3.5.

By Markov's inequality, bounded convergence (Lemma B.7), noting that the previous statements hold for all times in $\mathcal{T}_{n,L}$, we have

$$\max_{k \in \mathcal{T}_{n,L}} \left|\left| \hat{Q}^{\boldsymbol{R}}_{k,n} - Q^{\boldsymbol{R}}_{P,k,n} \right|\right|_{\text{tr}} = O_{\mathcal{P}}(r^{\delta}_{n,2}),$$

where the rate $r^{\delta}_{n,2}$ is defied as

$$r^{\delta}_{n,2} = T_{n,L} \tau_n^7 D_n^{-\frac{3}{2}} + T_{n,L} D_n^{-4} \tau_n^{14}.$$

Recall the original offset condition

$$\nu_n \gg \log(T_{n,L}) D_n \left(\tau_n^{-2} \left(T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} \hat{\delta}^{\frac{1}{2}}_{P,n} + T_{n,L}^{-1} \right) \right),$$

where

$$\hat{\delta}_{P,n} = O_{\mathcal{P}}(r_{n,1}^{\delta} + r_{n,2}^{\delta}),$$

and

$$r_{n,1}^{\delta} = D_n \left(\bar{\Gamma}_n^R L_n^{\frac{1}{2}} + T_{n,L}^{\frac{1}{2}} D_n^{\frac{1}{2}} L_n^{\frac{1}{2}} + T_{n,L} L_n^{-1} + T_{n,L} L_n^{2-\bar{\beta}^R} \right),$$

$$r_{n,2}^{\delta} = T_{n,L} \tau_n^7 D_n^{-\frac{3}{2}} + T_{n,L} D_n^{-4} \tau_n^{14}.$$

Observe that

$$\begin{aligned}
&T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} (r_{n,1}^{\delta})^{\frac{1}{2}} + T_{n,L}^{-1} \\
&\leq T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} \left(D_n^{\frac{1}{2}} ((\bar{\Gamma}_n^R)^{\frac{1}{2}} L_n^{\frac{1}{4}} + T_{n,L}^{\frac{1}{4}} D_n^{\frac{1}{4}} L_n^{\frac{1}{4}} + T_{n,L}^{\frac{1}{2}} L_n^{-\frac{1}{2}} + T_{n,L}^{\frac{1}{2}} L_n^{1-\frac{\bar{\beta}^R}{2}}) \right) + T_{n,L}^{-1} \\
&= T_{n,L}^{-\frac{1}{2}} (\bar{\Gamma}_n^R)^{\frac{1}{2}} L_n^{\frac{1}{4}} + T_{n,L}^{-\frac{1}{4}} D_n^{\frac{1}{4}} L_n^{\frac{1}{4}} + L_n^{-\frac{1}{2}} + L_n^{1-\frac{\bar{\beta}^R}{2}} + T_{n,L}^{-1} \\
&= \varphi_{n,1},
\end{aligned}$$

which comes from the covariance estimation error.

Next, we have

$$\begin{aligned}
&T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} (r_{n,2}^{\delta})^{\frac{1}{2}} \\
&\leq T_{n,L}^{-\frac{1}{2}} D_n^{-\frac{1}{2}} \left(T_{n,L}^{\frac{1}{2}} \tau_n^{\frac{7}{2}} D_n^{-\frac{3}{4}} + T_{n,L}^{\frac{1}{2}} \tau_n^7 D_n^{-2} \right) \\
&= \tau_n^{\frac{7}{2}} D_n^{-\frac{5}{4}} + \tau_n^7 D_n^{-\frac{5}{2}} \\
&= \varphi_{n,2},
\end{aligned}$$

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

The assumption on the offset condition (13) implies that

$$\nu_n \gg \log(T_{n,L}) D_n \left(\tau_n^{-2} \left(\varphi_{n,1} + \varphi_{n,2} \right) \right),$$

and therefore

$$\sup_{P \in \mathcal{P}_{0,n}^*} \mathbb{P}_P(\hat{\Delta}_{P,n} \geq \frac{1}{16} T_{n,L} \nu_n \tau_n^2) \to 0.$$

Combining the results from steps 1, 2, and 3, we obtain the final result

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{0,n}^*} \mathbb{P}_P(S_{n,p}(\hat{\boldsymbol{R}}_n) > \hat{q}_{1-\alpha+\nu_n} + \tau_n) \leq \alpha.$$

□

C.2 Proof of Theorem 4.1

The distribution-uniform results from Section B, notably the strong Gaussian approximation, are applicable to locally stationary time series in particular. It suffices to show that the following two things. First, that the sieve time-varying nonlinear regression estimator satisfies the convergence rate requirement from Theorem 3.1 with the basis functions for time $\{\phi_{\ell_1}(u)\}$ chosen to be Legendre polynomials, the basis functions for the covariate values $\{\varphi_{\ell_2}(z)\}$ chosen to be mapped Legendre polynomials with algebraic mappings, and the number of basis functions for time and the covariate values chosen to satisfy $\tilde{c}_n = O(\log(n))$, $\tilde{d}_n = O(\log(n))$, respectively. Second, that this sieve estimator along with Assumptions 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 from Theorem 4.1 imply Assumptions 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 for Theorem 3.1.

The causal representations or the observed processes and error processes from Assumptions 4.1, 4.4 are defined for all $u \in \mathcal{U}_n$ and thus for all $\{t/n\}_{t \in \mathcal{T}_n} \subset \mathcal{U}_n$ in particular. We see that Assumptions 3.1, 3.4 are satisfied by using the following notation. For a generic high-dimensional locally stationary observed process $W \in \{X, Y, Z\}$ and any time t, sample size n, dimension l, and time-offset d, we write

$$G_{t,n}^W(\cdot) = \tilde{G}_n^W(t/n, \cdot), \; G_{t,n,l}^W(\cdot) = \tilde{G}_{n,l}^W(t/n, \cdot), \; G_{t,n,l,d}^W(\cdot) = \tilde{G}_{n,l,d}^W(t/n, \cdot),$$

to respectively denote the causal representation for all dimensions of W at time t, the causal representation for dimension l of W at time t, and the causal representation for time-offset d of dimension l of W at time t. Similarly, for a generic high-dimensional locally stationary error process $e \in \{\varepsilon, \xi\}$ and any distribution P, time t, sample size n, dimension l, and time-offset d, we denote

$$G^e_{P,t,n}(\cdot) = \tilde{G}^e_{P,n}(t/n, \cdot),\ G^e_{P,t,n,l}(\cdot) = \tilde{G}^e_{P,n,l}(t/n, \cdot),\ G^e_{P,t,n,l,d}(\cdot) = \tilde{G}^e_{P,n,l,d}(t/n, \cdot),$$

to respectively denote the causal representation for all dimensions of e at time t, the causal representation for dimension l of e at time t, and the causal representation for time-offset d of dimension l of e at time t. Using this notation, Assumption 3.5 is the same as Assumption 4.5. Also, we can show that Assumption 3.6 is satisfied with $\bar{\Gamma}^R_n = D_n^{\frac{1}{2}}$ by using linearity of expectation and directly applying the stochastic Lipschitz condition for the product of errors from the discussion below Assumption 4.6 to each term in the sum. Similarly, we use the following notation for the regression functions. For any distribution P, time t, sample size n, dimensions i, j, time-offsets a, b, we write

$$f_{P,t,n,i,a}(\cdot) = f_{P,n,i,a}(t/n, \cdot),\ \hat{f}_{t,n,i,a}(\cdot) = \hat{f}_{t,n,i,a}(t/n, \cdot),$$
$$g_{P,t,n,j,b}(\cdot) = g_{P,n,j,b}(t/n, \cdot),\ \hat{g}_{t,n,j,b}(\cdot) = \hat{g}_{t,n,j,b}(t/n, \cdot),$$

to respectively denote the time-varying regression functions from Subsection 4.3 within the triangular array framework from Subsection 2.3.

For the sieve estimator described in Subsection 4.3, the predictors $\hat{f}_{t,n,i,a}$, $\hat{g}_{t,n,j,b}$ and corresponding predictions $\hat{f}_{t,n,i,a}(z)$, $\hat{g}_{t,n,j,b}(z)$ for all $t \in \mathcal{T}_n$, $z \in \mathbb{R}$ are a (Borel) measurable functions of $(X_{t,n,i,a}, \boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, $(Y_{t,n,j,b}, \boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively. The measurability of the causal mechanisms $G^X_{t,n,i,a}$, $G^Y_{t,n,j,b}$, $\boldsymbol{G}^{\boldsymbol{Z}}_{t,n}$ from Assumption 4.1 ensures that the predictors $\hat{f}_{t,n,i,a}$, $\hat{g}_{t,n,j,b}$ have the causal representations $G^{\mathcal{A}^{\hat{f}}}_{t,n,i,a}(\mathcal{H}^{\mathfrak{D}^{\hat{f}}}_{t,a})$, $G^{\mathcal{A}^{\hat{g}}}_{t,n,j,b}(\mathcal{H}^{\mathfrak{D}^{\hat{g}}}_{t,b})$ and corresponding predictions $\hat{f}_{t,n,i,a}(\boldsymbol{Z}_{t,n})$, $\hat{g}_{t,n,j,b}(\boldsymbol{Z}_{t,n})$ have the causal representations $G^{\hat{f}}_{t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a})$, $G^{\hat{g}}_{t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b})$ from Assumptions 3.2, 3.3, respectively. Further, the (Borel) measurability of the conditional expectations $f_{P,t,n,i,a}$, $g_{P,t,n,j,b}$ ensures that the prediction errors $\hat{w}^f_{P,t,n,i,a}$, $\hat{w}^g_{P,t,n,j,b}$ are (Borel) measurable functions of $(X_{t,n,i,a}, \boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, $(Y_{t,n,j,b}, \boldsymbol{Z}_{t,n})_{t\in\mathcal{T}_n}$, respectively. By the measurability of the causal mechanisms $G^X_{t,n,i,a}$, $G^Y_{t,n,j,b}$, $\boldsymbol{G}^{\boldsymbol{Z}}_{t,n}$ for the observed processes from Assumption 4.1, the prediction errors $\hat{w}^f_{P,t,n,i,a}$, $\hat{w}^g_{P,t,n,j,b}$ are ensured to have the causal representations $G^{\hat{w}^f}_{P,t,n,i,a}(\mathcal{H}^{\hat{f}}_{t,a})$, $G^{\hat{w}^g}_{P,t,n,j,b}(\mathcal{H}^{\hat{g}}_{t,b})$ from Assumption 3.3, respectively. Also, in view of the additive form for the time-varying regression functions in Subsection 4.3, the boundedness of the sieve predictions, and Assumption 4.2, there exists some $q \geq 2$ such that for all $n \in \mathbb{N}$, $t \in \mathcal{T}_n$, $(i, j, a, b) \in \mathcal{D}_n$ we have

$$\sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P(|\hat{w}^f_{P,t,n,i,a}|^q) < \infty,\quad \sup_{P\in\mathcal{P}^*_{0,n}} \mathbb{E}_P(|\hat{w}^g_{P,t,n,j,b}|^q) < \infty.$$

Hence, the prediction errors $\hat{w}^f_{P,t,n,i,a}$, $\hat{w}^g_{P,t,n,j,b}$ with the sieve estimator satisfy all the required conditions from Assumption 3.3.

Under Assumptions 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, the distribution-pointwise results from Theorem 3.2, Lemma C.1, Lemma C.2, and Lemma C.3 in Ding and Zhou [DZ21] hold for all $P \in \mathcal{P}^*_{0,n}$. Since the suprema (over all distributions in the collection) of the upper bounds are always finite, the corresponding distribution-uniform inequalities hold for $\mathcal{P}^*_{0,n}$ by basic properties of the supremum. Specifically, the sieve estimator with the basis functions for time $\{\phi_{\ell_1}(u)\}$ chosen to be Legendre polynomials, the basis functions for the covariate values $\{\varphi_{\ell_2}(z)\}$ chosen to be mapped Legendre polynomials with algebraic mappings, and the number of basis functions for time and the covariate values chosen to satisfy $\tilde{c}_n = O(\log(n))$, $\tilde{d}_n = O(\log(n))$, respectively, will satisfy

$$\sup_{P\in\mathcal{P}^*_{0,n}} \max_{i\in[d_X], a\in A_i} \max_{t\in\mathcal{T}_n} \mathbb{E}_P\left(\left|\hat{w}^f_{P,t,n,i,a}\right|^2\right)^{\frac{1}{2}} = o(T_n^{-\frac{1}{2+\delta}} \log^4(T_n)),$$

$$\sup_{P\in\mathcal{P}^*_{0,n}} \max_{j\in[d_Y], b\in B_j} \max_{t\in\mathcal{T}_n} \mathbb{E}_P\left(\left|\hat{w}^g_{P,t,n,j,b}\right|^2\right)^{\frac{1}{2}} = o(T_n^{-\frac{1}{2+\delta}} \log^4(T_n)),$$

for any $\delta > 0$. Since $D_n = O(T_n^{\frac{1}{6}})$ and $\tau_n = o(\log^{-(1+\delta)}(T_n))$, the sieve estimator achieves the convergence rates required by Theorem 3.1.

□