Testing Directed Acyclic Graph via Structural, Supervised and Generative Adversarial Learning

Chengchun Shi[†], Yunzhe Zhou[‡], and Lexin Li[‡]

[†]*London School of Economics and Political Science*
[‡]*University of California at Berkeley*

Abstract

In this article, we propose a new hypothesis testing method for directed acyclic graph (DAG). While there is a rich class of DAG estimation methods, there is a relative paucity of DAG inference solutions. Moreover, the existing methods often impose some specific model structures such as linear models or additive models, and assume independent data observations. Our proposed test instead allows the associations among the random variables to be nonlinear and the data to be time-dependent. We build the test based on some highly flexible neural networks learners. We establish the asymptotic guarantees of the test, while allowing either the number of subjects or the number of time points for each subject to diverge to infinity. We demonstrate the efficacy of the test through simulations and a brain connectivity network analysis.

Key Words: Brain connectivity networks; Directed acyclic graph; Hypothesis testing; Generative adversarial networks; Multilayer perceptron neural networks.

1 Introduction

Directed acyclic graph (DAG) is an important tool to characterize pairwise associations among multivariate and high-dimensional random variables. It has been frequently used in a wide range of scientific applications. One example is gene regulatory network analysis in genetics (Sachs et al., 2005), where the time-course expression data of multiple genes are measured over multiple cellular samples through microarray or RNA sequencing, and the goal is to understand the regulatory activation or repression relations among different genes. Another example is brain effective connectivity analysis in neuroscience (Garg et al., 2011), where the

time-course neural activities are measured at multiple brain regions for multiple experimental subjects through functional magnetic resonance imaging, and the goal is to infer the influences of brain regions exerting over each other under the stimulus.

There is a large body of literature studying penalized estimation of DAG given the observational data (see, e.g., Spirtes et al., 2000; van de Geer and Bühlmann, 2013; Zheng et al., 2018; Yuan et al., 2019, among many others). These works all impose some specific model structures, most often, linear models or additive models. There have recently emerged a number of proposals in the computer science literature that used neural networks or reinforcement learning to tackle nonlinear models and to estimate the associated DAG (Yu et al., 2019; Zheng et al., 2020; Zhu et al., 2020). While all these works have made crucial contributions, DAG model *estimation* is an utterly different problem from DAG inference. By inference, we mean hypothesis testing of individual edges throughout this article. The two problems are closely related, and both can, in effect, identify important links of a DAG. Besides, DAG inference usually relies on DAG estimation as a precedent step. Nevertheless, estimation does not produce an explicit quantification of statistical significance as inference does. Bayesian networks have been proposed for DAG estimation and inference. However, computationally, it is extremely difficult to search through all possible graph structures in a Bayesian network (Chickering et al., 2004), and as a result, the dimension of the Bayesian network is often small (Friston, 2011). There are very few frequentist inference solutions for inferring DAG structures. Only recently, Janková and van de Geer (2019) proposed a de-biased estimator to construct confidence intervals for the edge weights in a DAG, whereas Li et al. (2020) developed a constrained likelihood ratio test to infer individual edges or some given directed paths of a DAG. These works are probably the most relevant to our proposal. However, both have focused on Gaussian linear DAG, and cannot be easily extended to more general nonlinear DAG models. Moreover, all the above works considered the setting where the data observations are independent and identically distributed (i.i.d.). Learning DAG from time-dependent data remains largely unexplored.

There is another body of literature studying conditional independence testing (CIT); see Li and Fan (2019); Shah and Peters (2020); Shi et al. (2021) and the references therein. CIT is

closely related to DAG inference, and is to serve as a building block of our proposed testing procedure. On the other hand, naively performing CIT on two variables given the rest would fail to infer the directed edges of a DAG; see Section 2.2 for details. Besides, most CIT methods assume the data observations are independent, and are not suitable for the setting where the measurements are time-dependent.

In this article, we propose a novel statistical testing procedure for the inference of individual links or some given paths in a large and general DAG. The new test hinges upon some highly flexible neural networks-based machine learning techniques. The associations among the random variables can be either linear or nonlinear, the variables themselves can be either continuous or discrete-valued, and the observed data can be time-dependent.

Methodologically, we employ a number of state-of-the-art deep learning techniques that are highly flexible and can capture nonlinear associations among high-dimensional variables. We begin with a new characterization of directed edges under the additive noise structure (Peters et al., 2014); see Theorem 1. Based on this characterization, we propose a new testing procedure that integrates three key deep learning ingredients: (a) a DAG structural learning method based on neural networks or reinforcement learning to estimate the DAG; (b) a supervised learning method based on neural networks to estimate the conditional mean; and (c) a distribution generator produced by generative adversarial networks (Goodfellow et al., 2014, GANs) to approximate the conditional distribution of the variables in the DAG. We further couple these deep learning tools with some hypothesis testing strategies, including data splitting and cross-fitting to ensure a valid size control, and constructing a doubly robust test statistic as the maximum of multiple transformation functions to improve the power.

Theoretically, we establish the asymptotic size and power guarantees for the proposed test. The data-splitting and cross-fitting strategy ensures that our test achieves a valid type-I error control asymptotically under minimal conditions on those learning methods. As a result, our test procedure can work with a wide range of nonparametric estimators. Next, our DAG testing procedure requires a DAG estimation solution as a precedent step, which is common for almost all graph inference approaches (Cai, 2017). However, we do not assume the ordering of the

nodes is known a priori, but instead estimate this DAG ordering from the data using some DAG structural learning method. To establish the consistency of the proposed test, we require this ordering is consistently estimated; see condition (C1). Nevertheless, this order consistency is much weaker than requiring the initial DAG estimator to be selection consistent, or to satisfy the sure screening property. In other words, we only require a reasonably good initial estimator of DAG, which is order consistent but not necessarily selection consistent. We then develop a testing procedure that produces an explicit quantification of statistical significance for each individual link, and we show the test has the desired size and power guarantees. We also prove that the estimator from the DAG structural learning method we employ is indeed order consistent. Meanwhile, we discuss the impact on our test when this order consistency condition is not satisfied. Finally, for our theoretical analysis, we introduce a bidirectional asymptotic framework that allows either the number of subjects, or the number of time points for each subject, to diverge to infinity. This is useful for different types of applications. There are plenty of studies where the interest is about the general population, and thus it is reasonable to let the number of subjects or samples to diverge. Meanwhile, there are plenty of other applications, e.g., neuroimaging-based brain networks studies, where the number of subjects is almost always limited, but the scanning time and the temporal resolution can greatly increase. For those applications, it is more suitable to let the number of time points to diverge.

Our proposal is innovative and makes useful contributions in several ways.

First, rigorous inference of directed edges in DAG is a vital but also a long-standing open question. The existing solutions rely on particular model structures such as linear or additive models, and mostly deal with i.i.d. data. Such requirements can be restrictive in numerous applications, since the actual relations may be nonlinear and the data are correlated. By contrast, we only require an additive noise structure. To the best of our knowledge, our work is the first frequentist hypothesis testing solution for a general DAG with time-dependent data.

Second, we employ modern deep learning techniques such as neural networks and GANs to help address a classical statistical hypothesis testing problem. Such modern learning methods serve as nonparametric learners, and conceptually, play a similar role as splines and repro-

ducing kernels. Meanwhile, they are often more flexible and can handle more complex data structures. With increasingly efficient implementations of these methods and improved understandings of their theoretical properties (e.g., Bauer and Kohler, 2019; Farrell et al., 2021), this family of deep learning methods offer a powerful set of tools for classical statistical problems. Our proposal can be viewed as one of the early examples of harnessing such power, as the use of these deep learning techniques allows us to accurately estimate the DAG structure, the conditional means, as well as the distribution functions, and to improve the power of the test.

Third, even though the individual learning components such as neural networks, GANs and cross-fitting are not completely new, how to integrate them properly and effectively into a test with desired theoretical guarantees is highly nontrivial, and is one of the main contributions of this article. In effect, our proposed test achieves a parametric convergence rate and a parametric power guarantee while using nonparametric estimators. This is made possible mainly due to the innovative way we put together these learning components, which leads to a doubly robust test statistic (Tsiatis, 2007), in the sense that the proposed statistic is consistent, as long as either the conditional mean function in (b), or the distribution generator in (c) is correctly specified. In our solution, we propose to estimate both the conditional mean and the distribution generator fully nonparametrically. As such, the convergence rate of the two estimators, denoted by κ_1 and κ_2, respectively, may each be slower than the parametric rate. Nevertheless, we only require $\kappa_1 + \kappa_2 > 1/2$, which is totally achievable for the multilayer perceptron models and GANs; see the discussion after condition (C4). The key idea of our theoretically analysis is to show the bias of the estimating equation grows faster than the parametric rate. Thanks to the double robustness property of the test statistic, if we replace either estimator with its oracle value, the bias would be equal to zero. This observation, together with the Neyman orthogonality property of the estimating equation, ensures that the bias can be represented as a product of the difference between the two nonparametric estimators and their oracle values. Consequently, when $\kappa_1 + \kappa_2 > 1/2$, the test statistic converges at a parametric rate, the corresponding test controls the type-I error, and has a parametric power guarantee. We comment that, in their seminal work on double/debiased machine learning, Chernozhukov et al. (2018) proposed to

combine two machine learning estimators to infer the average treatment effect, which they showed to achieve a parametric convergence rate, even though each of the machine learning estimator converges at a nonparametric rate. Our result is similar in spirit as theirs, but targets a completely different problem, and thus is the first of its kind for DAG inference.

The rest of the article is organized as follows. We formally define the hypotheses, along with the model and data structure, in Section 2. We develop the testing procedure in Section 3, and establish the theoretical properties in Section 4. We study the empirical performance of the test through simulations and a real data example in Sections 5 and 6. We relegate several extensions, additional results, and all technical proofs to the Supplementary Appendix.

2 Problem Formulation

In this section, we first present the DAG model, based on which we formally define our hypotheses. We next propose an equivalent characterization of the hypotheses, for which we develop our testing procedure. Finally, we detail the data structure.

2.1 DAG model

Consider d random variables $X = (X_1, \ldots, X_d)^\top$, each with a finite fourth moment. We use a directed graph to characterize the relationships among these variables, where a node of the graph corresponds to a variable in X. For two nodes $i, j \in \{1, \ldots, d\}$, if an arrow is drawn from i to j, i.e., $i \to j$, then X_i is called a parent of X_j, and X_j a child of X_i. A directed path in the graph is a sequence of distinct nodes $i_1, \ldots, i_{d'}$, such that there is a directed edge $i_k \to i_{k+1}$ for all $k = 1, \ldots, d' - 1$. If there exists a directed path from i to j, then X_i is called an ancestor of X_j, and X_j a descendant of X_i. For node X_j, let PA_j, DS_j and AC_j denote the set of indices of the parents, descendants, and ancestors of X_j, respectively. Moreover, let $X_{\mathcal{M}}$ denote the sub-vector of X formed by those whose indices are in a subset $\mathcal{M} \subseteq \{1, \ldots, d\}$.

To rigorously formulate our problem, we make two assumptions.

(A1) The directed graph is acyclic; i.e., no variable is an ancestor of itself.

(A2) The DAG is identifiable from the joint distribution of X.

Condition (A1) has been commonly imposed in directed graph analysis. It does not permit any variable to be its own ancestor. As a result, the relationship between any two variables is unidirectional. Condition (A2) helps simplify the problem, and avoids dealing with the equivalence class of DAG. This condition is again frequently imposed in the DAG estimation literature (Zheng et al., 2018; Yuan et al., 2019; Li et al., 2020; Zheng et al., 2020). We discuss the extension to the equivalence class in Section A.4 of the Appendix.

We consider a class of structural equation models that follow an additive noise structure,

$$X_j = f_j(X_{\mathrm{PA}_j}) + \varepsilon_j, \quad \text{for any } j = 1, \ldots, d, \tag{1}$$

where $\{f_j\}_{j=1}^d$ are a set of continuous functions, and $\{\varepsilon_j\}_{j=1}^d$ are a set of independent zero mean random errors. Model (1) permits a fairly flexible structure. For instance, if each f_j is a linear function, then (1) reduces to a linear structural equation model. If each f_j is an additive function, i.e., $f_j(X_{\mathrm{PA}_j}) = \sum_{k \in \mathrm{PA}_j} f_{j,k}(X_k)$, then (1) becomes an additive model. In our test, we do *not* impose linear or additive model structures. Moreover, we can easily extend the proposed test to the setting of generalized linear model, where the X_j can be either continuous or discrete-valued. We discuss such an extension in Section A.3 of the Appendix.

Under model (1), the corresponding DAG is identifiable under some reasonable conditions. We consider three examples to discuss explicitly those conditions.

Example 1 (Gaussian graphical model)**.** Suppose $X_1, \ldots, X_d$ are jointly normal, and model (1) becomes $X_j = W_j^\top X_{\mathrm{PA}_j} + b_j + \varepsilon_j$, for some W_j and b_j. Then the corresponding DAG is identifiable, if the variance of the random error ε_j is the same for all $j = 1, \ldots, d$ (Bühlmann et al., 2014, Theorem 1).

Example 2 (Nonlinear graphical model with Gaussian noise)**.** Suppose $\varepsilon_1, \ldots, \varepsilon_d$ are jointly normal, but $X_1, \ldots, X_d$ are not. Then the corresponding DAG is identifiable, if each f_j is three times differentiable and not linear in any of its arguments (Peters et al., 2014, Corollary 31).

Example 3 (Nonlinear graphical model with general noise)**.** Suppose neither X_j nor ε_j is normal. Then the corresponding DAG is identifiable, if each f_j is non-constant in each of its arguments, and (1) is a restricted additive noise model (Peters et al., 2014, Definition 27).

2.2 Hypotheses and equivalent characterization

We next formally define the hypotheses we target, then give an equivalent characterization. For a given pair of nodes (j, k), $j, k = 1, \ldots, d, j \neq k$, we aim at the hypotheses:

$$H_0(j,k) : k \notin \text{PA}_j, \quad \text{versus} \quad H_1(j,k) : k \in \text{PA}_j. \tag{2}$$

When the alternative hypothesis holds, there is a link from X_k to X_j. In the following, we mainly focus on testing an individual link $H_0(j, k)$. We discuss the extension of testing a directed pathway, or a union of links, in Section A.1 and Section A.2 of the Appendix.

We next consider a pair of hypotheses that involve two variables that are *conditionally independent* (CI). The new hypotheses are closely related to (2), but are *not* exactly the same.

$$\begin{aligned} &H_0^*(j,k) : X_k \text{ and } X_j \text{ are CI given the rest of variables}, \quad \text{versus} \\ &H_1^*(j,k) : X_k \text{ and } X_j \text{ are } \textit{not} \text{ CI given the rest of variables}. \end{aligned} \tag{3}$$

We point out that, testing for (3) is generally *not* the same as testing for (2). To elaborate this, we consider a three-variable DAG with a v-structure.

Example 4 (v-structure)**.** Consider three random variables X_1, X_2, X_3 that form a v-structure, as illustrated in Figure 1(a), where X_1 and X_2 are the common parents of X_3. Even if X_1 and X_2 are *marginally* independent, they can be *conditionally* dependent given X_3. To better understand this, consider the following toy illustration. Either the ballgame or the rain could cause traffic jam, but they are uncorrelated. However, seeing traffic jam puts the ballgame and the rain in competition as a potential explanation. As such, these two events are conditionally dependent. Since X_2 is not a parent of X_1, both $H_0(1, 2)$ and $H_1^*(1, 2)$ hold. Consequently, testing for (3) can have an inflated type-I error for testing (2).

In this example, we see the reason that testing for (3) is not the same as for (2) is because the conditioning set of X_1 and X_2 contains their common descendant X_3. This key observation motivates us to consider a variant of (3), which we show is equivalent to (2) under certain conditions. We also remark that missing links in a DAG correspond to specific conditional independence between variables, but are not equivalent to marginal independence in general.

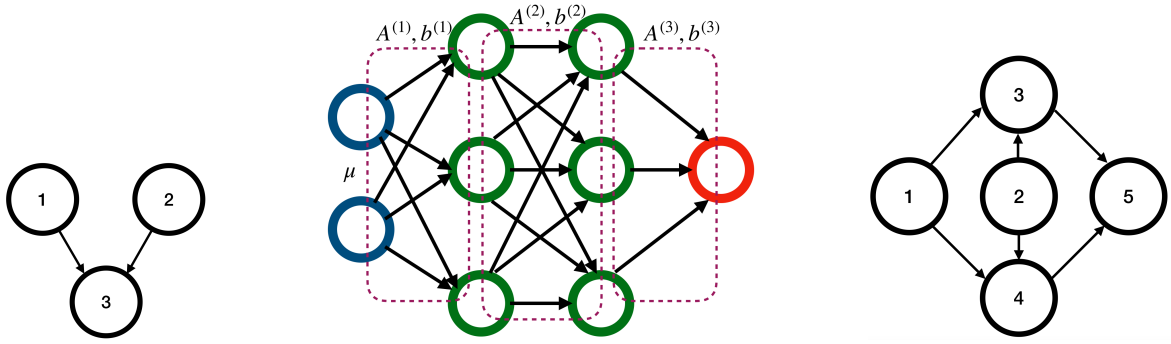

Figure 1: (a) A three-variable DAG with a v-structure; (b) A graphical illustration of a multi-layer perceptron, with two hidden layers, $m_0 = 2, m_1 = m_2 = 3$, where u is the input, $A^{(\ell)}$ and $b^{(\ell)}$ denote the corresponding parameters to produce the linear transformation for the $(\ell - 1)$th layer; (c) A five-variable DAG.

Specifically, for a given set of indices $\mathcal{M} \subseteq \{1, \ldots, d\}$ such that $j \notin \mathcal{M}$, and letting $X_{\mathcal{M}-\{k\}}$ denote the set of variables in $\mathcal{M} - \{k\}$, we consider the hypotheses:

$$\begin{aligned} &H_0^*(j, k|\mathcal{M}) : X_k \text{ and } X_j \text{ are CI given } X_{\mathcal{M}-\{k\}}, \quad \text{versus} \\ &H_1^*(j, k|\mathcal{M}) : X_k \text{ and } X_j \text{ are } \textit{not} \text{ CI given } X_{\mathcal{M}-\{k\}}, \end{aligned} \tag{4}$$

Proposition 1. *For a given pair of nodes (j, k) such that $j \in \text{DS}_k$, $j, k = 1, \ldots, d$, and for any $\mathcal{M}$ such that $j \notin \mathcal{M}$, $\text{PA}_j \subseteq \mathcal{M}$ and $\mathcal{M} \cap \text{DS}_j = \emptyset$, testing* (4) *is equivalent to testing* (2).

Proposition 1 forms the basis for our test. That is, to infer the directed links, we first restrict our attention to the pairs (j, k) such that $j \in \text{DS}_k$. Apparently, $H_0(j, k)$ does not hold when $j \notin \text{DS}_k$. Next, when devising a conditional independence test for $H_0(j, k)$, the conditioning set $\mathcal{M}$ is supposed to contain the parents of node j, but *cannot* contain any common descendants of j, k. Under these conditions, we establish the equivalence between (4) and (2). A similar idea of using CI tests for DAG structural learning was employed in Spirtes et al. (2000) too.

Next, we develop a test statistic for the hypotheses (4). We introduce a key quantity. Let h denote a square-integrable function that takes X_k and $X_{\mathcal{M}-\{k\}}$ as the input. Define

$$I(j, k|\mathcal{M}; h) = \mathbb{E}\left\{X_j - \mathbb{E}\left(X_j|X_{\mathcal{M}-\{k\}}\right)\right\}\left[h\left(X_k, X_{\mathcal{M}-\{k\}}\right) - \mathbb{E}\left\{h\left(X_k, X_{\mathcal{M}-\{k\}}\right)|X_{\mathcal{M}-\{k\}}\right\}\right].$$

Under the additive noise model (1), the next theorem connects this quantity with the null hypothesis $H_0^*(j, k|\mathcal{M})$ in (4). Together with Proposition 1, it shows that $I(j, k|\mathcal{M}; h)$ can serve

as a test statistic for (4), and equivalently, for (2) that we target.

Theorem 1. *Suppose* (1) *holds. For a given pair of nodes* (j,k) *such that* $j \in \text{DS}_k$, $j,k = 1,\ldots,d$, *for any* $\mathcal{M}$ *such that* $j \notin \mathcal{M}$, $\text{PA}_j \subseteq \mathcal{M}$ *and* $\mathcal{M} \cap \text{DS}_j = \emptyset$, *the null hypothesis* $H_0^*(j,k|\mathcal{M})$ *in* (4) *is equivalent to* $\sup_h |I(j,k|\mathcal{M};h)| = 0$ *where the supremum is taken over all square-integrable functions* h.

Theorem 1 immediately suggests a possible testing procedure for (4). That is, we first employ a DAG estimator to learn the ancestors and descendants for node j. We then consider a natural choice for h, where $h\left(X_k, X_{\mathcal{M}-\{k\}}\right) = X_k$. Then $I(j,k|\mathcal{M};h)$ becomes

$$I(j,k|\mathcal{M};h) = \mathbb{E}\left\{X_j - \mathbb{E}\left(X_j|X_{\mathcal{M}-\{k\}}\right)\right\}\left\{X_k - \mathbb{E}\left(X_k|X_{\mathcal{M}-\{k\}}\right)\right\}. \tag{5}$$

By Theorem 1, under the null hypothesis $H_0^*(j,k|\mathcal{M})$, a consistent estimator for (5) should be close to zero. A Wald type test can then be devised with i.i.d. data. That is, we first obtain an estimator $\widehat{I}_{j,k}$ for $I(j,k|\mathcal{M};h)$, by plugging in the estimators of the conditional mean functions, $\widehat{\mathbb{E}}\left(X_j|X_{\mathcal{M}-\{k\}}\right)$ and $\widehat{\mathbb{E}}\left(X_k|X_{\mathcal{M}-\{k\}}\right)$. We then get an estimator of its asymptotic variance $\widehat{\sigma}^2_{j,k}$, and obtain the Wald type test statistic, $\sqrt{N}\widehat{\sigma}^{-1}_{j,k}\widehat{I}_{j,k}$, where N is the number of samples. Such a test is similar in spirit as the tests of Zhang et al. (2018) and Shah and Peters (2020). Since it involves estimation of two conditional mean functions, we refer to it as the *double regression-based test*. We later numerically compare our proposed test with this test.

On the other hand, this double regression-based test has some limitations. One is that it requires the set $\mathcal{M}$ to be fixed. To meet the requirement in Proposition 1, $\mathcal{M}$ needs to be determined in a data-adaptive way. The resulting test may not control the type-I error due to the dependence between $\mathcal{M}$ and the estimator of the mean functions in $\widehat{I}_{j,k}$. Another limitation is that it may not have a sufficient power to detect $H_1(j,k)$. As an illustration, we revisit Example 4. For this example, consider the structural equation model: $X_1 = \varepsilon_1$, $X_2 = \varepsilon_2$, and $X_3 = X_1^2 + X_2 + \varepsilon_3$. Under this model, $H_1(1,3)$ holds. Meanwhile, $I(1,3) = \mathbb{E}(X_3 - X_2)X_1 = \mathbb{E}\varepsilon_1^3$. When the distribution of ε_1 is symmetric, $I(1,3) = 0$, despite the fact that X_1 is a parent of X_3. As such, for this example, the double regression-based test is to have no power at all.

To address the first limitation, we employ the sample splitting strategy to ensure its size control. To address the second limitation, we consider multiple transformation functions h, instead of a single h, to improve the power. We detail our idea in Section 3.

2.3 Time-dependent observational data

Throughout this article, we use X to denote the population variables, and $\mathbb{X}$ to denote the data realizations. Suppose the data come from an observational study, and are of the form, $\{\mathbb{X}_{i,t,j} : i = 1, \ldots, N, t = 1, \ldots, T_i, j = 1, \ldots, d\}$, where i indexes the ith subject, t indexes the tth time point, and j indexes the jth random variable. Suppose there are totally N subjects, with T_i observations for the ith subject. Write $\mathbb{X}_{i,t} = (\mathbb{X}_{i,t,1}, \ldots, \mathbb{X}_{i,t,d})^\top$, $i = 1, \ldots, N, t = 1, \ldots, T_i$. We consider the following data structure.

(B1) Across subjects, the measurements $\mathbb{X}_{1,t}$, ..., $\mathbb{X}_{N,t}$ are i.i.d.

(B2) Across time points, the random vectors $\mathbb{X}_{i,1}$, ..., $\mathbb{X}_{i,T_i}$ are stationary.

(B3) For any i, t, $\mathbb{X}_{i,t,1}$, ..., $\mathbb{X}_{i,t,d}$ are DAG-structured. In addition, their joint distribution is the same as that of $X_1, \ldots, X_d$.

Condition (B1) is reasonable, as the subjects are usually independent from each other. We do not study the scenario where the data come from the same families or clusters. Condition (B2) about the stationarity is common in numerous applications such as brain connectivity analysis (Bullmore and Sporns, 2009; Qiu et al., 2016; Wang et al., 2016). Condition (B3) brings the data into the DAG framework that we study. Note that (B3) does not allow directed edges from past to future observations. Meanwhile, we discuss the extensions of our test for non-stationary DAG, or for past to future edges, in Section A.5 of the Appendix.

3 Testing Procedure

In this section, we develop an inferential procedure for the hypotheses in (2) for a given pair (j, k), through (4), given the observational data $\mathbb{X}_{i,t}$. We first present the main ideas and the complete procedure, then detail the major steps. As our test is based on **S**tructural learning, s**U**pervised learning, and **G**enerative **A**dve**R**sarial networks, we call our method SUGAR.

3.1 The main algorithm

Our main idea is to construct a series of measures $\{I(j,k|\mathcal{M};h_b) : b = 1,\ldots,B\}$, for a large number of transformation functions $h_1,\ldots,h_B$, then take the maximum of some standardized version of $I(j,k|\mathcal{M};h_b)$. Toward that goal, our test involves three key components:

(a) A DAG structural learning method to learn the set of indices $\mathcal{M}$ that satisfy Proposition 1;

(b) A supervised learning method to estimate the conditional mean function $\mathbb{E}\left(X_j|X_{\mathcal{M}-\{k\}}\right)$;

(c) A distribution generator to approximate the conditional distribution of the variables.

For (a), we apply a structural learning algorithm to learn the underlying DAG $\mathcal{G}$ corresponding to X. The input of this step is the observed data $\{\mathbb{X}_{i,t,j} : i = 1,\ldots,N, t = 1,\ldots,T_i, j = 1,\ldots,d\}$, and the output is the estimated DAG. We then set $\mathcal{M}$ as the estimated set of ancestors of X_j. To capture possible sparsity and nonlinear associations in $\mathcal{G}$, we employ the DAG estimation method of Zheng et al. (2020). See Section 3.3 for details.

For (b), we employ a supervised learning algorithm. The input of this step is $X_{\mathcal{M}-\{k\}}$ that serves as the "predictors", and X_j that serves as the "response", and the output is the estimated mean function $\widehat{\mathbb{E}}\left(X_j|X_{\mathcal{M}-\{k\}}\right)$. We employ a multilayer perceptron learner, which has a good capacity of estimating complex high-dimensional mean, and the estimator has the desired consistency guarantees (Farrell et al., 2021). See Section 3.4 for details.

For (c), we propose to use generative adversarial networks (Goodfellow et al., 2014, GANs) to approximate the conditional distribution of X_k given $X_{\mathcal{M}-\{k\}}$. The input of this step is $\mathbb{X}_{i,t,\mathcal{M}-\{k\}}$ and multivariate Gaussian noise vectors, and the output is the learnt generator model, with a set of M pseudo samples $\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}$, $m = 1,\ldots,M$, that have a similar distribution as the training samples. We employ a generator model with the Sinkhorn divergence loss (Genevay et al., 2018) to mitigate the potential bias of GANs. See Section 3.5 for details.

Given the generated pseudo samples, we then proceed to estimate the conditional mean function $\mathbb{E}\big\{h_b\big(X_k, X_{\mathcal{M}-\{k\}}\big)|\ X_{\mathcal{M}-\{k\}}\big\}$ in (5), and construct the corresponding test statistic. We also incorporate the data-splitting and cross-fitting strategy (Romano and DiCiccio, 2019), to ensure a valid type-I error control for the test under minimal conditions for the above three

learners. Specifically, we randomly split the samples into two equal halves $\mathcal{I}_1 \cup \mathcal{I}_2$, where $\mathcal{I}_s$ denotes the set of subsample indices, $s = 1, 2$. We then compute the three learners in (a) to (c) using each half of the data separately. Based on these learners, we next use cross-fitting to estimate $\{I(j, k|\mathcal{M}; h_b)\}_{b=1}^{B}$, and their associated standard deviations. We construct our test statistic as the largest standardized version of $I(j, k|\mathcal{M}; h_b)$ in the absolute value. This leads to two Wald-type test statistics, one for each half of the data. Finally, we derive the p-values based on Gaussian approximation, and reject the null when either one of the p-value is smaller than $\alpha/2$. By Bonferroni's inequality, this yields a valid α-level test. See Section 3.2 for details.

A summary of the proposed testing procedure is given in Algorithm 1.

3.2 Test statistic and p-value

We begin with the presentation of our test, including the test statistic and the computation of the p-value, which are built on the three learners in (a) to (c) that we discuss in detail later.

First, for each half of the data, $s = 1, 2$, we begin with a bounded function class $\mathbb{H}^{(s)} = \left\{h_\omega^{(s)} : \omega \in \Omega^{(s)}\right\}$, indexed by some parameter ω. In our implementation, we consider the class of characteristic functions of X_k,

$$\mathbb{H}^{(1)} = \mathbb{H}^{(2)} = \mathbb{H} = \left\{ \cos(\omega X_k),\ \sin(\omega X_k) : \omega \in \mathbb{R} \right\}. \tag{6}$$

We note that (6) is not able to approximate the entire class of square integrable functions. Nevertheless, our numerical experiments have found that setting $\mathbb{H}^{(s)}$ according to (6) results in a good power empirically. Moreover, we note that one may set $\mathbb{H}^{(s)}$ to the class of characteristic functions of $(X_k, X_{\mathcal{M}^{(s)}})$. By the Fourier Theorem (Siebert, 1986), this alternative choice can approximate any square integrable function h, and the resulting test is consistent against all alternatives. We choose (6) for its simplicity as well as good empirical performance. Without loss of generality, we choose an even number for the total number of transformation functions B. We randomly generate i.i.d. standard normal variables $\omega_1, \ldots, \omega_{B/2}$, and set

$$h_b^{(s)}\left(X_k, X_{\mathcal{M}^{(s)}}\right) = \begin{cases} \cos(\omega_b X_k), & \text{for } b = 1, \ldots, B/2, \\ \sin(\omega_b X_k), & \text{for } b = B/2 + 1, \ldots, B. \end{cases}$$

Algorithm 1 Testing procedure for a given edge (j, k).

Step 1. Randomly split the data into two equal halves, $\{\mathbb{X}_{i,t,k}\}_{i\in\mathcal{I}_s, t=1,\ldots,T_i}$, $s = 1, 2$.

Step 2. For each half of the data, $s = 1, 2$,

(2a) Apply the structural learning method (9) to estimate the DAG $\mathcal{G}$. Denote the estimated set of ancestors of X_j by $\widehat{\text{AC}}_j^{(s)}$. Set $\mathcal{M}^{(s)} = \widehat{\text{AC}}_j^{(s)} - \{k\}$.

(2b) If $k \notin \widehat{\text{AC}}_j^{(s)}$, return the p-value, $p^{(s)}(j, k) = 1$.

Step 3. For $s = 1, 2$, apply the supervised learning method (10) to estimate the conditional mean function $\mathbb{E}\big(X_j|X_{\mathcal{M}^{(s)}}\big)$, and denote the estimator by $\widehat{g}^{(s)}$.

Step 4. For $s = 1, 2$, apply the GANs method to learn a generator model to approximate the conditional distribution of X_k given $X_{\mathcal{M}^{(s)}-\{k\}}$. It returns the learnt generator $\mathbb{G}^{(s)}$, and a set of pseudo samples $\big\{\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}\big\}_{i\in\mathcal{I}_s, t=1,\ldots,T_i, m=1,\ldots,M}$.

Step 5. Construct the test statistic:

(5a) Randomly generate B functions $\big\{h_b^{(s)}\big\}_{b=1}^{B}$ from the class $\mathbb{H}^{(s)}$ in (6).

(5b) For each (s, b), construct two standardized measures, $\widehat{T}_{b,\text{CF}}^{(s)}$ and $\widehat{T}_{b,\text{NCF}}^{(s)}$, with and without cross-fitting, using (7).

(5c) Select the index, $\widehat{b}^{(s)} = \arg\max_{b\in\{1,\ldots,B\}} \big|\widehat{T}_{b,\text{NCF}}^{(s)}\big|$, based on the measure without cross-fitting.

(5d) Set the test statistic as $\widehat{T}_{\widehat{b}^{(s)},\text{CF}}^{(s)}$, based on the measure with cross-fitting.

Step 6. Return the p-value:

(6a) Compute the p-value, $p^{(s)}(j, k) = 2\mathbb{P}\big\{Z_0 \geq \big|\widehat{T}_{\widehat{b}^{(s)},\text{CF}}^{(s)}\big|\big\}$, for each half of the data, $s = 1, 2$, where Z_0 is a standard normal random variable.

(6b) Return $p(j, k) = 2\min\big\{p^{(1)}(j, k), p^{(2)}(j, k)\big\}$.

Next, for each pair of (s, b), $b = 1, \ldots, B$, $s = 1, 2$, let $\widehat{\text{AC}}_j^{(s)}$, $\mathcal{M}^{(s)}$, $\widehat{g}^{(s)}$, and $\{\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}\}$ denote the estimated set of ancestors of X_j, the estimated set of indices $\mathcal{M}$, the estimated conditional mean function, and the generated pseudo samples, obtain from the components (a) to (c), respectively. We compute two estimators $\widehat{I}_{b,\text{CF}}^{(s)}$ and $\widehat{I}_{b,\text{NCF}}^{(s)}$ for the measure $I\left(j, k|\widehat{\text{AC}}_j^{(s)}, h_b^{(s)}\right)$, one *with* cross-fitting, and the other *without* cross-fitting. Specifically, we compute

$$\widehat{I}_{b,\text{CF}}^{(s)} = \Big(\sum_{i\in\mathcal{I}_s^c} T_i\Big)^{-1}\Big(\sum_{i\in\mathcal{I}_s^c} I_{i,t,b}^{(s)}\Big), \quad \widehat{I}_{b,\text{NCF}}^{(s)} = \Big(\sum_{i\in\mathcal{I}_s} T_i\Big)^{-1}\Big(\sum_{i\in\mathcal{I}_s} I_{i,t,b}^{(s)}\Big),$$

where

$$I_{i,t,b}^{(s)} = \left\{ \mathbb{X}_{i,t,j} - \widehat{g}^{(s)} \left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}} \right) \right\} \left\{ h_b^{(s)} \left(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}^{(s)}} \right) - \frac{1}{M} \sum_{m=1}^{M} h_b^{(s)} \left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}, \mathbb{X}_{i,t,\mathcal{M}^{(s)}} \right) \right\},$$

and M is the total number of pseudo samples. We note that, for $\widehat{I}_{b,\text{NCF}}^{(s)}$, we use the same subset of data to learn the graph, the generator, the condition mean function, and to construct $I_{i,t,b}^{(s)}$. By contrast, for $\widehat{I}_{b,\text{CF}}^{(s)}$, the data used for the DAG learner, the conditional mean learner and the generator are independent from the data used to construct $I_{i,t,b}^{(s)}$.

Next, we compute the corresponding standard errors $\widehat{\sigma}_{b,\text{CF}}^{(s)}$ and $\widehat{\sigma}_{b,\text{NCF}}^{(s)}$ for $\widehat{I}_{b,\text{CF}}^{(s)}$ and $\widehat{I}_{b,\text{NCF}}^{(s)}$, respectively. Since our data are time-dependent, the usual sample variance would not be a consistent estimator. Therefore, we employ the batched estimator common in time series analysis (Carlstein, 1986). That is, we divide the data associated with each subject into non-overlapping batches, with each batch containing at most K observations. For simplicity, suppose T_i is divisible by K for all $i = 1, \ldots, N$. We obtain the following standard error estimators,

$$\widehat{\sigma}_{b,\text{CF}}^{(s)} = \left[\frac{K}{\sum_{i \in \mathcal{I}_s^c} T_i} \sum_{i \in \mathcal{I}_s^c} \sum_{k=1}^{T_i/K} \left\{ \frac{\sum_{t=(k-1)K+1}^{kK} \left(I_{i,t,b}^{(s)} - \widehat{I}_{b,\text{CF}}^{(s)} \right)}{\sqrt{K}} \right\}^2 \right]^{1/2},$$

$$\widehat{\sigma}_{b,\text{NCF}}^{(s)} = \left[\frac{K}{\sum_{i \in \mathcal{I}_s} T_i} \sum_{i \in \mathcal{I}_s} \sum_{k=1}^{T_i/K} \left\{ \frac{\sum_{t=(k-1)K+1}^{kK} \left(I_{i,t,b}^{(s)} - \widehat{I}_{b,\text{NCF}}^{(s)} \right)}{\sqrt{K}} \right\}^2 \right]^{1/2}.$$

Putting $\widehat{I}_{b,\text{CF}}^{(s)}$ and $\widehat{I}_{b,\text{NCF}}^{(s)}$ together with their standard error estimators, we obtain two standardized measures,

$$\widehat{T}_{b,\text{CF}}^{(s)} = \sqrt{\sum_{i \in \mathcal{I}_s^c} T_i} \left(\widehat{\sigma}_{b,\text{CF}}^{(s)} \right)^{-1} \widehat{I}_{b,\text{CF}}^{(s)}, \text{ and } \widehat{T}_{b,\text{NCF}}^{(s)} = \sqrt{\sum_{i \in \mathcal{I}_s} T_i} \left(\widehat{\sigma}_{b,\text{NCF}}^{(s)} \right)^{-1} \widehat{I}_{b,\text{NCF}}^{(s)}. \quad (7)$$

We then select the index $\widehat{b}^{(s)}$ that maximizes the standardized measure without cross-fitting, $\widehat{T}_{b,\text{NCF}}^{(s)}$, in absolute value, i.e., $\widehat{b}^{(s)} = \arg\max_{b \in \{1,\ldots,B\}} \left| \widehat{T}_{b,\text{NCF}}^{(s)} \right|$. We take the measure with cross-fitting, $\widehat{T}_{\widehat{b}^{(s)},\text{CF}}^{(s)}$, under the selected $\widehat{b}^{(s)}$, as our final test statistic.

We make a few remarks. First, we use the cross-fitting measure to construct the test statistic $\widehat{T}_{\widehat{b}^{(s)},\text{CF}}^{(s)}$. This enables us to derive its limiting distribution more easily. Specifically, conditional on the data in $\mathcal{I}_s$, for each $b = 1, \ldots, B$, $\widehat{T}_{b,\text{CF}}^{(s)}$ converges in distribution to standard normal

under the null. Since $\widehat{b}^{(s)}$ is determined by $\widehat{T}^{(s)}_{b,\text{NCF}}$, the index $\widehat{b}^{(s)}$ depends solely on the data in $\mathcal{I}_s$. Consequently, conditional on the data in $\mathcal{I}_s$, $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{CF}}$ converges in distribution to standard normal under the null as well. By contrast, the limiting distribution of the no-cross-fitting measure $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{NCF}}$ is unclear, due to the complicated dependence between $\widehat{b}^{(s)}$ and $\widehat{T}^{(s)}_{b,\text{NCF}}$.

Second, we use the no-cross-fitting measure to select the index $\widehat{b}^{(s)}$. As we show in Section 4, when the estimated conditional mean function and the distributional generator belong to the VC type class (Chernozhukov et al., 2014, Definition 2.1), the index $\widehat{b}^{(s)}$ that maximizes the no-cross-fitting measure $\{\widehat{T}^{(s)}_{b,\text{NCF}}\}$ asymptotically maximizes the cross-fitting measure $\{\widehat{T}^{(s)}_{b,\text{CF}}\}$ as well. This choice of the index $\widehat{b}^{(s)}$ is to maximize the power of the resulting test.

Finally, the random binary data splitting may introduce some sampling uncertainty. This issue is mitigated in our test, since we construct two test statistics based on both data subsets, then combine them to derive the final decision rule. One may also consider the multiple binary-splits idea of Meinshausen et al. (2009), or the multi-split idea of Romano and DiCiccio (2019). We discuss a multiple binary-splits version of our test in Section B.2 of the Appendix.

3.3 DAG structural learning

We next discuss the three key learning components (a) to (c) of our proposed test. The first is to estimate the DAG $\mathcal{G}$ associated with $X = (X_1, \ldots, X_d)^\top$, and to construct $\mathcal{M}$. In our implementation, we employ the neural structural learning method of Zheng et al. (2020). Other methods, e.g., Yu et al. (2019); Zhu et al. (2020), can be used as well.

Consider a multilayer perceptron (MLP) with L hidden layers and an activation function σ:

$$\text{MLP}\left(u; A^{(1)}, b^{(1)}, \ldots, A^{(L)}, b^{(L)}\right) = A^{(L)}\sigma\left\{\ldots A^{(2)}\sigma\left(A^{(1)}\mu + b^{(1)}\right)\ldots + b^{(L-1)}\right\} + b^{(L)}, \tag{8}$$

where $u \in \mathbb{R}^{m_0}$ is the input signal of the MLP, $A^{(s)} \in \mathbb{R}^{m_\ell \times m_{\ell-1}}$, $b^{(s)} \in \mathbb{R}^{m_\ell}$ are the parameters that produce the linear transformation of the $(\ell-1)$th layer, the output is a scalar with $m_L = 1$, and there are m_ℓ nodes at layer ℓ, $\ell = 0, \ldots, L$. See Figure 1(b) for a graphical illustration.

We employ MLP to approximate the functions f_j's in our DAG model (1). In our theoretical analysis, we focus on the setting where f_j's are a set of continuous functions. Meanwhile, we

may also consider a family of piecewise smooth functions (Imaizumi and Fukumizu, 2019) for f_j's. In both cases, neural networks models such as MLP can consistently estimate f_j's. Let $\theta_j = \left\{A_j^{(\ell)}, b_j^{(\ell)} : 1 \leq \ell \leq L\right\}$ collect all the parameters for the jth MLP that approximates f_j, and let $\theta = \{\theta_j\}_{j=1}^d$. Accordingly, θ uniquely determines a graph structure, i.e., how the variables are dependent to each other in the graph. We call this structure the graph induced by θ, and denote it by $\mathcal{G}(\theta)$. For each half of the data, $s = 1, 2$, we estimate the DAG via

$$\min_{\theta} \sum_{i \in \mathcal{I}_s} \sum_{t,j} \left\{\mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t}; \theta_j)\right\}^2, \quad \text{subject to } \mathcal{G}(\theta) \text{ is a DAG.}$$

This optimization, however, is challenging to solve, mainly due to the fact that the search space scales super-exponentially with the dimension d. To resolve this issue, Zheng et al. (2020) proposed a novel characterization of the acyclic constraint, and showed that the DAG constraint can be represented by $\text{trace}[\exp\{W(\theta) \circ W(\theta)\}] = d$, where $\circ$ denotes the Hadamard product, $\exp(W)$ is the matrix exponential of W, $\text{trace}(W)$ is the trace of W, and $W(\theta)$ is a $d \times d$ matrix whose (k, j)th entry equals the Euclidean norm of the kth column of $A_j^{(1)}$. Based on this characterization, the above optimization problem becomes,

$$\begin{aligned} \min_{\theta} \sum_{j=1}^{d} \left[\sum_{i \in \mathcal{I}_s} \sum_{t=1}^{T_i} \left\{\mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t}; \theta_j)\right\}^2 + \lambda n_s \left\|A_j^{(1)}\right\|_{1,1} \right], \\ \text{subject to} \quad \text{trace}[\exp\{W(\theta) \circ W(\theta)\}] = d, \end{aligned} \tag{9}$$

where $n_s = \sum_{i \in \mathcal{I}_s} T_i$ is the number of observations in $\mathcal{I}_s$, $\left\|A_j^{(1)}\right\|_{1,1}$ is the sum of all elements in $A_j^{(1)}$ in absolute values, and $\lambda > 0$ is a sparsity tuning parameter. Note that the sparsity penalization is placed only on $A_j^{(1)}$, since this is the only layer that determines the sparsity of the input variables $X_1, \ldots, X_d$. This new optimization problem in (9) can be efficiently solved using the augmented Lagrangian method (Zheng et al., 2020).

Let $\widehat{\mathcal{G}}^{(s)}$ denote the estimated graph, and $\widehat{\text{AC}}_j$ and $\widehat{\text{PA}}_j$ denote the corresponding estimated set of ancestors and parents of X_j, respectively. If $k \notin \widehat{\text{AC}}_j^{(s)}$, then it follows from $\text{PA}_j \subseteq \widehat{\text{AC}}_j^{(s)}$ that $k \notin \text{PA}_j$. Consequently, we simply set the corresponding p-value $p^{(s)}(j, k) = 1$. Our subsequent testing procedure is to focus on the case where $k \in \widehat{\text{AC}}_j^{(s)}$, and we set $\mathcal{M}^{(s)} = \widehat{\text{AC}}_j^{(s)} - \{k\}$. We also remark that, to establish the consistency of our test, we only require

$\mathbb{P}(\text{PA}_j \subseteq \widehat{\text{AC}}_j^{(s)} \subseteq \text{DS}_j^c - \{j\}) \to 1$, where DS_j^c denotes the complement of the set DS_j. This essentially requires the order of the DAG to be consistently estimated. We later show in Section B.1 that this condition is satisfied when using the method of Zheng et al. (2020). Meanwhile, this order consistency is much weaker than requiring the DAG estimator $\widehat{\mathcal{G}}^{(s)}$ to be selection consistent, i.e., $\mathbb{P}(\text{PA}_j = \widehat{\text{PA}}_j) \to 1$, or to satisfy sure screening, i.e., $\mathbb{P}(\text{PA}_j \subseteq \widehat{\text{PA}}_j) \to 1$.

3.4 Supervised learning

The second key component of our test is to learn the conditional mean $g^{(s)}(x) = \mathbb{E}\left(X_j | X_{\mathcal{M}^{(s)}} = x\right)$. This is essentially a regression problem, and there are many choices, e.g., boosting, random forests, or neural networks. In our implementation, we use the MLP again, by seeking

$$\min_{\theta_j} \sum_{i \in \mathcal{I}_s} \sum_{t=1}^{T_i} \left\{ \mathbb{X}_{i,t,j} - \text{MLP}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}; \theta_j\right) \right\}^2, \tag{10}$$

where the learner $\text{MLP}(\cdot)$ is as defined in (8). The optimization problem in (10) can be solved using a stochastic gradient descent algorithm, or the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (Byrd et al., 1995).

3.5 Generative adversarial learning

The third key component of our test is to use GANs to learn a generator $\mathbb{G}^{(s)}(\cdot, \cdot)$, which generates a set of pseudo samples that have a similar distribution as the training samples. More accurately, in our setting, we learn the generator $\mathbb{G}(\cdot, \cdot)$ that takes $\mathbb{X}_{i,t,\mathcal{M}-\{k\}}$ and a set of multivariate Gaussian noise vectors as the input, and the output are a set of pseudo samples $\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}$. We train the generator such that the divergence between the conditional distribution of $\mathbb{X}_{i,t,k}$ given $\mathbb{X}_{i,t,\mathcal{M}-\{k\}}$ and that of $\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}$ given $\mathbb{X}_{i,t,\mathcal{M}-\{k\}}$ is minimized.

More specifically, we adopt Genevay et al. (2018) to learn the generator $\mathbb{G}^{(s)}$, by optimizing

$$\min_{\mathbb{G}} \max_{c} \widetilde{\mathcal{D}}_{c,\rho}(\mu, \nu), \tag{11}$$

where μ and ν denote the joint distribution of $\left(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)$ and $\left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}, \mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)$, respectively, and $\widetilde{\mathcal{D}}_{c,\rho}$ is the Sinkhorn loss function between two probability measures. The loss $\widetilde{\mathcal{D}}_{c,\rho}$

is with respect to a cost function c and a regularization parameter $\rho > 0$,

$$\begin{aligned}\widetilde{\mathcal{D}}_{c,\rho}(\mu,\nu) &= 2\mathcal{D}_{c,\rho}(\mu,\nu) - \mathcal{D}_{c,\rho}(\mu,\mu) - \mathcal{D}_{c,\rho}(\nu,\nu),\\ \mathcal{D}_{c,\rho}(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{x,y} \big\{c(x,y) - \rho H(\pi|\mu\otimes\nu)\big\}\pi(dx,dy),\end{aligned}$$

where $\Pi(\mu,\nu)$ is a set containing all probability measures π whose marginal distributions correspond to μ and ν, H is the Kullback-Leibler divergence, and $\mu\otimes\nu$ is the product measure of μ and ν. When $\rho = 0$, $\mathcal{D}_{c,0}(\mu,\nu)$ measures the optimal transport of μ into ν with respect to the cost function $c(\cdot,\cdot)$ (Cuturi, 2013). When $\rho \neq 0$, an entropic regularization is added to this optimal transport. As such, the objective function $\widetilde{\mathcal{D}}_{c,\rho}$ in (11) is a regularized optimal transport metric, where the regularization is to facilitate the computation, so that $\widetilde{\mathcal{D}}_{c,\rho}$ can be efficiently evaluated. Intuitively, the closer the two conditional distributions, the smaller the Sinkhorn loss. Therefore, maximizing $\widetilde{\mathcal{D}}_{c,\rho}$ with respect to the cost c learns a discriminator that can better discriminate μ and ν. On the other hand, minimizing the maximum cost with respect to the generator $\mathbb{G}$ makes the conditional distribution of $\widetilde{\mathbb{X}}^{(s,m)}_{i,t,k}$ given $\mathbb{X}_{i,t,\mathcal{M}^{(s)}}$ closer to that of $\mathbb{X}_{i,t,k}$ given $\mathbb{X}_{i,t,\mathcal{M}^{(s)}}$. This yields the minimax formulation in (11). In our implementation, we approximate the cost function c and the generator based on MLP (8). We approximate the distributions $\mu_{j,k}$ and $\nu_{j,k}$ in (11) by the empirical distributions of the data samples. We update the parameters in GANs by the Adam algorithm (Kingma and Ba, 2015).

We again make a few remarks. First, we choose the Gaussian noise as the input for GANs. We have found the performance of the generator is not overly sensitive to the choice of the distribution of the input noise. We present more discussion and some additional numerical results in Section B.3 of the Appendix. Besides, we choose GANs based on the Sinkhorn divergence loss to mitigate the potential bias of traditional GANs. Moreover, in addition to GANs, other deep generative learning approaches such as variational auto-encoders (Kingma and Welling, 2013) are equally applicable here. Second, we note that, based on the estimated conditional distribution from GANs, one can derive the joint distribution of all variables, then infer the corresponding DAG structure. However, this may be computational inefficient, due to the huge number of conditional dependence relations that must be learnt. Finally, we note

that, an alternative approach for this step is to separately apply a supervised learning method B times to estimate $\mathbb{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}$, for $b = 1, \ldots, B$. Nevertheless, when B is large, and in our implementation, $B = 2000$, this approach is computationally very expensive. Therefore, we choose the generative learning approach for this step.

4 Bidirectional Theory

In this section, we establish the asymptotic size and power of the proposed test. As a by-product, we also derive the oracle property of the DAG estimator produced by (9), which is needed to guarantee the validity of the test. In the interest of space, we report that result in Section B.1 of the Appendix. To simplify the theoretical analysis, we assume $T_1 = \ldots = T_n = T$. All the asymptotic results are derived when either the number of subjects N, or the number of time points T, diverges to infinity. Such results are new, provide useful theoretical guarantees for different types of applications, and are referred as the bidirectional theory.

We begin with a set of regularity conditions needed for the asymptotic consistency.

(C1) With probability approaching one, $\text{PA}_j \subseteq \widehat{\text{AC}}_j^{(s)} \subseteq \text{DS}_j^c - \{j\}$.

(C2) Suppose $\mathbb{E}\left|g^{(s)}\left(X_{\mathcal{M}^{(s)}}\right) - \widehat{g}^{(s)}\left(X_{\mathcal{M}^{(s)}}\right)\right|^2 = O\{(NT)^{-2\kappa_1}\}$ for some constant $\kappa_1 > 0$, and $\widehat{g}^{(s)}$ is uniformly bounded almost surely. Suppose $\mathbb{E}\sup_{\widetilde{B}\in\mathcal{B}}\left|\mathbb{P}\left\{X_k \in \widetilde{B}|X_{\mathcal{M}^{(s)}}\right\} - \mathbb{P}\left\{\mathbb{G}^{(s)}\left(X_{\mathcal{M}^{(s)}}, Z_{j,k}^{(m)}\right) \in \widetilde{B}|\ X_{\mathcal{M}^{(s)}}\right\}\right|^2 = O\{(NT)^{-2\kappa_2}\}$ for some constant $\kappa_2 > 0$, where $\mathcal{B}$ denotes the Borel algebra on $\mathbb{R}$. Suppose $\kappa_1 + \kappa_2 > 1/2$.

(C3) The random process $\{\mathbb{X}_{i,t}\}_{t\geq 0}$ is β-mixing if T diverges to infinity. The β-mixing coefficients $\{\beta(q)\}_q$ satisfy that $\sum_q q^{\kappa_3}\beta(q) < +\infty$ for some constant $\kappa_3 > 0$. Here, $\beta(q)$ denotes the β-mixing coefficient at lag q, which measures the time dependence between the set of variables $\{\mathbb{X}_{i,j}\}_{j\leq t}$ and $\{\mathbb{X}_{i,j}\}_{j\geq t+q}$.

(C4) Suppose the number of observations K in the batched standard error estimators $\widehat{\sigma}_{b,\text{CF}}^{(s)}$ and $\widehat{\sigma}_{b,\text{NCF}}^{(s)}$ satisfies that, $K = T$ if T is bounded, and $T^{(1+\kappa_3)^{-1}} \ll K \ll NT$ otherwise.

Condition (C1) concerns about the step of structural learning of DAG, which essentially requires the order of the DAG can be consistently estimated. We first remark that, this order

consistency is much weaker than the selection consistency. In other words, we only require a reasonably good initial DAG estimator that is order consistent, which is much easier to obtain than a DAG estimator that is selection consistent. In Section B.1, we show that (C1) holds when (9) is employed to estimate the DAG. Second, (C1) may not be a necessary condition to ensure the type-I error control. We next give two examples, where (C1) does not hold, but our proposed test can still control the type-I error. Moreover, in our simulation examples in Section 5, (C1) does not alway hold either. We report the percentage of times out of 500 data replications when (C1) holds for some selected nodes in Section B.4 of the Appendix. Nevertheless, our test still manages to achieve a competitive empirical performance. On the other hand, we keep (C1) in its current form, as it helps simplify the proof considerably.

Example 5 (missing parents)**.** We first consider an example where $\widehat{\text{AC}}_j^{(s)}$ misses some nodes in PA_j. The proposed test remains valid as long as these nodes have weak effects on X_j and X_k. More specifically, consider the five-variable example as illustrated in Figure 1(c). Our goal is to test whether there is a directed link from X_3 to X_4. Then $\text{PA}_j \subseteq \widehat{\text{AC}}_j^{(s)}$ requires that $\{1,2\} \subseteq \widehat{\text{AC}}_4^{(s)}$. Suppose X_1 has a weak effect on X_4, so that X_1 is not included in $\widehat{\text{AC}}_4^{(s)}$. Suppose $|\mathbb{E}(X_4|X_1,X_2) - \mathbb{E}(X_4|X_2)|^2 = O\{(NT)^{-2\kappa_1^*}\}$, for some $\kappa_1^* \geq \kappa_1$. When $\mathbb{E}\sup_{\widetilde{B}\in\mathcal{B}} |\mathbb{P}(X_3 \in \widetilde{B}|X_2) - \mathbb{P}(X_3 \in \widetilde{B}|X_1,X_2)| = O\{(NT)^{-2\kappa_2^*}\}$, for some $\kappa_2^* \geq \kappa_2$, under (C2)-(C4), the estimated conditional mean function and the distributional generator would converge to $\mathbb{E}(X_4|X_1,X_2)$ and $\mathbb{P}_{X_3|X_1,X_2}$ at the rate of $(NT)^{-\kappa_1}$ and $(NT)^{-\kappa_2}$, respectively. As such, the proposed test still works as if X_1 were included in $\widehat{\text{AC}}_4^{(s)}$.

Example 6 (including descendants)**.** We next consider an example where $\widehat{\text{AC}}_j^{(s)}$ includes some nodes in DS_j. The proposed test remains valid as long as none of these nodes is a descendant of X_k, or has a common descendant with X_k. In this case, X_k and X_j are d-separated given $\widehat{\text{AC}}_j^{(s)}$, as none of those falsely included nodes is a collider on any path between X_j and X_k; see the definition of d-separation and collider in Pearl (2009). As d-separation implies conditional independence, the proposed test is still able to control the type-I error. For the example in Figure 1(c), when $\{5\} \in \widehat{\text{AC}}_4^{(s)}$, (C1) is violated. However, when X_3 does not have affect X_5, the proposed test remains valid.

Condition (C2) concerns about the steps of learning the conditional mean function and the distribution generator. It requires the squared prediction loss of the supervised learner of the conditional mean, and the squared total variation norm between the conditional distributions of the observed and pseudo samples to satisfy some convergence rate, κ_1 and κ_2, respectively. We note that both estimators are nonparametric, and as such, both κ_1 and κ_2 can be slower than the parametric rate of $1/2$. However, (C2) only requires that $\kappa_1 + \kappa_2 > 1/2$. This is relatively easy to achieve when using the multilayer perceptron models and GANs, whose convergence rates have been established (see e.g., Schmidt-Hieber, 2017; Farrell et al., 2021; Liang, 2018; Bauer and Kohler, 2019; Chen et al., 2020). Moreover, we remark that, it is possible to further relax the requirement of $\kappa_1 + \kappa_2 > 1/2$ to $\kappa_1, \kappa_2 > 0$, by using the theory of higher order influence functions (Robins et al., 2017). However, the corresponding estimators would be considerably much more complicated, and thus we do not pursue those in this article.

Condition (C3) characterizes the dependence of the data observations over time, and is commonly imposed in the time series literature (Bradley, 2005). We also note that, (C3) is *not* needed when T is bounded but N diverges to infinity. Condition (C4) guarantees the consistency of the batched standard error estimators $\widehat{\sigma}^{(s)}_{b,\text{CF}}$ and $\widehat{\sigma}^{(s)}_{b,\text{NCF}}$, and is easily satisfied, since K is a parameter we specify. When T is bounded and is relatively small compared to a large sample size N, we can simply set $K = T$, i.e., treating the entire time series as one batch.

We next establish the asymptotic size of the propose testing procedure.

Theorem 2 (Size)**.** *Suppose model* (1)*, and conditions (C1)-(C4) hold. Suppose* $\min_b NT\, Var\Big(\widehat{I}^{(s)}_{b,CF} | \{\mathbb{X}_{i,t}\}_{i\in\mathcal{I}_s, 1\le t\le T}\Big) \ge \kappa_4$ *for some constant* $\kappa_4 > 0$*. If the constants* κ_1*,* κ_2*,* κ_3 *satisfy that* $\kappa_3 > \max[\{2\min(\kappa_1, \kappa_2)\}^{-1} - 1, 2]$*, then, as either* N *or* $T \to \infty$*,*

(a) The test statistic $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{CF}} \xrightarrow{d}$ *Normal*$(0, 1)$ *under* $H_0(j, k)$*.*

(b) The p-value satisfies that $\mathbb{P}\{p(j, k) \le \alpha\} \le \alpha + o(1)$*, for any nominal level* $0 < \alpha < 1$*.*

To establish the asymptotic size of the test, we require $\beta(q)$ to decay at a polynomial rate with respect to q. Such a condition holds for many common time series models (see, e.g., McDonald et al., 2015). We also require a minimum variance condition, which automatically holds

when the conditional variance of $h_b^{(s)}\big(X_k, X_{\mathcal{M}^{(s)}}\big) - \mathbb{E}\big\{h_b^{(s)}\big(X_k, X_{\mathcal{M}^{(s)}}\big)|X_{\mathcal{M}^{(s)}}\big\}$ given $X_{\mathcal{M}^{(s)}}$ is bounded away from zero. Under these conditions, we establish the asymptotic normality of the test statistic $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{CF}}$, which further implies that the p-value $p^{(s)}(j,k)$ converges to a uniform distribution on $[0,1]$. By Bonferroni's inequality, $p(j,k)$ is a valid p-value, and consequently, the proposed test achieves a valid control of type-I error.

Next, we study the asymptotic power of the test. We introduce a quantity to characterize the degree to which the alternative hypothesis deviates from the null for a given function class $\mathbb{H}$: $\Delta(\mathbb{H}) = \min_{\mathcal{M}} \sup_{h\in\mathbb{H}} |I(j,k|\mathcal{M};h)|$, where the minimum is taken over all subsets $\mathcal{M}$ that satisfy the conditions in Proposition 1. When $\mathbb{H}$ is taken over the class of characteristic functions of $(X_k, X_{\mathcal{M}})$, we have $\Delta(\mathbb{H}) > 0$. We also need the concept of the VC type class (Chernozhukov et al., 2014, Definition 2.1); see Section C.4 of the Appendix. To simplify the analysis, we suppose X_j is bounded, and without loss of generality, its support is $[0,1]$.

Theorem 3 (Power)**.** *Suppose the conditions in Theorem 2 hold, and the β-mixing coefficient $\beta(q)$ in (C3) satisfies that $\beta(q) = O(\kappa_5^q)$ for some constant $0 < \kappa_5 < 1$ when T diverges. Suppose $\Delta(\mathbb{H}) \gg (NT)^{-1/2}\log(NT)$ under $H_1(j,k)$. Suppose, with probability tending to one, $\widehat{g}^{(s)}$ and $\mathbb{G}^{(s)}$ belong to the class of VC type functions with bounded envelope functions and the bounded VC indices no greater than $O\{(NT)^{\min(2\kappa_1,2\kappa_2,1/2)}\}$, $s = 1,2$. If the number of transformation functions $B = \kappa_6(NT)^{\kappa_7}$ for some constants $\kappa_6 > 0, \kappa_7 \geq 1/2$, then, as either N or $T \to \infty$, $p(j,k) \xrightarrow{p} 0$ under $H_1(j,k)$.*

To establish the asymptotic power of the test, we require the function $\widehat{g}^{(s)}$ and the generator $\mathbb{G}^{(s)}$ to both belong to the VC type class. This is to help establish the concentration inequalities for the measure $\widehat{I}^{(s)}_{b,\text{NCF}}$ without cross-fitting. This condition automatically holds in our implementation where the MLP is used to model both (Farrell et al., 2021). We have also strengthened the requirement on $\beta(q)$, so that it decays exponentially with respect to q. This is to ensure the $\sqrt{NT}$-consistency of the proposed test when $T \to \infty$. This condition holds when the process $\{\mathbb{X}_{i,t}\}_{t\geq 0}$ forms a recurrent Markov chain with a finite state space. It also holds for more general state space Markov chains (see, e.g., Bradley, 2005, Section 3). Under these conditions,

Theorem 3 shows that our proposed test is consistent against some local alternatives that are $\sqrt{NT}$-consistent to the null up to some logarithmic term.

We remark that, Theorems 2 and 3 show that the proposed test controls the type-I error and achieves a parametric power guarantee, even though we estimate the three key components, the DAG structure, the conditional mean, and the distribution generator, all using fully nonparametric methods. This is achieved mainly due to the fact that our test statistic $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}$ is doubly robust, in that it is consistent as long as either the conditional mean or the distribution generator is correctly specified. Together with the Neyman orthogonality of the estimating equation, we show that the bias can be represented as a product of the difference between the two nonparametric estimators and their oracle values; see Step 3 of the proof of Theorem 2 in Section C.3 of the Appendix. Consequently, as long as $\kappa_1 + \kappa_2 > 1/2$, the test statistic converges at a parametric rate, and the test has a parametric power guarantee.

We also remark that, in our theory, the dimension d of the DAG is allowed to diverge to infinity with the sample size. Note that there is no explicit specification on d in the statements of Theorems 2 and 3. It is implicitly imposed due to the requirement that $\kappa_1 + \kappa_2 > 1/2$, as the convergence rates would become slower as the dimension d increases.

5 Simulations

In this section, we examine the finite-sample performance of the proposed testing procedure.

We begin with a discussion of some implementation details. Our test employs three neural networks-based learners, which involve numerous tuning parameters. Many of these parameters are common, e.g., the number of hidden layers and hidden nodes, the activation function, batch size, and epoch size, and we set them at the typical values recommended in the literature. For the DAG learning step, one tuning parameter is the sparsity parameter λ in (9). Following Zheng et al. (2020), we fix $\lambda = 0.025$ in our implementation to speed up the computation. We have also experimented with a number of values of λ and find the results are not overly sensitive. It can also be tuned via cross-validation. For the supervised learning step, we employ the multilayer perceptron regressor implementation of Pedregosa et al. (2011). For the

GANs training step, we follow the implementation of Genevay et al. (2018). There are three additional parameters associated with our test, including the number of transformation functions B, the number of pseudo samples M, and the number of observations K in the batched standard error estimators. We have found that the results are not sensitive to the choice of M and K, and we fix $M = 100$ and $K = 20$. For B, a larger value generally improves the power of the test, but also increases the computational cost. In our implementation, we set $B = 2000$, which achieves a reasonable balance between the test accuracy and the computational cost.

We compare the proposed test with two alternative solutions, the double regression-based test (DRT) as outlined in Section 2.2, and the constrained likelihood ratio test (LRT) proposed by Li et al. (2020) for linear DAGs. The implementation of DRT is similar to our proposed method. The main difference lies in that DRT uses the MLP regressor to first estimate the conditional mean function $\mathbb{E}(X_k|X_{\mathcal{M}_{j,k}^{(j)}})$ in Step 4, then plugs in this estimate to construct the test statistic in Step 5, with $B = 1$ and $h_1^{(s)}(X_k, X_{\mathcal{M}_{j,k}^{(j)}}) = X_k$.

We consider the following nonlinear DAG model,

$$X_{t,j} = \sum_{\substack{k_1,k_2 \in \mathrm{PA}_j \\ k_1 \le k_2}} c_{j,k_1,k_2} f^{(1)}_{j,k_1,k_2}(X_{t,k_1}) f^{(2)}_{j,k_1,k_2}(X_{t,k_2}) + \sum_{k_3 \in \mathrm{PA}_j} c_{j,k_3} f^{(3)}_{j,k_3}(X_{t,k_3}) + \varepsilon_{t,j}. \tag{12}$$

The data generation follows that of Zhu et al. (2020). Specifically, $f^{(1)}_{j,k_1,k_2}$, $f^{(2)}_{j,k_1,k_2}$, and $f^{(3)}_{j,k_3}$ in (12) are randomly set to be sine or cosine function with equal probability, whereas c_{j,k_1,k_2} and c_{j,k_3} are randomly generated from uniform $[0.5\delta, 1.5\delta]$ or $[-1.5\delta, -0.5\delta]$ with an equal probability, where $\delta > 0$ denotes some constant that controls the signal strength. The error $\varepsilon_{t,j}$ is an AR(1) process with the autoregressive coefficient equal to 0.5 and a standard normal white noise. The DAG structure is determined by a $d \times d$ lower triangular binary adjacency matrix, in which each entry is randomly sampled from a Bernoulli distribution with probability ζ. We vary four sets of key parameters in our simulations: (a) the number of subjects N from $\{10, 20, 40\}$; (b) the number of time points T from $\{50, 100, 200\}$; (c) the signal strength δ from $\{0.5, 1, 2\}$, and (d) the dimension d and the Bernoulli probability ζ from $(d, \zeta) = \{(50, 0.10), (100, 0.04), (150, 0.02)\}$. When we vary one set of the parameters, we keep the rest fixed at their default values of $N = 20, T = 100, \delta = 1, d = 50, \zeta = 0.10$.

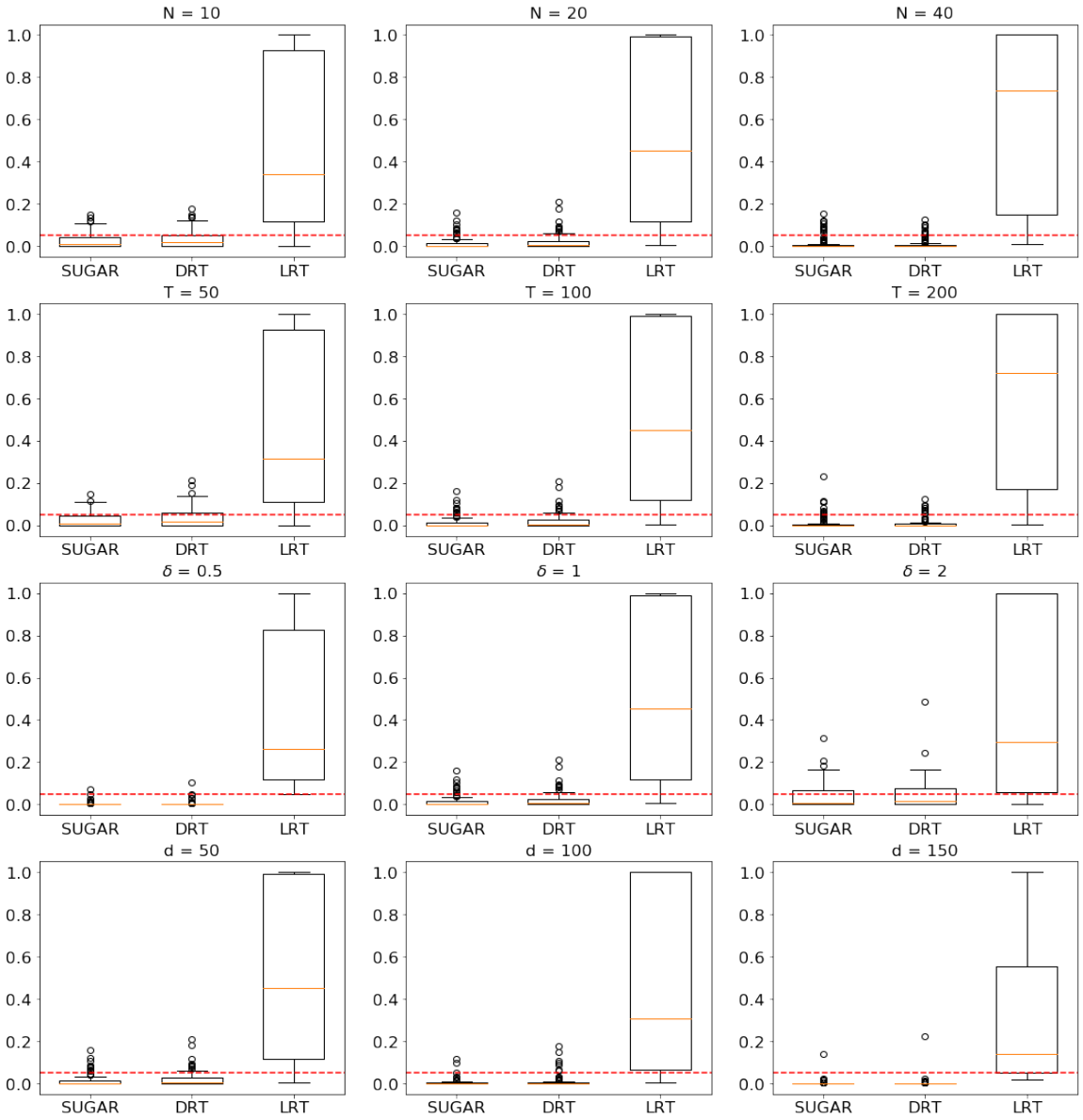

Figure 2: The boxplots of the empirical size of three methods: our proposed test (SUGAR), the double regression-based test (DRT), and the constrained likelihood ratio test (LRT), under four sets of varying parameters: first row $N = \{10, 20, 40\}$, second row $T = \{50, 100, 200\}$, third row $\delta = \{0.5, 1, 2\}$, and fourth row $(d, \zeta) = \{(50, 0.10), (100, 0.04), (150, 0.02)\}$.

For each scenario, we randomly sample 100 pairs of nodes where the null hypothesis holds, and another 100 pairs of nodes where the alternative hypothesis holds. We then apply the proposed test to these pairs, and record the empirical size and power of the test, i.e., the percentage of the times out of 200 data replications when the p-value is smaller than the nominal level

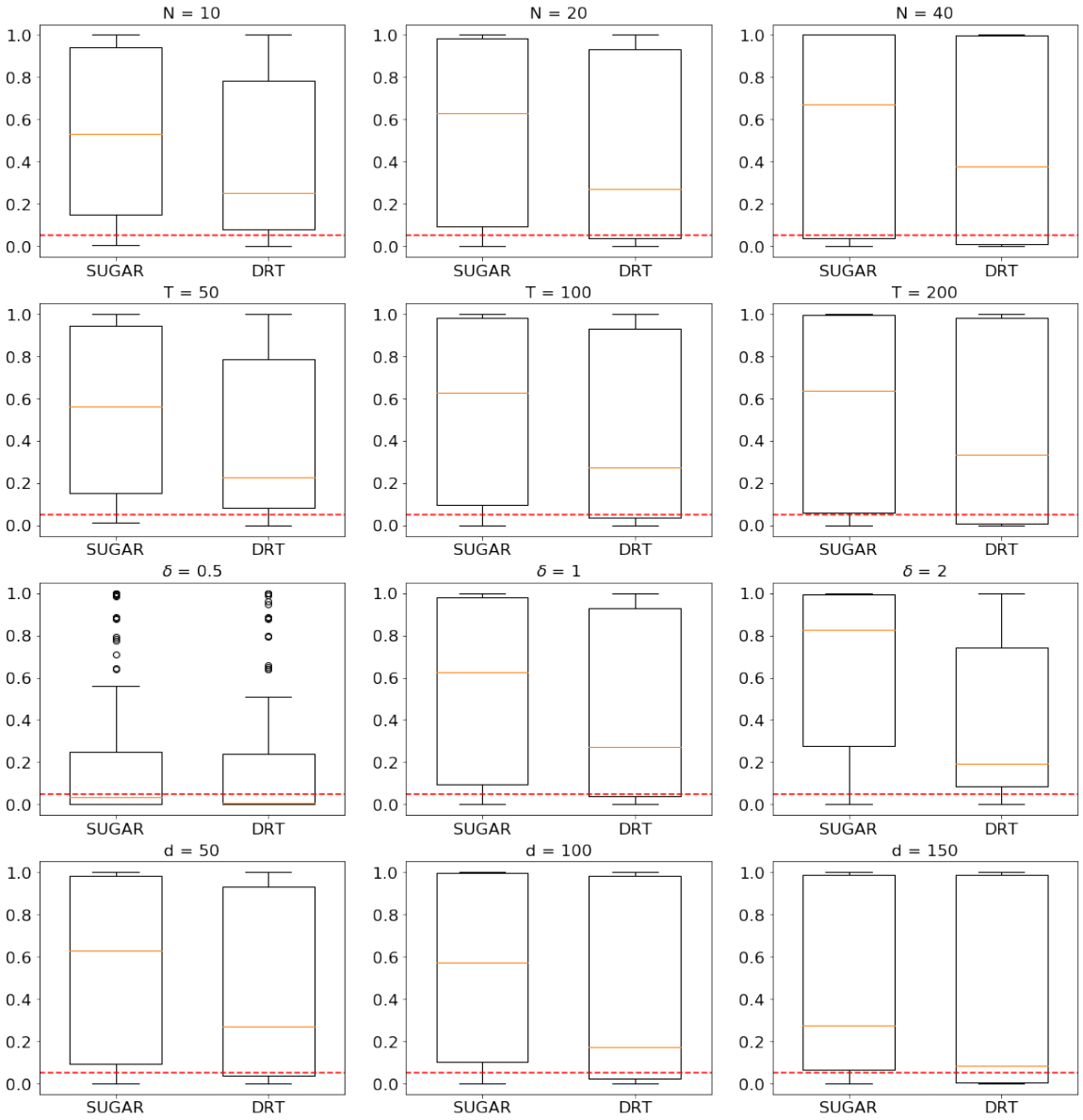

Figure 3: The boxplots of the empirical power of two methods: our proposed test (SUGAR), and the double regression-based test (DRT), under four sets of varying parameters: first row $N = \{10, 20, 40\}$, second row $T = \{50, 100, 200\}$, third row $\delta = \{0.5, 1, 2\}$, and fourth row $(d, \zeta) = \{(50, 0.10), (100, 0.04), (150, 0.02)\}$.

$\alpha = 0.05$. Figure 2 shows the boxplots of the empirical size for the pairs when the null holds, and Figure 3 shows the boxplots of the empirical power for the pairs when the alternative holds. We further report the difference of the powers of SUGAR and DRT in Figure 5 in Section B.5 of the Appendix. We do not report the power of LRT, because it fails to control the type-I error,

and thus its empirical power becomes meaningless. We make the following observations from these plots. In terms of the empirical size, both SUGAR and DRT manage to control the type-I error, but LRT does not. The reason is that LRT requires the graph to have a linear structure and the samples to be independent, but none is satisfied in our simulation model. On the other hand, in terms of the empirical power, SUGAR achieves generally a higher power than DRT, over 75% of the times in all scenarios as seen from Figure 5. Finally, as the key model parameters vary, the power of both SUGAR and DRT increases as the number of subjects N, or the number of time points T increases, since more data information becomes available, and the power of both tests decreases as the dimension d increases, since the graph becomes bigger and the problem more challenging. Meanwhile, the power of SUGAR increases as the signal strength δ increases, but that of DRT is not monotonic with respect to δ, because DRT is not guaranteed to be consistent in general, as we have commented earlier.

In terms of the computational time, our testing procedure consists of two main parts: the DAG estimation in Step 2 of Algorithm 1, and the rest in Steps 3 to 6. The DAG estimation is the most time consuming step, but it only needs to be learnt once for all pairs of edges in the graph. We implemented the DAG estimation step on the NVIDIA Tesla T4 GPU, and it took about 5 to 20 minutes when d ranges from 50 to 150 for one data replication. We implemented the rest of the testing procedure on the N1 standard CPU, and it took about 2 minutes for one data replication. A Python implementation of our method is available at `https://github.com/yunzhe-zhou/SUGAR`.

6 Brain Effective Connectivity Analysis

We next illustrate our method with a brain effective connectivity analysis of task-evoked functional magnetic resonance imaging (fMRI) data. The brain is a highly interconnected dynamic system, and it is of great interest to understand the relations among different brain regions through fMRI, which measures synchronized blood oxygen level dependent brain signals. The dataset we analyze is part of the Human Connectome Project (HCP, Van Essen et al., 2013), whose overarching objective is to understand brain connectivity patterns of healthy adults. We

Table 1: The number of identified significant within-module and between-module connections of the four functional modules for the low-performance and high-performance groups. The number of brain regions of each functional module is reported in the parenthesis.

	Auditory (13)		Default mode (58)		Visual (31)		Fronto-parietal (25)	
	low	high	low	high	low	high	low	high
Auditory (13)	20	17	0	0	0	1	2	0
Default mode (58)	0	0	68	46	3	2	11	23
Visual (31)	0	0	3	2	56	46	0	1
Fronto-parietal (25)	2	1	11	23	0	1	22	27

study the fMRI scans of a group of individuals who undertook a story-math task. The task consisted of blocks of auditory stories and addition-subtraction calculations, and required the participant to answer a series of questions. An accuracy score was given at the end. We analyze two subsets of individuals with matching age and sex. One set consists of $N = 28$ individuals who scored below 65 out of 100, and the other set consists of $N = 28$ individuals who achieved the perfect score of 100. All fMRI scans have been preprocessed following the pipeline of Glasser et al. (2013) that summarized each fMRI scan as a matrix of time series. Each row is a time series with length $T = 316$, and there are 264 rows corresponding to 264 brain regions (Power et al., 2011). Those brain regions are further grouped into 14 functional modules (Smith et al., 2009). Each module possesses a relatively autonomous functionality, and complex tasks are believed to perform through coordinated collaborations among the modules. In our analysis, we concentrate on $d = 127$ brain regions from four functional modules: auditory, visual, frontoparietal task control, and default mode, which are generally believed to be involved in language processing and problem solving domains (Barch et al., 2013).

We apply the proposed test to the two datasets separately. We control the false discovery at 0.05 using the standard Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Table 1 reports the number of identified significant within-module and between-module connec-

tions. We first note that, we identify many more within-module connections than the between-module connections. The partition of the brain regions into the functional modules has been fully based on the biological knowledge, and our finding lends some numerical support to this partition. In addition, we identify more within-module connections for the frontoparietal task control module for the high-performance subjects than the low-performance subjects, while we have identified fewer within-module connections for the default mode and visual modules for the high-performance subjects. These findings generally agree with the neuroscience literature. Particularly, the frontoparietal network is known to be involved in sustained attention, complex problem solving and working memory (Menon, 2011), and the high-performance group exhibits more active connections for this module. Meanwhile, the default mode network is more active during passive rest and mind-wandering, which usually involves remembering the past or envisioning the future rather than the task being performed (Van Praag et al., 2017), and the high-performance group exhibits fewer active connections for this module.

Acknowledgement

Li's research was partially supported by NSF grant CIF-2102227, and NIH grants R01AG061303, and R01AG062542. Shi's research was partially supported by EPSRC grant EP/W014971/1.

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034.

Barch, D. M., Burgess, G. C., et al. (2013). Function in the human connectome: Task-fmri and individual differences in behavior. *NeuroImage*, 80:169 – 189. Mapping the Connectome.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and

powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.*, 57:289–300.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *The Annals of Statistics*, 37(1):157–183.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Survey*, 2:107–144. Update of, and a supplement to, the 1986 original.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, 42(6):2526–2556.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–198.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208.

Cai, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application*, 4:423–446.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179.

Chakrabortty, A., Nandy, P., and Li, H. (2018). Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*.

Chen, M., Liao, W., Zha, H., and Zhao, T. (2020). Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv preprint arXiv:2002.03938*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.

Chickering, D. M., Heckerman, D., and Meek, C. (2004). Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Dedecker, J. and Louhichi, S. (2002). Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pages 137–159. Birkhäuser Boston, Boston, MA.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Friedrich, F., Kempe, A., Liebscher, V., and Winkler, G. (2008). Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224.

Friston, K. J. (2011). Functional and effective connectivity: A review. *Brain Connectivity*, 1(1):13–36.

Garg, R., Cecchi, G., and Rao, R. (2011). Full-brain auto-regressive modeling (farm) using fmri. *NeuroImage*, 58:416–41.

Genevay, A., Peyré, G., and Cuturi, M. (2018). Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617.

Glasser, M. F., Sotiropoulos, S. N., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Imaizumi, M. and Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In *The 22nd international conference on artificial intelligence and statistics*, pages 869–878. PMLR.

Janková, J. and van de Geer, S. (2019). Inference in high-dimensional graphical models. In *Handbook of graphical models*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 325–349. CRC Press, Boca Raton, FL.

Kalisch, M., Hauser, A., Maechler, M., Colombo, D., Entner, D., Hoyer, P., Hyttinen, A., Peters, J., Andri, N., Perkovic, E., et al. (2021). pcalg: Methods for graphical models and causal inference. *R Package retrieved from https://CRAN. R-project. org/package= pcalg*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kourogenis, N. and Pittis, N. (2011). Mixing conditions, central limit theorems, and invariance principles: a survey of the literature with some new results on heteroscedastic sequences. *Econometric Reviews*, 30(1):88–108.

Li, C. and Fan, X. (2019). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1489.

Li, C., Shen, X., and Pan, W. (2020). Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, 115(531):1304–1319.

Liang, T. (2018). On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*.

McDonald, D. J., Shalizi, C. R., and Schervish, M. (2015). Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9(2):2855–2883.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends in Cognitive Sciences*, 15(10):483–506.

Nandy, P., Maathuis, M. H., and Richardson, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, reasoning, and inference.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15:2009–2053.

Power, J. D., Cohen, A. L., et al. (2011). Functional network organization of the human brain. *Neuron*, 72(4):665–678.

Qiu, H., Han, F., Liu, H., and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society Series B.*, 78(2):487–504.

Rio, E. (2013). Inequalities and limit theorems for weakly dependent sequences. In *3rd cycle*, page 170. France.

Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., and van der Vaart, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987.

Romano, J. and DiCiccio, C. (2019). Multiple data splitting for testing. Technical report, Technical report.

Sachs, K., Perez, O., Peter, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Shi, C., Wan, R., Song, G., Luo, S., Song, R., and Zhu, H. (2020). *Spatiotemporal Causal Effects Evaluation: A Multi-Agent Reinforcement Learning Framework*. Under review.

Shi, C., Xu, T., Bergsma, W., and Li, L. (2021). Double generative adversarial networks for conditional independence testing. *The Journal of Machine Learning Research*, 22(1):13029–13060.

Siebert, W. M. (1986). *Circuits, signals, and systems*. MIT press.

Smith, S. D., Fox, P. T., Miller, K., Glahn, D., Fox, P., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A., and Beckmann, C. F. (2009). Correspondence of the brain; functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, 106:13040–5.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79.

Van Praag, C. D. G., Garfinkel, S. N., Sparasci, O., Mees, A., Philippides, A. O., Ware, M., Ottaviani, C., and Critchley, H. D. (2017). Mind-wandering and alterations to default mode network connectivity when listening to naturalistic versus artificial sounds. *Scientific Reports*, 7:45273.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. (2016). An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Frontiers in Neuroscience*, 10:1–17.

Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163.

Yuan, Y., Shen, X., Pan, W., and Wang, Z. (2019). Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika*, 106(1):109–125.

Zhang, H., Zhou, S., and Guan, J. (2018). Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. P. (2020). Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*.

Zhu, S., Ng, I., and Chen, Z. (2020). Causal discovery with reinforcement learning. In *International Conference on Learning Representations*.

In this appendix, Section A discusses several extensions of the proposed test. Section B presents additional theoretical and numerical results. Section C gives the detailed proofs.

A Extensions

In the article, we have primarily focused on testing a particular pair of nodes (j, k) in the DAG model, $j, k = 1, \ldots, d$. Next, we discuss the extensions to test a directed pathway, a union of directed edges, and the categorical X_j following a generalized linear model. We also outline the extensions to the Markov equivalence class, and non-stationary and time-varying DAG.

A.1 Extension to a directed pathway

Suppose our goal is to test a given directed pathway, $j_1 \to j_2 \to \ldots \to j_K$, where $j_1, j_2, \ldots, j_K$ are a sequence of nodes in the DAG. The problem can be formulated as the pair of hypotheses:

$$\begin{aligned} &H_{p0}: \ H_0(j_k, j_{k+1}) \text{ holds for some } k, \quad \text{versus} \\ &H_{p1}: \ H_0(j_k, j_{k+1}) \text{ does } \textit{not} \text{ hold for any } k = 1, \ldots, d. \end{aligned} \tag{13}$$

Under the alternative, each individual null hypothesis $H_0(j_k, j_{k+1})$ does not hold, and thus there exists such a directed pathway. The hypotheses in (13) can be tested using the union-intersection principle. Specifically, let $p(j_k, j_{k+1})$ denote the p-value for $H_0(j_k, j_{k+1})$ from the proposed test. Then it is straightforward to show that $\max_k p(j_k, j_{k+1})$ is a valid p-value for (13). Based on Theorems 2 and 3, we can also show that such a test is consistent.

A.2 Extension to a union of directed edges

Suppose our goal is to test a union of the hypotheses $\cup_{l \in \mathcal{L}} H_0(j_l, k_l)$. We first apply the proposed test to construct two standardized measures, $\widehat{T}^{(s)}_{b,\text{CF}}(j_l, k_l)$ and $\widehat{T}^{(s)}_{b,\text{NCF}}(j_l, k_l)$, with and without cross-validation, for each $b = 1, \ldots, B, s = 1, 2$, and $l \in \mathcal{L}$. Then for each s, we select the indices $\widehat{b}^{(s)}$ and $\widehat{l}^{(s)}$ that yield the largest measure $\max_{b,l} |\widehat{T}^{(s)}_{b,\text{NCF}}(j_l, k_l)|$ in the absolute value. We then construct the Wald type test statistic $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\left(j_{\widehat{l}^{(s)}}, k_{\widehat{l}^{(s)}}\right)$. Based on Theorems 2 and 3, we can establish the consistency of this test.

A.3 Extension to generalized linear model

We can further extend the proposed test to the following class of models:

$$\mathbb{E}(X_j|X_{\text{PA}_j}) = \phi_j\left\{f_j(X_{\text{PA}_j})\right\}, \quad \text{for any } j = 1, \ldots, d,$$

where the link function ϕ_j is pre-specified while the function f_j is unspecified. For instance, when X_j is binary, we may set ϕ_j as the logistic function. Similar to Theorem 1, we can show that the null hypothesis in (4) is equivalent to $I(j, k|\mathcal{M}; h) = 0$, for all square-integrable function h. Therefore, the proposed test can be applied to this class of models as well.

A.4 Extension to Markov equivalence class

In the article, we have mainly focused on the case when the underlying DAG is identifiable. In this section, we discuss the extension to the Markov equivalence class. We first outline the key steps of the extension, then consider a way to expedite the computation. We further discuss the relation between our test and the DAGs in the equivalence class. Meanwhile, we leave the full investigation of the inference for the equivalence class as future research.

Outline of the extension: Suppose there exists an equivalence class of DAGs that could generate the same joint distribution of the variables. Such a class can be uniquely represented by a completed partially directed acyclic graph (CPDAG). For each DAG $\mathcal{G}$ that belongs to the equivalence class, we define $\text{PA}_j(\mathcal{G})$ as the set of parents of node j in $\mathcal{G}$. Then, we aim to test the hypotheses:

$$\begin{aligned} &H_{e0}(j,k):\ k \notin \text{PA}_j(\mathcal{G}), \quad \text{versus} \\ &H_{e1}(j,k):\ k \in \text{PA}_j(\mathcal{G}'), \quad \text{for some } \mathcal{G}' \text{ that belongs to the equivalence class.} \end{aligned} \tag{14}$$

To test the hypotheses in (14), we first estimate the equivalence class given each half of the data. Next, for each DAG $\mathcal{G}$ that belongs to the estimated equivalence class, we employ supervised learning and generative adversarial learning to compute the standardized measures, $\{\widehat{T}^{(s)}_{b,\text{CF}}(\mathcal{G})\}_{b=1}^{B}$, and $\{\widehat{T}^{(s)}_{b,\text{NCF}}(\mathcal{G})\}_{b=1}^{B}$. We then select the index $(\widehat{b}^{(s)}, \widehat{\mathcal{G}}^{(s)})$ that maximizes $|\widehat{T}^{(s)}_{b,\text{NCF}}(\mathcal{G})|$, and take $|\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{NCF}}(\widehat{\mathcal{G}}^{(s)})|$ as the final test statistic. Finally, we compute the p-value as $p(j,k) = 2\min\left[\ \Phi\left\{Z_0 > |\widehat{T}^{(1)}_{\widehat{b}^{(1)},\text{NCF}}(\widehat{\mathcal{G}}^{(1)})|\right\}, \Phi\left\{Z_0 > |\widehat{T}^{(2)}_{\widehat{b}^{(2)},\text{NCF}}(\widehat{\mathcal{G}}^{(2)})|\right\}\ \right]$, where Z_0 is a

standard normal variable. This testing procedure is similar as Algorithm 1, except that the index is now selected among all possible pairs of $(b, \mathcal{G})$, whereas the index is selected among b only in Algorithm 1.

We can show the above test is consistent, following a similar approach as the test for an identifiable DAG in Section 4. We remark that, to establish the type-I error control, we only require each DAG estimator in the estimated equivalence class to be order consistent to some DAG in the true equivalence class. By contrast, to establish the power guarantee, we further require a one-to-one correspondence between the estimated and the true equivalence class.

Computation acceleration: When the graph is large, we recognize that it is computationally intensive to enumerate *all* the DAGs within the equivalence class. To accelerate the computation, we propose to focus on those DAGs that are only "locally" different.

Specifically, we first observe that our proposed algorithm depends on the estimated DAG $\mathcal{G}$ only through the index set $\mathcal{M} = \widehat{\text{AC}}_j(\mathcal{G}) - \{k\}$. As such, we can speed up the computation by directly calculating the multi-set of the ancestor sets,

$$\widetilde{\text{AC}}_{j,k} = \left\{ \widehat{\text{AC}}_j(\mathcal{G}) : \mathcal{G} \text{ that belongs to the equivalence class and } k \in \widehat{\text{AC}}_j(\mathcal{G}) \right\}.$$

Moreover, for a graph $\mathcal{G}$, denote a subset of its estimated ancestor set $\widehat{\text{AC}}_j(\mathcal{G})$ up to G generations by $\widehat{\text{AC}}_j^{(G)}(\mathcal{G})$. For instance, $\widehat{\text{AC}}_j^{(1)}(\mathcal{G})$ denotes all the estimated parent nodes, and $\widehat{\text{AC}}_j^{(2)}(\mathcal{G})$ denotes all the estimated parent and grandparent nodes. Along with some other mild conditions, if the following condition holds,

$$\text{PA}_j(\mathcal{G}) \subseteq \widehat{\text{AC}}_j^{(G)}(\mathcal{G}), \tag{15}$$

then the corresponding test remains to be consistent. On the other hand, while the ancestor sets of two DAGs may not be completely the same, their ancestor sets up to certain generations, e.g., the parent sets or the grandparent sets, may be the same. This motivates us to consider the following multi-set to further speed up the computation,

$$\widetilde{\text{AC}}_{j,k}^{(G)} = \left\{ \widehat{\text{AC}}_j^{(G)}(\mathcal{G}) : \mathcal{G} \text{ that belongs to the equivalence class and } k \in \widehat{\text{AC}}_j^{(G)}(\mathcal{G}) \right\}.$$

Correspondingly, the number of elements in $\widetilde{\text{AC}}_{j,k}^{(G)}$ can potentially be much smaller than that of $\widetilde{\text{AC}}_{j,k}$. In other words, we focus on the ancestors of node j for the graphs in the equivalence

class up to G generations only, instead of all the generations. Here G represents a trade-off between the computational cost and the sufficient condition to ensure the consistency of the test. When G is large, it is easier for the condition (15) to hold, but it is computationally more expensive. When G is small, it is harder for (15) to hold, but it allows us to focus on the DAGs that are only "locally" different around the link (j, k), and thus accelerates the computation.

To implement the above idea, we first use each half of the data to obtain a CPDAG. This can be achieved by directly applying some existing structural learning method, e.g., the PC algorithm (Spirtes et al., 2000), or by first applying the method in Section 3.3, then converting the learnt DAG to a CPDAG (Kalisch et al., 2021). Next, based on the estimated CPDAG, we select those nodes that are ancestors of j up to G generations. Let $\mathcal{N}^{(G)}$ denote these nodes. We then apply Algorithm 3 of Nandy et al. (2017) to obtain the multi-set of the parent sets of $\mathcal{N}^{(G)} \cup \{j\}$,

$$\left\{ \left\{ \widehat{\text{PA}}_l(\mathcal{G}) : l \in \mathcal{N}^{(G)} \cup \{j\} \right\} : \mathcal{G} \text{ that belongs to the equivalence class and } k \in \widehat{\text{AC}}_j^{(G)}(\mathcal{G}) \right\}.$$

For each $\mathcal{G}$, the parent set of $\mathcal{N}^{(G)} \cup \{j\}$, i.e., $\left\{ \widehat{\text{PA}}_l(\mathcal{G}) : l \in \mathcal{N}^{(G)} \right\}$ essentially contains all parents for each node in $\mathcal{N}^{(G)} \cup \{j\}$, based on which we can derive $\widehat{\text{AC}}_j^{(G)}(\mathcal{G})$, and subsequently $\widetilde{\text{AC}}_{j,k}^{(G)}$. Nandy et al. (2017) and Chakrabortty et al. (2018) noted that it is much more computationally efficient to obtain the multi-set than to enumerate all DAGs.

Equivalence class: We remark that our proposed test is built upon testing the conditional independence, and can test if a link exists in a DAG in an equivalence class. However, our test is generally not able to distinguish different DAGs in an equivalence class. We consider the following example to further elaborate.

Example 7 (Equivalence class)**.** Consider three DAGs depicted in Figure 4. All three DAGs have the same skeleton, none has colliders, and thus they belong to the same equivalence class following Verma and Pearl (1990). Each DAG has three variables, which are all binary, and

are generated in the following three ways for the three DAGs, respectively:

$$
\begin{aligned}
\mathcal{G}_1: \quad & \mathbb{P}(X_2=1)=p_0, \quad \mathbb{P}(X_1=X_2|X_2)=p_1, \quad \mathbb{P}(X_3=X_1|X_1)=p_2;\\
\mathcal{G}_2: \quad & \mathbb{P}(X_1=1)=p_0p_1+(1-p_0)(1-p_1), \quad \mathbb{P}(X_3=X_1|X_1)=p_2,\\
& \mathbb{P}(X_2=X_1|X_1)=\begin{cases} \dfrac{p_0p_1}{p_0p_1+(1-p_0)(1-p_1)}, & \textit{if } X_1=1,\\ \dfrac{(1-p_0)p_1}{(1-p_0)p_1+p_0(1-p_1)}, & \textit{otherwise;}\end{cases}\\
\mathcal{G}_3: \quad & \mathbb{P}(X_3=1)=p_0p_2+(1-p_0)(1-p_2),\\
& \mathbb{P}(X_1=X_3|X_3)=\begin{cases} \dfrac{p_0p_2}{p_0p_2+(1-p_0)(1-p_2)}, & \textit{if } X_1=1,\\ \dfrac{(1-p_0)p_2}{(1-p_0)p_1+p_0(1-p_2)}, & \textit{otherwise,}\end{cases}\\
& \mathbb{P}(X_2=X_1|X_1)=\begin{cases} \dfrac{p_0p_1}{p_0p_1+(1-p_0)(1-p_1)}, & \textit{if } X_1=1,\\ \dfrac{(1-p_0)p_1}{(1-p_0)p_1+p_0(1-p_1)}, & \textit{otherwise;}\end{cases}
\end{aligned}
$$

for some $p_0 \in (0,1)$, $p_1, p_2 \in (0, 0.5) \cup (0.5, 1)$. It can be shown that (X_1, X_2, X_3) has the same likelihood function, and the three DAGs are not identifiable.

Suppose we test whether there is an edge from X_1 to X_2, i.e., we test the hypotheses in (2) with $j = 2, k = 1$. We first apply the structural learning to estimate the DAG. When the estimated DAG equals $\mathcal{G}_1$, since X_1 is not in the ancestor set of X_2, following Step 2b of Algorithm 1, our test returns the p-value of 1 directly, and thus would not reject the null hypothesis. When the estimated DAG equals $\mathcal{G}_2$, since the ancestor set of X_2 contains X_1, while $\mathcal{M} = \widehat{\text{AC}}_j - \{k\} = \emptyset$, the problem becomes testing the marginal independence between X_1 and X_2. Following Steps 3 to 6 of Algorithm 1, our test would reject the null, as there is a link from X_1 to X_2. When the estimated DAG equals $\mathcal{G}_3$, since the ancestor set of X_2 contains both X_1 and X_3, and $\mathcal{M} = \widehat{\text{AC}}_j - \{k\} = \{3\}$, the problem becomes testing the conditional independence between X_1 and X_2 given X_3. Again, following Steps 3 to 6 of Algorithm 1, our test would reject the null. In this example, we are not able to differentiate $\mathcal{G}_2$ and $\mathcal{G}_3$ in the equivalence class from our testing result alone. Even though the testing result is different when the estimated DAG equals $\mathcal{G}_1$, we still do not know if the estimated DAG corresponds to the true DAG in the equivalence class where the data is generated from.

Therefore, without specific distributional assumptions, it is generally impossible to distin-

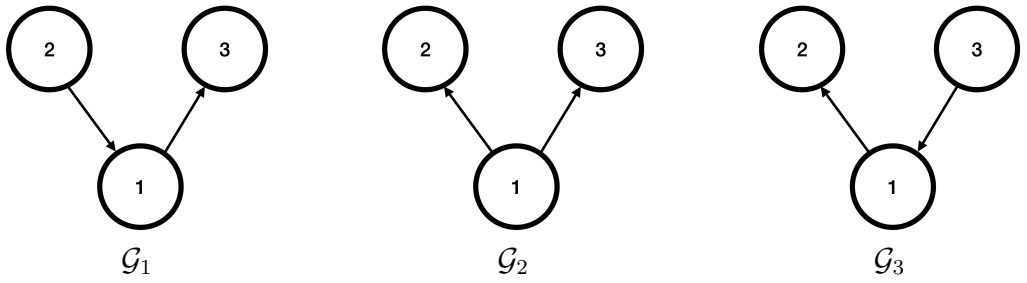

Figure 4: Three DAGs that belong to the same equivalence class, and each with three variables.

guish DAGs in an equivalence class, and our test alone cannot either. The main reason is that there is no way to tell if the estimated DAG actually corresponds to the true DAG. Although our test result depends on the estimated DAG, or say, the estimated ordering of the nodes, it is independent of the true DAG that generates the data.

A.5 Extension to non-stationary and time-varying DAG

In Section 2.3, we have focused on the case when DAG is stationary, as imposed by condition (B2). We have also excluded the case when DAG is time-varying, as implied by condition (B3). In this section, we again outline the key steps of extensions, first to non-stationary DAG, then to time-varying DAG. We leave the full investigation as possible future research. To simplify the presentation, we assume $T_1 = T_2 = \cdots = T_N = T$. In addition, we denote the random variable X_j at time t as $X_{j,t}$, for $j = 1, \ldots, d, t = 1, \ldots, T$.

We first consider a non-stationary DAG, and relax the stationarity condition (B2). Toward that end, suppose the DAG structure is piecewise constant over time. That is, there exist some change points, $1 = \tau_1 < \tau_2 < \ldots < \tau_M = T$, such that the random vectors $\mathbb{X}_{i,\tau_m}, \mathbb{X}_{i,\tau_m+1}, \ldots, \mathbb{X}_{i,\tau_{m+1}-1}$ are stationary for any $m = 1, \ldots, M-1$. Then, our goal is to test if there exists a directed edge from $X_{k,t}$ to $X_{j,t}$, for some $\tau_m \leq t < \tau_{m+1}$.

To test the hypotheses, we first estimate the change point locations and the graph structures

given each half of the data. We consider the following optimization,

$$\min_{\theta} \sum_{j=1}^{d} \left[\sum_{i \in \mathcal{I}_s} \sum_{m=1}^{M} \sum_{t=\tau_m}^{\tau_{m+1}-1} \left\{ \mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t}; \theta_{j,m}) \right\}^2 + \sum_{m} \frac{\lambda n_s (\tau_m - \tau_{m-1})}{T} \left\| A_{j,m}^{(1)} \right\|_{1,1} \right] + \gamma n_s M, \quad \text{subject to } \text{trace}[\exp\{W_m(\theta) \circ W_m(\theta)\}] = d, \tag{16}$$

for all $m = 1, \ldots, M$, where $\theta_{j,m} = \{A_{j,m}^{(l)}, b_{j,m}^{(l)}\}_l$ denotes the parameters in MLP that models the conditional mean function of $\mathbb{X}_{i,t,j}$ when t belongs to the time interval $[\tau_m, \tau_{m+1})$, and $W_m(\theta)$ is a $d \times d$ matrix whose (k, j)th entry equals the Euclidean norm of the kth column of $A_{j,m}^{(1)}$. The first penalty in (16) is placed on $\|A_{j,m}^{(1)}\|_{1,1}$ and is to impose the sparsity structure on the estimated DAG. The second penalty in (16) is placed on M, and is to penalize the total number of change points. Dynamic programming method such as Friedrich et al. (2008) can be employed to solve the optimization problem (16). Let $\widehat{\tau}_m$ denote the estimated change point locations, and $\widehat{\mathcal{G}}_{\widehat{\tau}_m}^{(s)}$ denote the estimated graphs, $m = 1, \ldots, \widehat{M}$, where $\widehat{M}$ denotes the corresponding estimator for M. Let $\widehat{\text{AC}}_j^{(s)}$ denote the set of ancestors of j based on $\widehat{\mathcal{G}}_{\widehat{\tau}_m}^{(s)}$, and $\mathcal{M}^{(s)} = \widehat{\text{AC}}_j^{(s)} - \{k\}$. We apply Steps 3 to 6 of Algorithm 1 to $\{\mathbb{X}_{i,t}\}_{1 \le i \le N, \widehat{\tau}_m \le t < \widehat{\tau}_{m+1}}$, and derive the corresponding p-value.

We can again show that the above test is consistent. This is based on the following key observation. Under the piecewise stationary structure, the number of change points can be consistently estimated, and the estimated change point locations converge at a faster rate than the estimated DAG. This phenomenon is well-known in the time series literature (see e.g., Boysen et al., 2009), where the estimated change point converges at a rate of $O_p(n^{-1} \log n)$, and this rate is much faster than the parametric rate. As a consequence, our test is to behave as well as if the true change point locations were known in advance.

Next, we briefly consider a time-varying DAG, which allows to test directed links from past to future observations. Suppose at time t, a given node not only depends on other nodes at the same time, but also on past variables at time $t-1, t-2, \cdots, t-Q$ as well. Our goal is to test if there exists a directed edge from $X_{k,t-q}$ to $X_{j,t}$, for some $0 \le q \le Q$. We can essentially apply Algorithm 1 to this problem, and can establish the consistency of the test similarly.

B Additional Results

B.1 Oracle property of the DAG learner

As a by-product of our theoretical analysis, we derive the oracle property of the DAG estimator produced by (9). This result is to guarantee $\mathbb{P}\left(\cap_{j\in\{1,\cdots,d\}}\{\text{PA}_j \subseteq \widehat{\text{AC}}_j^{(s)}\}\right) \to 1$, which was not available in Zheng et al. (2020). It implies that the ordering of the true DAG can be consistently estimated, which in turn ensures the validity of (C1). In this section, for simplicity, we assume the DAG dimension d is fixed. Nevertheless, we can extend our proof to the high-dimensional setting in a relatively straightforward fashion, by imposing a certain Hölder smoothness assumption on $\{f_j\}_j$; see, e.g., Farrell et al. (2021, Assumption 2).

We first define the oracle estimator. For an ordering $\pi = (\pi_1, \ldots, \pi_d)$ for a given DAG, consider the estimator $\widetilde{\theta}^{(s)}(\pi) = \{\widetilde{\theta}_1^{(s)}(\pi), \ldots, \widetilde{\theta}_d^{(s)}(\pi)\}$, where each $\widetilde{\theta}_j^{(s)}(\pi)$ is obtained by

$$\underset{\theta_j=\left\{A_j^{(1)},b^{(1)},\ldots,A_j^{(L)},b^{(L)}\right\}\text{supp}\left(A_j^{(1)}\right)\in\{\pi_1,\ldots,\pi_{j-1}\}}{\arg\min} \sum_{i\in\mathcal{I}_s}\sum_{t=1}^{T}\left\{\mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t};\theta_j)\right\}^2 + \frac{\lambda NT}{2}\left\|A_j^{(1)}\right\|_{1,1},$$

where $\text{supp}\left(A_j^{(1)}\right) \in \{\pi_1, \ldots, \pi_{j-1}\}$ means that, for any l that does not belong to this set, the lth column of $A_j^{(1)}$ equals zero. In other words, the estimator $\widetilde{\theta}_j^{(s)}(\pi)$ is computed as if the order π were known in advance.

Next, let Π^* denote the set of all true orderings. This means, for any true ordering $\pi^* \in \Pi^*$, $\text{PA}_j \subseteq \{\pi_1^*, \ldots, \pi_{j-1}^*\}$, for any $j = 1, \ldots, d$. In other words, the parents of each node should appear before the occurrence of this node under π^*. It is also worth mentioning that, the true ordering is *not* necessarily unique, even though the underlying DAG is unique. For instance, consider Example 4 with a v-structure as shown in Figure 1(a). In this example, both $(1, 2, 3)$ and $(1, 3, 2)$ are the true orderings, as there are no directional edges between nodes X_2 and X_3.

Next, we introduce some additional conditions. For any ordering π, define a least squares loss function, $\mathcal{L}(\pi) = \sum_{j=0}^{d-1}\mathbb{E}\left\{X_{j+1} - \mathbb{E}\left(X_{j+1}|X_{\{\pi_1,\ldots,\pi_j\}}\right)\right\}^2$. Moreover, we focus on neural networks with a ReLU activation function, $\sigma(x) = \max(0, x)$.

(C5) All minimizers of $\mathcal{L}(\pi)$ are contained in Π^*.

(C6) The widths of all layers in the MLP share a common asymptotic order H. Besides,

the number of layers L and the asymptotic order H diverge with NT, in that $HL = O\{(NT)^{\kappa_8}\}$, for some constant $\kappa_8 < 1/2$.

(C7) Suppose $\text{MLP}\big\{\cdot; \widetilde{\theta}^{(s)}(\pi)\big\}$ is bounded for any π.

Condition (C5) is reasonable and holds in numerous scenarios. One example is when all the random errors $\{\varepsilon_j\}_{j=1}^d$ in model (1) are normally distributed with equal variance. In that case, the least squares loss $\mathcal{L}$ is proportional to the expected value of the log-likelihood of X. Since the underlying DAG is identifiable, any ordering that minimizes the expected log-likelihood belongs to Π^*. Condition (C6) is also mild, as both H and L are the parameters that we specify. The part that $HL = O\{(NT)^{\kappa_8}\}$ ensures that the stochastic error resulting from the parameter estimation in the MLP is negligible. Condition (C7) ensures that the optimizer would not diverge in the ℓ_∞ sense. Similar assumptions are common in the literature to derive the convergence rates of deep learning estimators (see e.g. Farrell et al., 2021).

Now we show that the estimator $\widehat{\theta}^{(s)}$ obtained from (9) satisfies the oracle property, i.e., $\widehat{\theta}^{(s)} = \widetilde{\theta}^{(s)}(\pi^*)$, for some $\pi^* \in \Pi^*$. In other words, $\widehat{\theta}^{(s)}$ is computed as if one of the true ordering were known in advance. By the definition of Π^*, Condition (C1) holds for our estimated DAG. Moreover, we note that the oracle property does *not* imply the selection consistency, i.e., $\text{PA}_j = \widehat{\text{PA}}_j$, nor the sure screening property, in that $\text{PA}_j \subseteq \widehat{\text{PA}}_j$, for any $j = 1, \ldots, d$.

Theorem 4. *Suppose $\{f_j\}_j$ in model* (1) *are a set of continuous functions, (C5)-(C7) hold, the β-mixing coefficient $\beta(q)$ in (C4) decays exponentially with q, and $\lambda \to 0$. Then, with probability approaching one, $\widehat{\theta}^{(s)} = \widetilde{\theta}^{(s)}(\pi^*)$, for some $\pi^* \in \Pi^*$, as either N or $T \to \infty$.*

B.2 Sample splitting

We employ the data splitting and cross-fitting strategy for our test, and use a binary-split in Section 3. To mitigate sample randomization arising from a single binary-split, in this section, we develop a version of our test based on multiple binary-splits. The main idea is to apply the binary-split in Algorithm 1 multiple times, then combine the p-values from all splits. In addition, we may also adopt the multi-split strategy of Romano and DiCiccio (2019). These modifications may help reduce the sampling randomization, and may potentially improve the

power of the test, but also come with a price of increased computations. Specifically, we carry out the binary-split R times. For the rth binary-split, we randomly split all samples $\{1, \ldots, N\}$ into two disjoint subsets $\mathcal{I}_{r,1} \cup \mathcal{I}_{r,2}$ of equal sizes. We then apply Algorithm 1 to compute the p-values, $\widehat{p}^{(r,1)}$ and $\widehat{p}^{(r,2)}$, respectively, for each half of the data. We next combine these p-values by,

$$\widehat{p} = \min\Big(1, q_\gamma\left[\{\gamma^{-1}\widehat{p}^{(r,s)}(0,q), r = 1, \ldots, R, s = 1, 2\}\right]\Big),$$

where $0 < \gamma < 1$ is a constant, and q_γ is the empirical γ-quantile. We recommend to set γ to a small value, such as 0.1 or 0.2. This follows a similar idea as Meinshausen et al. (2009).

B.3 Gaussian versus non-Gaussian input noise for GANs

When learning the distribution generator in Section 3.5, we take the Gaussian noise as the input of GANs. One may also use other non-Gaussian noises, e.g., uniformly distributed random vectors over a unit hypercube. In general, the performance of the generator computed via GANs is not overly sensitive to the choice of the distribution of the input noise. This is partly because, the objective of the GAN step is to learn a generator $\mathbb{G}$, such that the conditional distribution of X_k given $X_{\mathcal{M}^{(s)}}$ can be well approximated by that of $\mathbb{G}(X_{\mathcal{M}^{(s)}}, Z_{j,k})$ given $X_{\mathcal{M}^{(s)}}$, where $Z_{j,k}$ is the Gaussian noise. Suppose we use some non-Gaussian noise $V_{j,k}$ with the same dimension. Under some regularity conditions, there exists a transformation function ϕ, such that $\phi(V_{j,k})$ has the same distribution as $Z_{j,k}$. Define $\mathbb{G}_\phi(X_{\mathcal{M}^{(s)}}, V_{j,k}) = \mathbb{G}(X_{\mathcal{M}^{(s)}}, \phi(V_{j,k}))$. Then, $\mathbb{G}_\phi$ has the same smoothness properties as $\mathbb{G}$. As such, the estimated distribution generator for $\mathbb{G}_\phi$ is expected to have similar statistical properties as that for $\mathbb{G}$ (Chen et al., 2020).

We also conduct a simulation to examine the empirical performance of our test under two distributions, Gaussian and uniform, for the input noise. We adopt the nonlinear model (12) in Section 5, with $N = 20, T = 100, \delta = 1, d = 50, \zeta = 0.1$. Table 2 reports the empirical size and power, i.e., the percentage of times out of 500 data replications when the p-value is smaller than the nominal level $\alpha = 0.05$ and $\alpha = 0.10$, respectively, for some pairs of nodes. It is clearly seen from the table that the results are very similar for two input noise distributions.

B.4 Condition (C1)

To establish the consistency of the proposed test, we require the initial DAG estimator can estimate the ordering consistently; see condition (C1) in Section 4. However, even when (C1) does not hold, our proposed test may still control the type-I error. Actually, in our simulation examples in Section 5, (C1) does not alway hold. Table 3 reports the percentage of times out of 500 data replications when (C1) holds for those selected nodes reported in Table 2 for the nonlinear model (12). It is seen that, for numerous nodes, (C1) only holds for a small fraction of times.

B.5 Power comparison

To compare the power of the two testing methods, we further report the empirical power of our SUGAR method minus that of DRT in Figure 5. It is seen that SUGAR achieves generally a

Table 2: The empirical size and power of the proposed testing method SUGAR under two distributions, Gaussian and uniform, for the input noise in GANs.

Edge	$j = 35, k = 5$		$j = 35, k = 31$		$j = 40, k = 16$	
Hypothesis	$\mathcal{H}_0$		$\mathcal{H}_0$		$\mathcal{H}_0$	
Input Noise	Normal	Uniform	Normal	Uniform	Normal	Uniform
$\alpha = 0.05$	0.050	0.046	0.012	0.022	0.016	0.016
$\alpha = 0.10$	0.078	0.078	0.032	0.046	0.032	0.022
Edge	$j = 45, k = 14$		$j = 45, k = 15$		$j = 50, k = 14$	
Hypothesis	$\mathcal{H}_0$		$\mathcal{H}_0$		$\mathcal{H}_0$	
Input Noise	Normal	Uniform	Normal	Uniform	Normal	Uniform
$\alpha = 0.05$	0.014	0.020	0.032	0.030	0.030	0.034
$\alpha = 0.10$	0.030	0.032	0.058	0.052	0.046	0.052
Edge	$j = 35, k = 4$		$j = 35, k = 30$		$j = 40, k = 15$	
Hypothesis	$\mathcal{H}_1$		$\mathcal{H}_1$		$\mathcal{H}_1$	
Input Noise	Normal	Uniform	Normal	Uniform	Normal	Uniform
$\alpha = 0.05$	0.534	0.524	0.992	0.992	0.550	0.550
$\alpha = 0.10$	0.546	0.552	0.992	0.992	0.550	0.550
Edge	$j = 45, k = 12$		$j = 45, k = 13$		$j = 50, k = 13$	
Hypothesis	$\mathcal{H}_1$		$\mathcal{H}_1$		$\mathcal{H}_1$	
Input Noise	Normal	Uniform	Normal	Uniform	Normal	Uniform
$\alpha = 0.05$	0.946	0.952	0.808	0.824	0.670	0.670
$\alpha = 0.10$	0.948	0.954	0.816	0.832	0.672	0.670

Table 3: The percentage of times out of 500 data replications when (C1) holds for selected nodes for four simulation models.

Nonlinear model (12) with $d = 50, \zeta = 0.10$				
Node j	35	40	45	50
Percentage	11.6%	44.0%	16.4 %	2.2%
Nonlinear model (12) with $d = 100, \zeta = 0.04$				
Node j	80	85	90	
Percentage	48 %	1.9 %	0 %	
Nonlinear model (12) with $d = 150, \zeta = 0.02$				
Node j	132	135	137	140
Percentage	37.1%	20.0%	46.5 %	91.8%

higher power than DRT, over 75% of the times in all scenarios.

C Proofs

We present the technical proofs of Proposition 1, Theorems 1, 2 and 3, followed by an auxiliary lemma needed for the proof of Theorem 3. To simplify the notation, we use O_s to denote the data subset $\{\mathbb{X}_{i,t}\}_{i\in\mathcal{I}_s, 1\le t\le T}$ throughout this section.

C.1 Proof of Proposition 1

We first show that $\mathcal{H}_0(j,k)$ implies $\mathcal{H}_0^*(j,k|\ \mathcal{M})$. Under model (1), it follows from Theorem 1.4.1 of Pearl (2009) that the joint distribution of $(X_1,\ldots,X_d)$ is Markov with respect to the graph. This suggests that the d-separation implies the conditional independence (Pearl, 2009). Under $\mathcal{H}_0(j,k)$, X_j and X_k are d-separated by X_{PA_j}. Under the given conditions on $\mathcal{M}$, we obtain that X_j and X_k are d-separated by $X_{\mathcal{M}-\{k\}}$ as well. Consequently, $\mathcal{H}_0^*(j,k|\mathcal{M})$ holds.

We next show that $\mathcal{H}_0^*(j,k|\mathcal{M})$ implies $\mathcal{H}_0(j,k)$. Under $\mathcal{H}_0^*(j,k|\mathcal{M})$, we have $\mathbb{E}(X_j|X_{\mathcal{M}},X_k) = \mathbb{E}\left(X_j|X_{\mathcal{M}-\{k\}}\right)$. Since $j\in \text{DS}_k$ and $\mathcal{M}\cap \text{DS}_j = \emptyset$, the additive noise ε_j is independent of X_k and $X_{\mathcal{M}}$. Under model (1), we obtain that $\mathbb{E}\{f_j(X_{\text{PA}_j})|X_{\mathcal{M}},X_k\} = \mathbb{E}\left[f_j(X_{\text{PA}_j})|X_{\mathcal{M}-\{k\}}\right]$. Since $\text{PA}_j\subseteq\mathcal{M}$, we have $\mathbb{E}\{f_j(X_{\text{PA}_j})|X_{\mathcal{M}},X_k\} = f_j(X_{\text{PA}_j})$. Consequently, we have $f_j(X_{\text{PA}_j}) = \mathbb{E}\left[f_j(X_{\text{PA}_j})|X_{\mathcal{M}-\{k\}}\right]$. As such, we have $k\notin \text{PA}_j$. Otherwise, there would exist two structural equation models with different graphs that lead to the same joint distribution of $(X_1,\ldots,X_d)$,

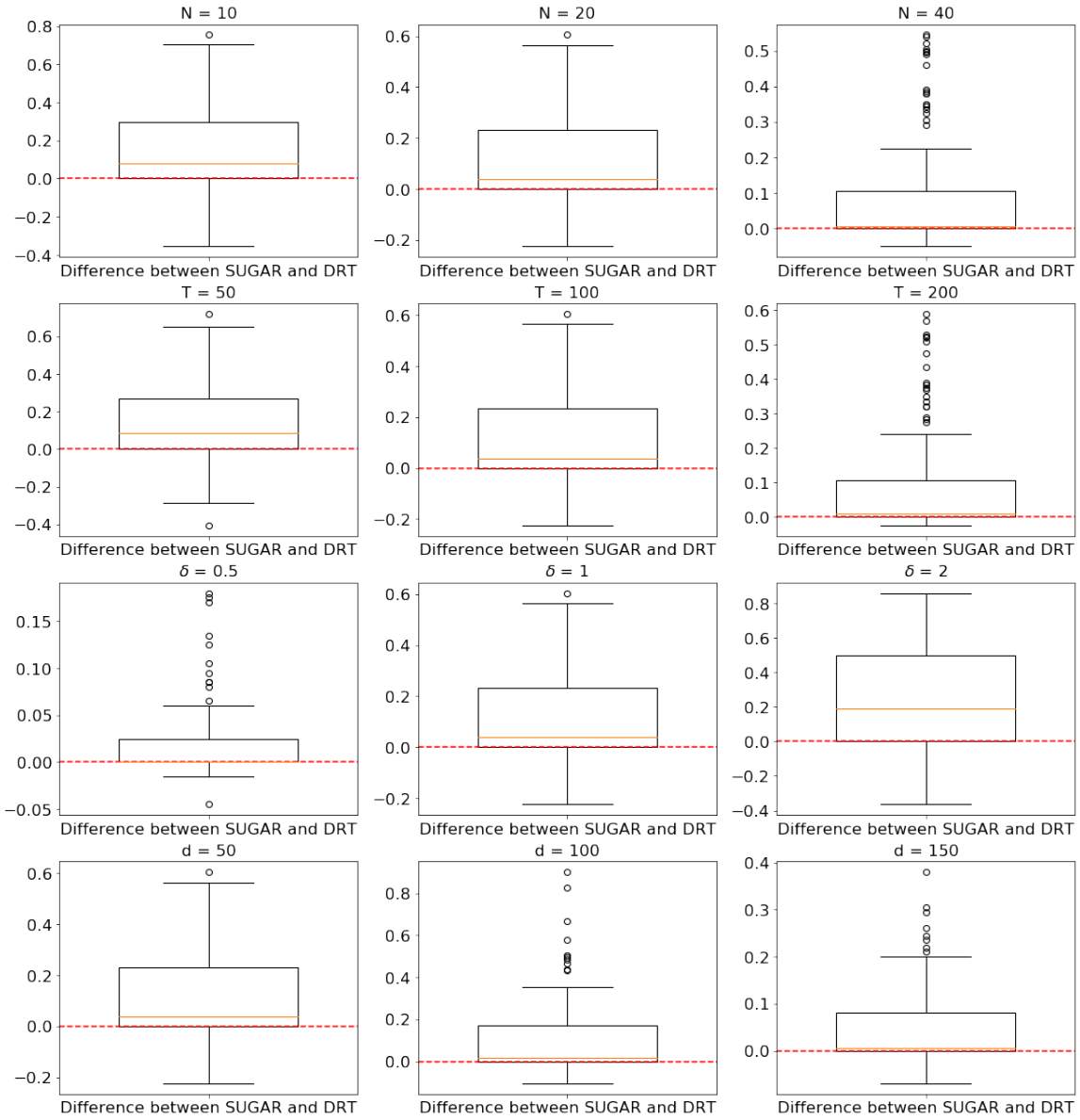

Figure 5: The boxplots of the difference of the empirical power of our proposed test (SUGAR) and that of the double regression-based test (DRT), under four sets of varying parameters: first row $N = \{10, 20, 40\}$, second row $T = \{50, 100, 200\}$, third row $\delta = \{0.5, 1, 2\}$, and fourth row $(d, \zeta) = \{(50, 0.10), (100, 0.04), (150, 0.02)\}$.

and the identifiability condition would have been violated. Therefore, $\mathcal{H}_0(j, k)$ holds.

This completes the proof of Proposition 1. □

C.2 Proof of Theorem 1

It suffices to show that the null hypothesis in (2) is sufficient and necessary to $I(j,k|\mathcal{M};h)=0$ for all square integrable functions h.

The sufficiency follows immediately from Proposition 1 and the definition of the conditional independence.

To prove the necessity, it suffices to show there exists some function h such that $I(j,k|\mathcal{M},h)\neq 0$ under $\mathcal{H}_1(j,k)$. Since X_j has a finite second moment, it follows from model (1) and Jensen's inequality that $\mathbb{E}\left\{f_j^2(X_k,X_{\text{PA}_j})\right\}$ is also finite. Define the function, $h^*(X_k,X_{\mathcal{M}-\{k\}})=f_j(X_k,X_{\text{PA}_j})-\mathbb{E}\big\{f_j(X_k,X_{\text{PA}_j})|X_{\mathcal{M}-\{k\}}\big\}$. It follows that h^* is square integrable. Also by definition,

$$I(j,k|\mathcal{M},h^*)=\mathbb{E}\left[f_j(X_k,X_{\text{PA}_j})-\mathbb{E}\left\{f_j(X_k,X_{\text{PA}_j})|X_{\mathcal{M}-\{k\}}\right\}\right]^2.$$

This measure is not zero. Otherwise, we would have $f_j(X_k,X_{\text{PA}_j})=\mathbb{E}\big\{f_j(X_k,X_{\text{PA}_j})|X_{\mathcal{M}-\{k\}}\big\}$, which would further imply that the data can be generated by another structural equation model such that X_k is not a direct cause of X_j. This would have violated the identifiability condition.

This completes the proof of Theorem 1. □

C.3 Proof of Theorem 2

We begin with a definition. Define

$$\begin{aligned}\widehat{I}^{(s)*}_{b,\text{CF}} &= \frac{2}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T} I^{(s)*}_{i,t,b}, \quad \text{where}\\ I^{(s)*}_{i,t,b} &= \left\{\mathbb{X}_{i,t,j}-g^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\right\}\left[h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\right.\\ &\qquad\left.-\mathbb{E}\left\{h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)|\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right\}\right].\end{aligned}$$

Note that $\left|\widehat{I}^{(s)}_{b,\text{CF}}-\widehat{I}^{(s)*}_{b,\text{CF}}\right|\le\sum_{l=1}^3\left|\eta^{(s)}_{b,l}\right|$, where

$$
\begin{aligned}
\eta_{b,1}^{(s)} &= \frac{2}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T}\left\{\mathbb{X}_{i,t,j}-g^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\right\} \\
&\quad\times\left[\frac{1}{M}\sum_{m=1}^{M}h_b^{(s)}\left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)-\mathbb{E}\left\{h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\mid\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right\}\right], \\
\eta_{b,2}^{(s)} &= \frac{2}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T}\left\{g^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)-\widehat{g}^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\right\} \\
&\quad\times\left[h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)-\mathbb{E}\left\{h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\mid\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right\}\right], \\
\eta_{b,3}^{(s)} &= \frac{2}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T}\left\{g^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)-\widehat{g}^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\right\} \\
&\quad\times\left[\frac{1}{M}\sum_{m=1}^{M}h_b^{(s)}\left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)-\mathbb{E}\left\{h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right)\mid\mathbb{X}_{i,t,\mathcal{M}^{(s)}}\right\}\right].
\end{aligned}
$$

Condition (C1) implies that the set $\widehat{\mathrm{AC}}_j^{(s)}$ meets the conditions of Proposition 1.

We next divide the proof of this theorem into 6 steps. In Steps 1 to 3, we show that $\eta_{\widehat{b}^{(s)},l}^{(s)} = o_p\{(NT)^{-1/2}\}$, for $l=1,2,3$, respectively. In Step 4, we show that, conditional on O_s,

$$
\frac{\widehat{I}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)*}}{\sqrt{\mathrm{Var}\left(\widehat{I}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)*}\mid O_s\right)}}\xrightarrow{d}N(0,1). \tag{17}
$$

In Step 5, we show that the batched mean estimator $\widehat{\sigma}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}$ converges to the standard deviation of $\sqrt{(NT)/2}\widehat{I}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}$ given O_s and the indices of the data subsets $\mathcal{I}_s,\mathcal{I}_s^c$. This together with Step 4 yields that $\sqrt{(NT)/2}\widehat{I}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}/\widehat{\sigma}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}\xrightarrow{d}N(0,1)$ given O_s, $\mathcal{I}_s$ and $\mathcal{I}_s^c$. Hence, $\sqrt{(NT)/2}\widehat{I}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}/\widehat{\sigma}_{\widehat{b}^{(s)},\mathrm{CF}}^{(s)}$ converges to a standard normal distribution unconditionally as well. In Step 6, we put all the above results together to complete the proof. In the following, we assume the data O_s is fixed. The expectation and variance are taken with respect to the data $\{\mathbb{X}_{i,t}\}_{i\in\mathcal{I}_s^c,1\le t\le T}$ conditional on O_s.

Step 1. We first use Berbee's coupling lemma (Dedecker and Louhichi, 2002, Lemma 4.1) to approximate $\eta_{\widehat{b}^{(s)},1}^{(s)}$ by a sum of independent random variables. We then derive the convergence rate of $\eta_{\widehat{b}^{(s)},1}^{(s)}$. Since we assume the data O_s is fixed, the index $\widehat{b}^{(s)}$ is fixed as well.

Denote $\mathcal{I}_\ell^c=\{\ell_1,\ell_2,\ldots,\ell_{N/2}\}$ and $Q=NT/2$. Consider the sequence $\{\mathbb{X}_{(n)}\}_{1\le n\le Q}$ formed by $\{\mathbb{X}_{\ell_i,t}\}_{1\le i\le N/2,1\le t\le T}$, such that $\mathbb{X}_{\ell_i,t}=\mathbb{X}_{((\ell_i-1)T+t)}$ for any i,t. By Condition (C3),

each sequence $\{\mathbb{X}_{i,t}\}_t$ is exponentially β-mixing, and so is $\{\mathbb{X}_{(n)}\}_n$. Following the discussion after Lemma 4.1 of Dedecker and Louhichi (2002), we can construct a sequence of random vectors $\{\mathbb{X}^0_{(n)}\}_n$, such that, with probability at least $1 - Q\beta(q)/q$,

$$\eta^{(s)}_{b,1} = \frac{1}{Q}\sum_{n=1}^{Q}\left\{\mathbb{X}^0_{(n),j} - g^{(s)}\left(\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\right\} \times \left[\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_b\left(\widetilde{\mathbb{X}}^{(m)}_{(n),k}, \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) - \mathbb{E}\left\{h^{(s)}_b\left(\mathbb{X}^0_{(n),k}, \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) \mid \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right],$$

for any b, where we use $\widetilde{\mathbb{X}}^{(m)}_{((\ell_i-1)T+t),k}$ to denote $\widetilde{\mathbb{X}}^{(m)}_{\ell_i,t,k}$, and that the sequences $\{U^0_{2n} : n \geq 0\}$ and $\{U^0_{2n+1} : n \geq 0\}$ are i.i.d., with $U^0_{2n+1} = (\mathbb{X}^0_{(nq)}, \mathbb{X}^0_{(nq+1)}, \ldots, \mathbb{X}^0_{(nq+q-1)})$.

Let $\mathcal{I}_r = \{q\lfloor Q/q\rfloor + 1, q\lfloor Q/q\rfloor + 2, \ldots, Q\}$, we have

$$\left|\eta^{(s)}_{\widehat{b}^{(s)},1}\right| \leq \left|\frac{1}{Q}\sum_{\tau=1}^{\lfloor Q/q\rfloor}\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right| + \left|\frac{1}{Q}\sum_{\tau\in\mathcal{I}_r}\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right| \equiv \delta_1 + \delta_2,$$

with probability $1 - Q\beta(q)/q$, where

$$\eta^{(s)}_{b,1,\tau} = \sum_{n=(\tau-1)q+1}^{\tau q}\left\{\mathbb{X}^0_{(n),j} - g^{(s)}\left(\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\right\} \times \left[\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_b\left(\widetilde{\mathbb{X}}^{(m)}_{(n),k}, \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) - \mathbb{E}\left\{h^{(s)}_b\left(\mathbb{X}^0_{(n),k}, \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) \mid \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right],$$

for $b = 1, \ldots, B$. We next bound δ_1 and δ_2, respectively.

For δ_2, since $\mathbb{H}^{(s)}$ is bounded, we have that,

$$\delta_2 \leq \frac{1}{Q}\sum_{n=(\tau-1)q+1}^{\tau q}\left|\mathbb{X}^0_{(n),j} - g^{(s)}\left(\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\right|.$$

The expectation of the above random variable is of the order $O(qN^{-1}T^{-1})$. Consequently, $\delta_2 = O_p(qN^{-1}T^{-1})$.

For δ_1, without loss of generality, suppose $\lfloor Q/q\rfloor$ is divisible by two. By construction,

$$\delta_1 \leq \left|\frac{1}{Q}\sum_{\tau=1}^{\lfloor Q/q\rfloor/2}\eta^{(s)}_{\widehat{b}^{(s)},1,2\tau-1}\right| + \left|\frac{1}{Q}\sum_{\tau=1}^{\lfloor Q/q\rfloor/2}\eta^{(s)}_{\widehat{b}^{(s)},1,2\tau}\right|,$$

where each of the above two terms corresponds to a sum of independent random variables. Since the data observations are stationary, it follows from Chebyshev's inequality that these

two terms can be upper bounded by $O\left\{(NTq)^{-1/2}\text{Var}^{1/2}\left(\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right)\right\}$. Next, it suffices to bound the variance term $\text{Var}\left(\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right)$.

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\text{Var}\left(\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right) &\leq q^2\mathbb{E}\left\{\mathbb{X}^0_{(n),j} - g^{(s)}\left(\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\right\}^2 \\ &\times\left[\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_{\widehat{b}^{(s)}}\left(\widetilde{\mathbb{X}}^{(m)}_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) - \mathbb{E}\left\{h^{(s)}_{\widehat{b}^{(s)}}\left(\mathbb{X}^0_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\mid \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right]^2.\end{aligned}$$

Under $\mathcal{H}_0(j,k)$ and model (1), the residual $\mathbb{X}^0_{(n),j} - g^{(s)}\left(\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)$ is independent of the variables on the second line. Consequently,

$$\begin{aligned}\text{Var}\left(\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}\right) \;\leq\; & O(1)q^2\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_{\widehat{b}^{(s)}}\left(\widetilde{\mathbb{X}}^{(m)}_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\right. \\ & \left.-\mathbb{E}\left\{h^{(s)}_{\widehat{b}^{(s)}}\left(\mathbb{X}^0_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\mid \mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right]^2,\end{aligned}$$

where $O(1)$ denotes some positive constant. Since

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_{\widehat{b}^{(s)}}\left(\widetilde{\mathbb{X}}^{(s)}_{(n),k,m},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) - \mathbb{E}\left\{h^{(s)}_{\widehat{b}^{(s)}}\left(X^0_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)|\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right]^2 \\ &= \mathbb{E}\left[\text{Var}\left\{\frac{1}{M}\sum_{m=1}^{M} h^{(s)}_{\widehat{b}^{(s)}}\left(\widetilde{\mathbb{X}}^{(s)}_{(n),k,m},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)\middle|\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right] \\ &\quad+\mathbb{E}\left[\mathbb{E}\left\{h^{(s)}_{\widehat{b}^{(s)}}\left(\widetilde{\mathbb{X}}^{(s)}_{(n),k,m},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right) - h^{(s)}_{\widehat{b}^{(s)}}\left(X^0_{(n),k},\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right)|\mathbb{X}^0_{(n),\mathcal{M}^{(s)}}\right\}\right]^2.\end{aligned}$$

By the boundedness of $\mathbb{H}^{(s)}$ and that M is proportional to NT, the second line is of the order $O(N^{-1}T^{-1})$. The third line is of the order $O_p\{(NT)^{-2\kappa_2}\}$ under (C2). Without loss of generality, suppose $\kappa_2 \leq 1$. It follows that $\text{Var}(\eta^{(s)}_{\widehat{b}^{(s)},1,\tau}) = O_p\{q^2(NT)^{-2\kappa_2}\}$. Consequently, $\delta_1 = O_p\left\{q^{1/2}(NT)^{-1/2-\kappa_2}\right\}$.

Putting together the bounds for δ_1 and δ_2, we have that,

$$\left|\eta^{(s)}_{\widehat{b}^{(s)},1}\right| = O_p\left\{q^{1/2}(NT)^{-1/2-\kappa_2}\right\},$$

with probability at least $1 - Q\beta(q)/q$. Since $\beta(q) = O(q^{-\kappa_3})$, set q to be proportional to $\{(NT)\log(NT)\}^{1/(1+\kappa_3)}$. It then follows that $Q\beta(q)/q = O\{\log^{-1}(NT)\} \to 0$. In addition,

since $\kappa_3 > \{2\min(\kappa_1, \kappa_2)\}^{-1} - 1$, we obtain $\left|\eta^{(s)}_{\widehat{b}^{(s)},1}\right| = o_p\{(NT)^{-1/2}\}$. This completes Step 1.

Step 2. This step is derived similarly as Step 1, and the details are omitted.

Step 3. Following similar arguments as in Step 1, we can show that

$$\left|\eta^{(s)}_{\widehat{b}^{(s)},3} - \mathbb{E}\eta^{(s)}_{\widehat{b}^{(s)},3}\right| = o_p\left\{(NT)^{-1/2}\right\}.$$

It then suffices to show $\mathbb{E}\eta^{(s)}_{\widehat{b}^{(s)},3} = o_p\{(NT)^{-1/2}\}$, or equivalently, $\delta_3 = o_p\{(NT)^{-1/2}\}$, where

$$\begin{aligned}\delta_3 \;\equiv\; & \max_{b\in\{1,\ldots,B\}} \left|\mathbb{E}\left\{g^{(s)}\left(X_{\mathcal{M}^{(s)}}\right) - \widehat{g}^{(s)}\left(X_{\mathcal{M}^{(s)}}\right)\right\}\right. \\ & \left. \times\, \mathbb{E}\left\{h^{(s)}_b\left(\widetilde{\mathbb{X}}^{(m)}_k, X_{\mathcal{M}^{(s)}}\right) - h^{(s)}_b\left(X_k, X_{\mathcal{M}^{(s)}}\right) \mid X_{\mathcal{M}^{(s)}}\right\}\right|.\end{aligned}$$

By Cauchy-Schwarz inequality, we have that,

$$\begin{aligned}\delta_3 \;\leq\; & \sqrt{\mathbb{E}|g^{(s)}(X_{\mathcal{M}^{(s)}}) - \widehat{g}^{(s)}(X_{\mathcal{M}^{(s)}})|^2} \\ & \times \max_{b\in\{1,\ldots,B\}} \sqrt{\mathbb{E}\left|\mathbb{E}[\{h^{(s)}_b(\widetilde{\mathbb{X}}^{(m)}_k, X_{\mathcal{M}^{(s)}}) - h^{(s)}_b(X_k, X_{\mathcal{M}^{(s)}})\}|X_{\mathcal{M}^{(s)}}]\right|^2},\end{aligned}$$

where the first term on the right-hand-side is $O\{(NT)^{-\kappa_1}\}$ by condition (C2), and the second term is $O\{(NT)^{-\kappa_2}\}$ by condition (C2). Since $\kappa_1 + \kappa_2 > 1/2$, we have $\delta_3 = o_p\{(NT)^{-1/2}\}$. This completes Step 3.

Step 4. In this step, we aim to establish (17) for $\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}$ under the bidirectional asymptotic framework. Conditional on the data O_s, the index $\widehat{b}^{(s)}$ is fixed. We next show that (17) holds under two scenarios, one with N bounded, and the other with N diverging.

Scenario 4.1: N is bounded and $T \to \infty$. Condition (C3) implies that each $\{\mathbb{X}_{i,t}\}_t$ is strong mixing. Since X_j has the bounded fourth moment, and $\mathbb{H}$ is a bounded function class, it follows from (Rio, 2013, Equation (1.12b)) that $\mathrm{cov}\left(I^{(s)*}_{i,t,\widehat{b}^{(s)}}, I^{(s)*}_{i,t+q,\widehat{b}^{(s)}}\right) = O(\beta^{1/2}(q))$, with respect to q. Since $\beta(q) = O(q^{-\kappa_3})$ and $\kappa_3 > 2$, it follows that $\mathrm{cov}\left(I^{(s)*}_{i,t,\widehat{b}^{(s)}}, I^{(s)*}_{i,t+q,\widehat{b}^{(s)}}\right)$ decays at the rate of $q^{-\kappa^*_3}$ for some $\kappa^*_3 > 1$. Consequently,

$$\sum_{q=-\infty}^{+\infty} \mathrm{cov}\left(I^{(s)*}_{i,t,\widehat{b}^{(s)}}, I^{(s)*}_{i,t+q,\widehat{b}^{(s)}}\right) < +\infty. \tag{18}$$

For each $i \in \mathcal{I}_\ell^c$, the process $T^{-1}\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}$ meets the requirements of Theorem 3 in Kourogenis and Pittis (2011). Consequently, for each $i \in \mathcal{I}_\ell^c$,

$$\frac{\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}}{\sqrt{\text{Var}(\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}})}} \xrightarrow{d} N(0,1). \tag{19}$$

Since the processes $\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}, \ldots, \sum_{1\le t\le T} I^{(s)*}_{N,t,\widehat{b}^{(s)}}$ are i.i.d., we have,

$$\mathbb{E}\exp\left\{iu\frac{\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}}{\sqrt{N\text{Var}\left(\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}\right)/2}}\right\} = \left[\mathbb{E}\exp\left\{iu\frac{\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}}{\sqrt{N\text{Var}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)/2}}\right\}\right]^{N/2}.$$

Since N is bounded, it follows from (19) that

$$\mathbb{E}\exp\left\{iu\frac{\sum_{i\in\mathcal{I}_\ell^c}\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}}{\sqrt{N\text{Var}\left(\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}\right)/2}}\right\} \xrightarrow{d} \left\{\exp\left(-\frac{u^2}{N}\right)\right\}^{N/2} = \exp(-u^2/2),$$

for any u. This completes the proof for this scenario.

Scenario 4.2: $N \to \infty$. We apply the Lindeberg central limit theorem for triangle arrays to derive our results. It suffices to verify the Lindeberg's condition, i.e.,

$$\frac{2}{N\text{Var}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)}\sum_{i\in\mathcal{I}_\ell^c}\mathbb{E}\left(\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}\right)^2$$
$$\times\mathbb{I}\left\{\left|\sum_{1\le t\le T} I^{(s)*}_{i,t,\widehat{b}^{(s)}}\right| \ge \epsilon\sqrt{N\text{Var}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)/2}\right\} \to 0,$$

for any $\epsilon > 0$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

Under the conditions of Theorem 2, we have that,

$$\text{Var}\left(\sqrt{NT}\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}|O_s\right) \ge \kappa_4/2, \tag{20}$$

with probability tending to 1. By (20), and that $\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}, \sum_{1\le t\le T} I^{(s)*}_{2,t,\widehat{b}^{(s)}}, \ldots, \sum_{1\le t\le T} I^{(s)*}_{N,t,\widehat{b}^{(s)}}$ are identically distributed, it suffices to show

$$\frac{4}{\kappa_4 T}\mathbb{E}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)^2\mathbb{I}\left(\left|\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right| \ge \epsilon\sqrt{\kappa_4 NT/4}\right) \to 0,$$

for any $\epsilon > 0$, or equivalently,

$$T^{-1}\mathbb{E}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)^2 \mathbb{I}\left(\left|\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right| > \epsilon N^{1/2}T^{1/2}\right) \to 0.$$

By (18), we have $\mathbb{E}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)^2 = T\mathbb{E}\left(\sum_{1\le t\le T} I^{(s)*}_{1,1,\widehat{b}^{(s)}}\right)^2 + O(T) = O(T)$. By the dominated convergence theorem, it suffices to show

$$T^{-1}\left(\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right)^2 \mathbb{I}\left(\left|I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right| > \epsilon N^{1/2}T^{1/2}\right) = o_p(1),$$

or equivalently,

$$\mathbb{P}\left(\left|\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right| > \epsilon N^{1/2}T^{1/2}\right) \to 0. \tag{21}$$

By Chebyshev's inequality, (21) holds, because

$$\mathbb{P}\left(\left|\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right| > \epsilon N^{1/2}T^{1/2}\right) \le \frac{\mathbb{E}\left|\sum_{1\le t\le T} I^{(s)*}_{1,t,\widehat{b}^{(s)}}\right|^2}{\epsilon^2 NT} = O(N^{-1}) = o(1),$$

as N diverges to infinity. This completes Step 4.

Step 5. In this step, we establish the consistency of the batched mean estimator. We consider three scenarios, when N is bounded and $T \to \infty$, when T is bounded and $N \to \infty$, and when both $N, T \to \infty$.

Scenario 5.1: N is bounded and $T \to \infty$. Note that

$$\begin{aligned}
\widehat{\sigma}^2_{\widehat{b}^{(s)},\mathrm{CF}} &= \frac{2K}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{k=1}^{T/K}\left\{\frac{\sum_{t=(k-1)K+1}^{kK}\left(I^{(s)}_{i,t,\widehat{b}^{(s)}} - \widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}\right)}{\sqrt{K}}\right\}^2 \\
&= \frac{2K}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{k=1}^{T/K}\left\{\frac{\sum_{t=(k-1)K+1}^{kK}\left(I^{(s)}_{i,t,\widehat{b}^{(s)}} - \mathbb{E}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}\right)}{\sqrt{K}}\right\}^2 - K\left\{\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}} - \mathbb{E}\left(\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}\right)\right\}^2 \\
&\equiv \delta_4 - \delta_5.
\end{aligned}$$

Following similar arguments as in Step 4, we can show that

$$\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}} - \mathbb{E}\left(\widehat{I}^{(s)}_{\widehat{b}^{(s)},\mathrm{CF}}\right) = O_p\{(NT)^{-1/2}\}.$$

Since $K \ll NT$, we have $\delta_5 = o_p(1)$. Consequently, it suffices to show that

$$\delta_4 \xrightarrow{P} \frac{NT}{2}\text{Var}\left(\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right) = \frac{1}{T}\text{Var}\left(\sum_{t=1}^{T} I_{i,t,\widehat{b}^{(s)}}\right). \tag{22}$$

Since N is bounded and $K \gg T^{1/(1+\kappa_3)}$, we have $K \gg (NT)^{1/(1+\kappa_3)}$. Without loss of generality, suppose T/K is divisible by 2. Following similar arguments as in Step 1, we approximate δ_4 by

$$\frac{K}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{k=1}^{T/(2K)} \underbrace{\left\{\frac{\sum_{t=(2k-2)K+1}^{(2k-1)K}\left(I^{(s)0}_{i,t,\widehat{b}^{(s)}} - \mathbb{E}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)}{\sqrt{K}}\right\}^2}_{\phi^{(s)}_{i,k,1}}$$
$$+ \quad \frac{K}{NT}\sum_{i\in\mathcal{I}_\ell^c}\sum_{k=1}^{T/(2K)} \underbrace{\left\{\frac{\sum_{t=(2k-1)K+1}^{2kK}\left(I^{(s)0}_{i,t,\widehat{b}^{(s)}} - \mathbb{E}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)}{\sqrt{K}}\right\}^2}_{\phi^{(s)}_{i,k,2}},$$

with probability tending to 1, where $\left\{I^{(s)0}_{i,t,\widehat{b}^{(s)}}\right\}_{i,t}$ denotes the version of $\left\{I^{(s)}_{i,t,\widehat{b}^{(s)}}\right\}_{i,t}$ such that $\left\{\phi^{(s)}_{i,k,m}\right\}_{i,k,m}$ are independent across different pairs (i,k) for any $m = 1, 2$. By condition (C3), using the weak law of large numbers, δ_4 converge in probability to

$$\mathbb{E}\left\{\frac{\sum_{t=1}^{K}\left(I^{(s)}_{1,t,\widehat{b}^{(s)}} - \mathbb{E}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)}{\sqrt{K}}\right\}^2 = \mathbb{E}\left\{\frac{\sum_{t=1}^{K}\left(I^{(s)}_{1,t,\widehat{b}^{(s)}} - \mathbb{E}I^{(s)}_{1,t,\widehat{b}^{(s)}}\right)}{\sqrt{K}}\right\}^2. \tag{23}$$

Similar to (18), we can show that both the right-hand-side of (23) and $T^{-1}\text{Var}\left(\sum_{t=1}^{T} I^{(s)}_{i,t,\widehat{b}^{(s)}}\right)$ are bounded. In addition, their difference is asymptotically negligible as K and T increases to infinity. This yields (22), and completes the proof for this scenario.

Scenario 5.2: T is bounded and $N \to \infty$. By condition (C4), we have $K = T$ under this setting. Then $\mathbb{E}\widehat{\sigma}^2_{\widehat{b}^{(s)},\text{CF}}$ is nearly unbiased to the variance of $\sqrt{(NT)/2}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}$. The consistency follows from the law of large numbers. This completes the proof for this scenario.

Scenario 5.3: Both T and N diverge to infinity. It suffices to show (22). Since N diverges to infinity, δ_4 converges to

$$\mathbb{E}\left\{\frac{\sum_{t=1}^{K}\left(I^{(s)}_{1,t,\widehat{b}^{(s)}} - \mathbb{E}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)}{\sqrt{K}}\right\}^2 = \mathbb{E}\left\{\frac{\sum_{t=1}^{K}\left(I^{(s)}_{1,t,\widehat{b}^{(s)}} - \mathbb{E}I^{(s)}_{1,t,\widehat{b}^{(s)}}\right)}{\sqrt{K}}\right\}^2.$$

Following similar arguments as in Scenario 5.1, we can show (22) holds. This completes Step 5.

Step 6. Putting together the results that $\eta^{(s)}_{\widehat{b}^{(s)},l} = o_p\{(NT)^{-1/2}\}$, $l = 1, 2, 3$, we obtain that $\left|\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}} - \widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}\right| = o_p\{(NT)^{-1/2}\}$. Following similar arguments, we can show that

$$\left|\text{Var}\left(\sqrt{NT}\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}|O_s\right) - \text{Var}\left(\sqrt{NT}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}|O_s\right)\right| = o_p(1).$$

By (20), we have that,

$$\frac{\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}} - \widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}}{\sqrt{\text{Var}\left(\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}|O_s\right)}} = o_p\{(NT)^{-1/2}\}.$$

Note that, under $\mathcal{H}_0(j,k)$, $\mathbb{E}\left(\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}|O_s\right) = 0$. By Step 4, we have that, conditional on O_s, (17) holds. Since the limiting distribution is independent to the data O_s, (17) also holds unconditionally. By Step 5, we have that $\left(\widehat{\sigma}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)^2$ is consistent to the conditional variance of $\sqrt{(NT)/2}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}$. As $\sqrt{(NT)/2}\widehat{I}^{(s)}_{\widehat{b}^{(s)},\text{CF}}$ and $\sqrt{(NT)/2}\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}$ are asymptotically negligible, we can show that $\left(\widehat{\sigma}^{(s)}_{\widehat{b}^{(s)},\text{CF}}\right)^2$ is consistent to the conditional variance of $\sqrt{(NT)/2}\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}$ as well. By Slutsky's theorem, we have that,

$$\frac{\sqrt{(NT)/2}\widehat{I}^{(s)*}_{\widehat{b}^{(s)},\text{CF}}}{\widehat{\sigma}^{(s)}_{\widehat{b}^{(s)},\text{CF}}} \xrightarrow{d} N(0,1),$$

or equivalently, $\widehat{T}^{(s)}_{\widehat{b}^{(s)},\text{CF}} \xrightarrow{d} N(0,1)$. This completes the proof of Theorem 2. □

C.4 Proof of Theorem 3

We first introduce the notion of the VC type class (Chernozhukov et al., 2014, Definition 2.1). Specifically, let $\mathcal{F}$ denote a class of measurable functions, with a measurable envelope function F such that $\sup_{f\in\mathcal{F}}|f| \le F$. For any probability measure Q, let e_Q denote a semi-metric on $\mathcal{F}$ such that $e_Q(f_1, f_2) = \|f_1 - f_2\|_{Q,2} = \sqrt{\int |f_1 - f_2|^2 dQ}$. An ϵ-net of the space $(\mathcal{F}, e_Q)$ is a subset $\mathcal{F}_\epsilon$ of $\mathcal{F}$, such that for every $f \in \mathcal{F}$, there exists some $f_\epsilon \in \mathcal{F}_\epsilon$ satisfying $e_Q(f, f_\epsilon) < \epsilon$. We say that $\mathcal{F}$ is a VC type class with envelope F, if there exist constants $c_0 > 0, c_1 \ge 1$, such that $\sup_Q \mathbb{N}\left(\mathcal{F}, e_Q, \epsilon\|F\|_{Q,2}\right) \le (c_0/\epsilon)^{c_1}$, for all $0 < \epsilon \le 1$, where the supremum is taken

over all finitely discrete probability measures on the support of $\mathcal{F}$, and $\mathbb{N}(\mathcal{F}, e_Q, \epsilon\|F\|_{Q,2})$ is the infimum of the cardinality of $\epsilon\|F\|_{Q,2}$-nets of $\mathcal{F}$. We refer to c_1 as the VC index of $\mathcal{F}$.

We next present the proof. Throughout the proof, we assume the indices of the data subsets $\mathcal{I}_s$ and $\mathcal{I}_s^c$ are fixed, and show the p-value converges to 1 in probability, given $\mathcal{I}_s$ and $\mathcal{I}_s^c$. As such, unconditionally, the p-value converges to 1 in probability as well. We begin with a definition,

$$\widehat{I}_{b,\text{NCF}}^{(s)*} = 2(NT)^{-1} \sum_{i \in \mathcal{I}_\ell} \sum_{1 \le t \le T} I_{i,t,b}^{(s)*},$$

where $I_{i,t,b}^{(s)*}$ is as defined in the proof of Theorem 2. Note that $g^{(s)}$ depends on s only through the set $\widehat{\text{AC}}_j^{(s)}$. Thus, we use the notation $g_{j,k,\mathcal{M}}$ to denote $g^{(s)}$. For a given set $\mathcal{M}$, define

$$\begin{aligned}
\zeta_{b,1}^{(s)}(\mathcal{M}) &= \frac{2}{NT} \sum_{i \in \mathcal{I}_\ell} \sum_{1 \le t \le T} \{\mathbb{X}_{i,t,j} - g_{j,k,\mathcal{M}}(\mathbb{X}_{i,t,\mathcal{M}})\} \\
&\quad \times \left[\frac{1}{M} \sum_{m=1}^{M} h_b^{(s)} \left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}, \mathbb{X}_{i,t,\mathcal{M}} \right) - \mathbb{E} \left\{ h_b^{(s)} \left(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}} \right) \mid \mathbb{X}_{i,t,\mathcal{M}} \right\} \right], \\
\zeta_{b,2}^{(s)}(\mathcal{M}) &= \frac{2}{NT} \sum_{i \in \mathcal{I}_\ell} \sum_{1 \le t \le T} \left\{ g_{j,k,\mathcal{M}}(\mathbb{X}_{i,t,\mathcal{M}}) - \widehat{g}^{(s)}(\mathbb{X}_{i,t,\mathcal{M}}) \right\} \\
&\quad \times \left[h_b^{(s)}(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}}) - \mathbb{E} \left\{ h_b^{(s)}(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}}) | \mathbb{X}_{i,t,\mathcal{M}} \right\} \right], \\
\zeta_{b,3}^{(s)}(\mathcal{M}) &= \frac{2}{NT} \sum_{i \in \mathcal{I}_\ell} \sum_{1 \le t \le T} \left\{ g_{j,k,\mathcal{M}}(\mathbb{X}_{i,t,\mathcal{M}}) - \widehat{g}^{(s)}(\mathbb{X}_{i,t,\mathcal{M}}) \right\} \\
&\quad \times \left[\frac{1}{M} \sum_{m=1}^{M} h_b^{(s)} \left(\widetilde{\mathbb{X}}_{i,t,k}^{(s,m)}, \mathbb{X}_{i,t,\mathcal{M}} \right) - \mathbb{E} \left\{ h_b^{(s)} \left(\mathbb{X}_{i,t,k}, \mathbb{X}_{i,t,\mathcal{M}} \right) \mid \mathbb{X}_{i,t,\mathcal{M}} \right\} \right].
\end{aligned}$$

We next divide the proof of this theorem into 5 steps. In Steps 1 to 3, we show that $\max_{\mathcal{M} \in \mathbb{M}} \max_b |\zeta_{b,l}^{(s)}(\mathcal{M} - \{k\})| = O_p\{(NT)^{-1/2} \log(NT)\}$ for $l = 1, 2, 3$, respectively, where $\mathbb{M}$ denotes the class of subsets $\mathcal{M}$ that meets the requirements of Proposition 1. In Step 4, we show that

$$\left| I \left(j, k | \widehat{\text{AC}}_j^{(s)}; h_{\widehat{b}^{(s)}}^{(s)} \right) \right| \gg N^{-1/2} T^{-1/2}, \tag{24}$$

with probability approaching one. In Step 5, we put all the above results together to complete the proof.

Step 1. It suffices to show $\max_b \left|\zeta_{b,1}^{(s)}(\mathcal{M}-\{k\})\right| = O_p\{(NT)^{-1/2}\log(NT)\}$ for any $\mathcal{M} \in \mathbb{M}$. To simplify the presentation, when there is no confusion, we write $\zeta_{b,l}^{(s)}(\mathcal{M}-\{k\})$ and $g_{j,k,\mathcal{M}-\{k\}}$ as $\zeta_{b,l}^{(s)}$ and $g_{j,k}$, respectively.

To bound $\max_b |\zeta_{b,1}^{(s)}|$, we apply Lemma 1 (see Section C.6). Note that $\widetilde{\mathbb{X}}_{i,t,k,m}^{(s)}$ can be written as $\mathbb{G}^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}-\{k\}}, Z_{j,k}^{(m)}\right)$. Note that the generator $\mathbb{G}^{(s)}$ belongs to a VC type class $\{f : f \in \mathcal{F}\}$ with a bounded envelop function F. Define the function,

$$\begin{aligned}
\tau_{b,f}(\mathbb{X}_{i,t}, Z_{i,t}) &= \{\mathbb{X}_{i,t,j} - \mathrm{g}_{j,k}(\mathbb{X}_{i,t,\mathcal{M}-\{k\}})\} \times \mathbb{E}\big\{\cos(\omega_b f(\mathbb{X}_{i,t,\mathcal{M}-\{k\}}, Z_{i,t})) \\
&\quad - \cos(\omega_b \mathbb{X}_{i,t,k})|\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\big\}, \ \text{ for } 1 \le b \le B/2, \\
\tau_{b,f}(\mathbb{X}_{i,t}, Z_{i,t}) &= \{\mathbb{X}_{i,t,j} - \mathrm{g}_{j,k}(\mathbb{X}_{i,t,\mathcal{M}-\{k\}})\} \times \mathbb{E}\big\{\sin(\omega_b f(\mathbb{X}_{i,t,\mathcal{M}-\{k\}}, Z_{i,t})) \\
&\quad - \sin(\omega_b \mathbb{X}_{i,t,k})|\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\big\}, \ \text{ for } B/2 < b \le B.
\end{aligned}$$

where $\{Z_{i,t}\}_{i,t}$ are i.i.d., and are independent of the observed data. Therefore,

$$\max_b \left|\zeta_{b,1}^{(s)}\right| \le \max_b \sup_{f\in\mathcal{F}}\{2/(NT)\} \left|\sum_{i\in\mathcal{I}_\ell}\sum_{1\le t\le T} \tau_{b,f}(\mathbb{X}_{i,t}, Z_{i,t})\right|.$$

By Lemma A.6 of Chernozhukov et al. (2014), for each b, we can show the class of functions $\{\tau_{b,f} : f \in \mathcal{F}\}$ corresponds to a VC type class with envelop function uniformly bounded by $O(1)|\omega_b|$, where $O(1)$ denotes some positive constant. In addition, we have $\sup_{b,f} \mathrm{Var}(\tau_{b,f}) = O\{(NT)^{-2\kappa_2}\}$ under the given conditions. By setting $q = \kappa\log(NT)$ with some proper choice of κ, it follows from the auxiliary Lemma 1 given in Section C.6, and the given condition on the VC index that, we have, with probability at least $1 - o\{(NT)^{-\kappa_7}\}$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{2}{NT}\left|\sum_{i\in\mathcal{I}_\ell}\sum_{1\le t\le T} \tau_{b,f}(\mathbb{X}_{i,t}, Z_{i,t})\right| &\le O(1)\left[\frac{\omega^*\{\log(NT)+\log\omega^*\}}{NT}\right. \\
&\quad \left. + \frac{\log(NT)+\sqrt{\log(NT)}\sqrt{\log\omega^*}}{\sqrt{NT}}\right].
\end{aligned}$$

where $\omega^* = \max_{1\le b\le B}|\omega_b|$.

By Bonferroni's inequality and the condition that $B = O\{(NT)^{\kappa_7}\}$,

$$\begin{aligned}
&\max_b \sup_{f\in\mathcal{F}} \frac{2}{NT}\left|\sum_{i\in\mathcal{I}_\ell}\sum_{1\le t\le T}\tau_{b,f}(\mathbb{X}_{i,t}, Z_{i,t})\right| \\
\le\ & O(1)\left[\frac{\omega^*\{\log(NT)+\log\omega^*\}}{NT} + \frac{\log(NT)+\sqrt{\log(NT)}\sqrt{\log\omega^*}}{\sqrt{NT}}\right] \\
\le\ & (NT)^{-1/2}\log(NT).
\end{aligned}$$

The last inequality is due to the fact that, each ω_b is standard normal, and $B = O\{(NT)^{\kappa_7}\}$, therefore, $\omega^* = O_p\{\sqrt{\log(NT)}\}$. This yields that $\max_b |\zeta_{b,1}^{(s)}| = O_p\{(NT)^{-1/2}\log(NT)\}$, which completes Step 1.

Step 2. This step is derived similarly as Step 1, and the details are omitted

Step 3. Similar to Step 1, it suffices to bound $\max_b \left|\zeta_{b,3}^{(s)}(\mathcal{M}-\{k\})\right|$ for each $\mathcal{M}\in\mathbb{M}$. By Cauchy-Schwarz inequality and following similar arguments as in the proof of Theorem 3 of Shi et al. (2020), we have, up to some logarithmic terms,

$$\sqrt{\sum_{i\in\mathcal{I}_\ell,1\le t\le T}\mathbb{E}\left|g_{j,k}\left(\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\right)-\widehat{g}_{i,k}^{(s)}\left(\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\right)\right|^2} \le O\{(NT)^{1/2-2\kappa_1}\},$$

$$\max_b\sqrt{\sum_{i,t}\mathbb{E}\left|\left\{h_b^{(s)}\left(\widetilde{\mathbb{X}}_{i,t,k,m}^{(s)},\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\right)-h_b^{(s)}\left(\mathbb{X}_{i,t,k},\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\right)\right\}\mid\mathbb{X}_{i,t,\mathcal{M}-\{k\}}\right|^2} \le O\{(NT)^{1/2-2\kappa_2}\}.$$

Under the condition that $\kappa_1+\kappa_2>1/2$, we obtain that $\max_b\left|\zeta_{b,3}^{(s)}(\mathcal{M}-\{k\})\right| = o_p\{(NT)^{-1/2}\}$ for each $\mathcal{M}\in\mathbb{M}$, which completes Step 3.

Step 4. Based on the results from Steps 1-3, we obtain that

$$\max_b\left|\widehat{I}_{b,\text{NCF}}^{(s)}-\widehat{I}_{b,\text{NCF}}^{(s)*}\right| \le \max_{\mathcal{M}\in\mathbb{M}}\max_b|\zeta_{b,l}^{(s)}(\mathcal{M}-\{k\})| = O_p\{(NT)^{-1/2}\log(NT)\}.$$

In the proof of Theorem 2, we have shown that $\min_b \text{Var}\left(\sqrt{NT}\widehat{I}_{b,\text{CF}}^{(s)*}|O_s\right) \ge \kappa_4/2$. Since $\text{Var}\left(\sqrt{NT}\widehat{I}_{b,\text{CF}}^{(s)*}|O_s\right)$ depends on O_s only though $\mathcal{M}^{(s)}$, we obtain that, with probability approaching one,

$$\min_b \text{Var}(\sqrt{NT}\widehat{I}_{b,\text{CF}}^{(s)*}|\mathcal{M}^{(s)}) \ge \kappa_4/2. \tag{25}$$

Following similar arguments as in the proof of the first three steps, we can show that

$$\max_b \left| \left(\widehat{\sigma}_{b,\text{NCF}}^{(s)} \right)^2 - \text{Var}\left(\sqrt{NT/2} \widehat{I}_{b,\text{CF}}^{(s)*} | \mathcal{M}^{(s)} \right) \right| = o_p(1).$$

Therefore, $\min_b \widehat{\sigma}_{b,\text{NCF}}^{(s)} \geq \sqrt{\kappa_4}/4$. It then follows that,

$$\max_b \left| \frac{\widehat{I}_{b,\text{NCF}}^{(s)*}}{\sqrt{NT\text{Var}\left(\widehat{I}_{b,\text{CF}}^{(s)*} | \mathcal{M}^{(s)} \right)/2}} - \frac{\widehat{I}_{b,\text{NCF}}^{(s)}}{\widehat{\sigma}_{b,\text{NCF}}^{(s)}} \right| = O_p\left\{ \frac{\log(NT)}{\sqrt{NT}} \right\}. \tag{26}$$

Following similar arguments as in Step 1, we can show that $\max_b \left| \widehat{I}_{b,\text{NCF}}^{(s)*} - I\left(j,k|\widehat{\text{AC}}_j^{(s)}; h_b^{(s)} \right) \right| = O_p\{(NT)^{-1/2}\log(NT)\}$. This together with (25) and (26) yields that,

$$\max_b \left| \frac{I\left(j,k|\widehat{\text{AC}}_j^{(s)}; h_b^{(s)} \right)}{\sqrt{NT\text{Var}\left(\widehat{I}_{b,\text{CF}}^{(s)*} | \mathcal{M}^{(s)} \right)/2}} - \frac{\widehat{I}_{b,\text{NCF}}^{(s)}}{\widehat{\sigma}_{b,\text{NCF}}^{(s)}} \right| = O_p\left\{ \frac{\log(NT)}{\sqrt{NT}} \right\}. \tag{27}$$

Next, since $\Delta(\mathbb{H}) \gg (NT)^{-1/2}\log(NT)$, there exists some ω_0, such that one of the following two inequalities hold, $\min_{\mathcal{M}\in\mathbb{M}} I(j,k|\mathcal{M};\cos(\omega_0\cdot)) \gg (NT)^{-1/2}\log(NT)$, or $\min_{\mathcal{M}\in\mathbb{M}} I(j,k|\mathcal{M};\sin(\omega_0\cdot)) \gg (NT)^{-1/2}\log(NT)$. Without loss of generality, suppose the former holds.

Note that the objective function $\min_{\mathcal{M}\in\mathbb{M}} I(j,k|\mathcal{M};\cos(\omega\cdot))$ is Lipschitz continuous in ω, and any ω within the interval $[\omega_0 - (NT)^{-1/2}\log(NT), \omega_0 + (NT)^{-1/2}\log(NT)]$ satisfies that

$$\min_{\mathcal{M}\in\mathbb{M}} I(j,k|\mathcal{M};\cos(\omega\cdot)) \gg (NT)^{-1/2}\log(NT).$$

Since each ω_b is normally distributed, the probability that ω_b falls into this interval is lower bounded by $c(NT)^{-1/2}\log(NT)$ for some constant $c > 0$. Since we randomly generate $B/2$ many ω, the probability that at least one of the ω falls into this interval is lower bounded by

$$1 - \{1 - c(NT)^{-1/2}\log(NT)\}^{B/2} \geq 1 - \exp\{-cB(NT)^{-1/2}\log(NT)/2\}.$$

The above probability tends to 1 under the condition that $B = \kappa_6(NT)^{\kappa_7}$ for some $\kappa_7 \geq 1/2$. Consequently, we obtain that,

$$\max_b \min_{\mathcal{M}\in\mathbb{M}} I(j,k|\mathcal{M};h_b^{(s)}) \gg (NT)^{-1/2}\log(NT).$$

This together with (25) yields that

$$\max_b \left| \frac{I\left(j,k|\widehat{\text{AC}}_j^{(s)}; h_b^{(s)}\right)}{\sqrt{NT\text{Var}\left(\widehat{I}_{b,\text{CF}}^{(s)*}|\mathcal{M}^{(s)}\right)/2}} \right| \gg (NT)^{-1/2}\log(NT).$$

By (27), we have that,

$$\max_b \left| \frac{\widehat{I}_{b,\text{NCF}}^{(s)}}{\widehat{\sigma}_{b,\text{NCF}}^{(s)}} \right| \gg (NT)^{-1/2}\log(NT).$$

By definition, we have that,

$$\left| \frac{\widehat{I}_{\widehat{b}^{(s)},\text{NCF}}^{(s)}}{\widehat{\sigma}_{\widehat{b}^{(s)},\text{NCF}}^{(s)}} \right| \gg (NT)^{-1/2}\log(NT).$$

Using (27) again, we obtain that,

$$\left| \frac{I\left(j,k|\widehat{\text{AC}}_j^{(s)}; h_{\widehat{b}^{(s)}}^{(s)}\right)}{\sqrt{NT\text{Var}\left(\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)*}|\mathcal{M}^{(s)}\right)/2}} \right| \gg (NT)^{-1/2}\log(NT).$$

This together with (25) yields (24). This completes Step 4.

Step 5. Following similar arguments as in the proof of Steps 1-3 in Theorem 2, we can show that $\left|\mathbb{E}\left(\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)} - \widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)*}|O_s\right)\right| = O_p\{(NT)^{-1/2}\}$. By (24), we have $\left|\mathbb{E}\left(\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)*}|O_s\right)\right| \gg N^{-1/2}T^{-1/2}$ with probability approaching one. Following similar arguments as in the proof of Theorem 2, we have that $\sqrt{NT}\left\{\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)} - \mathbb{E}\left(\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)}|O_s\right)\right\} = O_p\{(NT)^{-1/2}\}$. Therefore, $\sqrt{NT}\left|\widehat{I}_{\widehat{b}^{(s)},\text{CF}}^{(s)}\right|$ diverges to infinity with probability approaching one. Consequently, we obtain that $p^{(s)}(j,k) \xrightarrow{p} 0$ for each s. This completes the proof of Theorem 3. □

C.5 Proof of Theorem 4

Under the acyclicity constraint in (9), we have $\widehat{\theta}^{(s)} = \widetilde{\theta}^{(s)}(\widehat{\pi}^{(s)})$ for some ordering $\widehat{\pi}^{(s)}$. We aim to show $\widehat{\pi}^{(s)} \in \Pi^*$ with probability approaching one.

For any ordering π, define the objective function,

$$\mathcal{L}(\pi) = \sum_{j=0}^{d-1} \inf_{f_j} \mathbb{E}\left\{X_{j+1} - f_j\left(X_{\{\pi_1,\ldots,\pi_j\}}\right)\right\}^2,$$

where the minimum is taken over all square integrable functions, and the function f_0 equals zero almost surely. It is straightforward to show that $\mathcal{L}(\pi) = \sum_{j=1}^{d} \mathcal{L}_j(\pi)$, where

$$\mathcal{L}_j(\pi) = \mathbb{E}\left\{X_j - \mathbb{E}\left(X_j | X_{\{\pi_1,\ldots,\pi_{j-1}\}}\right)\right\}^2.$$

Let $\widehat{\mathcal{L}}(\pi) = \sum_{j=1}^{d} \widehat{\mathcal{L}}_j(\pi)$, where $\widehat{\mathcal{L}}_j(\pi)$ is the penalized least squares objective,

$$\min_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \frac{2}{NT} \sum_{i\in\mathcal{I}_\ell} \sum_{1\le t\le T} \{\mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t};\theta_j)\}^2 + \lambda \|A_j^{(1)}\|_{1,1}.$$

Note that, any ordering π that minimizes the objective function $\mathcal{L}(\pi)$ belongs to Π^*. As such, there exists some $\epsilon > 0$, such that

$$\mathcal{L}(\pi^*) \le \min_{\pi\notin\Pi^*} \mathcal{L}(\pi) - \epsilon, \quad \forall \pi^* \in \Pi^*. \tag{28}$$

We next divide the proof of this theorem into 2 steps. In Step 1, we show that $\widehat{\mathcal{L}}(\pi^*)$ converges to $\mathcal{L}(\pi^*)$ for all $\pi^* \in \Pi^*$. In Step 2, we show that $\widehat{\mathcal{L}}(\pi) \ge \mathcal{L}(\pi) + o_p(1)$ for all $\pi \notin \Pi^*$, which ultimately leads to the conclusion of this theorem. Note that the DAG dimension d is fixed in our proof.

Step 1. It suffices to show $\widehat{\mathcal{L}}_j(\pi^*) = \mathcal{L}_j(\pi^*) + o_p(1)$, or equivalently, $|\widehat{\mathcal{L}}_j(\pi^*) - \mathcal{L}_j(\pi^*)| \le \epsilon$ for all $j = 1,\ldots,d$, $\pi^* \in \Pi^*$, and any sufficiently small $\epsilon > 0$.

Fix an $0 < \epsilon < 1$. Since f_j is continuous, it follows from Stone-Weierstrass theorem that there exists a multivariate polynomial function f_j^* such that the absolute value of the residual $f_j - f_j^*$ is uniformly bounded by $\epsilon/6$. Since $\pi^* \in \Pi^*$, $f_j^*(X)$ can be written as a function of $X_{\{\pi_1,\ldots,\pi_{j-1}\}}$.

By Theorem 1 of Yarotsky (2017), there exists a feedforward neural network with a bounded number of hidden units that uniformly approximates f_j^*, with the approximation error uniformly bounded by $\epsilon/6$ in absolute value. By Lemma 1 of Farrell et al. (2021), such a feedforward network can be embedded into an MLP with a bounded number of hidden units. Since we allow H and L to diverge, such an MLP can be further embedded into an MLP with $L-1$ layers and the widths of all layers being proportional to H. Denote this MLP by MLP*, let $A_j^{(1)*},\ldots,A_j^{(L-1)*}$ denote the weight matrices at each layer, and $b_j^{(1)*},\cdots,b_j^{(L-1)*}$ the corresponding bias vectors. We can embed MLP* into another MLP, with L layers, by setting

$A_j^{(l)} = A_j^{(l-1)*}$ and $b_j^{(l)} = b_j^{(l-1)*}$ for $l = 2, \ldots, L$, $A_j^{(1)}$ such that its submatrix formed by columns in $\{\pi_1, \cdots, \pi_{j-1}\}$ and rows in $\{1, \cdots, j-1\}$ is set to an identity matrix and other entries are set to zero, and $b_j^{(1)}$ to a zero vector. The resulting MLP satisfies $\|A_j^{(1)}\|_{1,1} = j-1$, which is finite. Therefore, f_j can be approximated by an MLP with finite $\|A_j^{(1)}\|_{1,1} = j-1$ such that the approximation error is uniformly bounded by $\epsilon/3$ in absolute value. In addition, its weight matrix in the first layer $A_j^{(1)}$ satisfies that $\|A_j^{(j)}\|_{1,1} = j-1$. In other words, there exists some θ_j, such that

$$\left|\mathbb{E}(X_j|X_{\{\pi_1,\ldots,\pi_{j-1}\}} - \text{MLP}(X;\theta_j)\right| \leq \epsilon/3, \tag{29}$$

almost surely, and that

$$\text{supp}(A_j^{(1)}) \in \{\pi_1, \ldots, \pi_{j-1}\} \quad \text{and} \quad \|A_j^{(1)}\|_{1,1} = j-1. \tag{30}$$

It follows from (29) that

$$\begin{aligned}
& \left|\mathbb{E}\left\{X_j - \text{MLP}(X;\theta_j)\right\}^2 - \mathcal{L}_j(\pi)\right| \\
\leq\ & \left|\mathbb{E}\left(X_j|X_{\{\pi_1,\ldots,\pi_{j-1}\}}\right) - \text{MLP}(X;\theta_j)\right|^2 \\
& +2\left|X_j - \mathbb{E}\left(X_j|X_{\{\pi_1,\ldots,\pi_{j-1}\}}\right)\right|\left|\mathbb{E}\left(X_j|X_{\{\pi_1,\ldots,\pi_{j-1}\}}\right) - \text{MLP}(X;\theta_j)\right| \\
\leq\ & \epsilon^2/9 + 2(\epsilon/3)(1+\epsilon/3) < \epsilon.
\end{aligned}$$

This together with (30) and the condition $\lambda \to 0$ yields that,

$$\inf_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \left[\mathbb{E}\sum_{1\leq t\leq T}\left\{X_j - \text{MLP}(X;\theta_j)\right\}^2 + \lambda\|A_j^{(1)}\|_{1,1}\right] - \mathcal{L}_j(\pi) < \epsilon.$$

By definition, we have that,

$$\mathcal{L}_j(\pi) \leq \inf_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \left[\mathbb{E}\sum_{1\leq t\leq T}\left\{X_j - \text{MLP}(X;\theta_j)\right\}^2 + \lambda\|A_j^{(1)}\|_{1,1}\right].$$

It follows that,

$$\left|\inf_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \left[\mathbb{E}\sum_{1\leq t\leq T}\left\{X_j - \text{MLP}(X;\theta_j)\right\}^2 + \lambda\|A_j^{(1)}\|_{1,1}\right] - \mathcal{L}_j(\pi)\right| < \epsilon.$$

To show $|\widehat{\mathcal{L}}_j(\pi^*) - \mathcal{L}_j(\pi^*)| \leq \epsilon$, it suffices to show that,

$$\left| \inf_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \left[\mathbb{E} \sum_{1\leq t\leq T} \{X_j - \text{MLP}(X;\theta_j)\}^2 + \lambda \|A_j^{(1)}\|_{1,1} \right] - \widehat{\mathcal{L}}_j(\pi) \right| = o_p(1).$$

Under the conditions of the theorem, we can further restrict the parameter space to the class of θ_j, such that $\text{MLP}(\cdot;\theta_j)$ is bounded by some constant. As such, the above is upper bounded by

$$\begin{aligned} \sup_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\in\{\pi_1,\ldots,\pi_{j-1}\}}} \Bigg| & \mathbb{E} \sum_{1\leq t\leq T} \{X_j - \text{MLP}(X;\theta_j)\}^2 \\ & - \frac{2}{NT} \sum_{i\in\mathcal{I}_\ell} \sum_{1\leq t\leq T} \{\mathbb{X}_{i,t,j} - \text{MLP}(\mathbb{X}_{i,t};\theta_j^{(s)}(\pi^*))\}^2 \Bigg|, \end{aligned} \tag{31}$$

where the supremum is taken over all θ_j such that $\text{MLP}(\cdot;\theta_j)$ is bounded by some constant. It then suffices to show that (31) is $o_p(1)$. Following Step 1 of Theorem 2, we can first approximate (31) by a sum of independent random variables. This allows us to upper bounded (31) by $O\{\log(NT)\}$ many Radamacher complexity terms, under the exponential β-mixing condition. Following similar arguments as in Section A.2.2 of Liang (2018), each of these Radamacher complexity terms can be upper bounded by $O\{(NT)^{-\kappa_7}\}$ for some $\kappa_7 > 0$, under the given conditions on L and H. This completes Step 1.

Step 2. Following similar arguments as in Step 1, we can show that

$$\widehat{\mathcal{L}}_j(\pi) \geq \min_{\substack{\theta_j=(A_j^{(1)},\ldots,A_j^{(h)}) \\ \text{supp}(A_j^{(1)})\leq\{\pi_1,\ldots,\pi_{j-1}\}}} \mathbb{E}\{X_j - \text{MLP}(X;\theta_j)\}^2 + \lambda \|A_j^{(1)}\|_{1,1} - o_p(1),$$

for any π and $j = 1,\ldots,d$. Since the penalty term is non-negative, and the first term on the right-hand-side is lower bounded by $\mathcal{L}_j(\pi) = \mathbb{E}\left\{X_{j+1} - \mathbb{E}\left(X_{j+1}|X_{\{\pi_1,\ldots,\pi_j\}}\right)\right\}^2$, we obtain that,

$$\widehat{\mathcal{L}}_j(\pi) \geq \mathcal{L}_j(\pi) - o_p(1),$$

for any π and $j = 1,\ldots,d$.

Since d is fixed, so is the number of orderings. In view of (28), we obtain,

$$\widehat{\mathcal{L}}(\pi^*) \leq \min_{\pi \notin \Pi^*} \widehat{\mathcal{L}}(\pi) - \epsilon/2, \quad \text{for any } \ \pi^* \in \Pi^*,$$

with probability approaching one. Note that $\widehat{\pi}^{(s)}$ minimizes the empirical objective function $\widehat{\mathcal{L}}(\pi)$. We thus obtain that $\widehat{\pi}^{(s)} \in \Pi^*$ with probability approaching one. This completes the proof of Theorem 4. □

C.6 An auxiliary lemma

We present a useful lemma that is needed in Step 1 of the proof of Theorem 3. We first briefly introduce the setup. Let $\{Z_t : t \geq 0\}$ be a stationary β-mixing process with the β-mixing coefficient $\{\beta(q) : q \geq 0\}$. Let $\mathcal{F}$ be a pointwise measurable class of functions that take Z_t as input, and has a measurable envelope function F. For any $f \in \mathcal{F}$, suppose $\mathbb{E}\{f(Z_0)\} = 0$. Let $\sigma^2 > 0$ be a positive constant, such that $\sup_{f\in\mathcal{F}} \mathbb{E}\{f^2(Z_0)\} \leq \sigma^2 \leq \mathbb{E}\{F^2(Z_0)\}$. In the next lemma, we provide an exponential inequality for the empirical process $\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right|$.

Lemma 1. *Suppose the envelop function is uniformly bounded by some constant $C > 0$. In addition, suppose $\mathcal{F}$ belongs to the class of VC-type class such that $\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \leq (A/\varepsilon)^\nu$ for some $A \geq e, \nu \geq 1$. Then there exist some constants $c_1, c_2 > 0$, such that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > c_1\sqrt{\nu q\sigma^2 T\log\left(\frac{AC}{\sigma}\right)} + c_1\nu C\log\left(\frac{AC}{\sigma}\right) + c_1 q\tau + Cq\right)$$
$$\leq c_2 q\exp\left(-\frac{\tau^2 q}{c_2 T\sigma^2}\right) + c_2 q\exp\left(-\frac{\tau}{c_2 C}\right) + \frac{T\beta(q)}{q},$$

for any $\tau > 0$ and $1 \leq q < T/2$.

Proof: We divide the proof of this lemma into three steps. In Step 1, we use Berbee's coupling lemma (see Lemma 4.1 in Dedecker and Louhichi, 2002) to approximate $\sup_{f\in\mathcal{F}} |\sum_{t=0}^{T-1} f(Z_t)|$ by the sum of i.i.d. variables. In Step 2, we apply the tail inequality in Lemma 1 of Adamczak (2008) to bound the deviation between the empirical process and its mean. In Step 3, we apply the maximal inequality in Corollary 5.1 of Chernozhukov et al. (2014) to bound the expectation of the empirical process.

Step 1. Following the discussion below Lemma 4.1 of Dedecker and Louhichi (2002), we can construct a sequence of random variables $\{Z_t^0 : t \geq 0\}$, such that

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| = \sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right|, \tag{32}$$

with probability at least $1-T\beta(q)/q$, and that the sequences $\{U_{2i}^0 : i \geq 0\}$ and $\{U_{2i+1}^0 : i \geq 0\}$ are i.i.d., with $U_i^0 = (Z_{iq}^0, Z_{iq+1}^0, \cdots, Z_{iq+q-1}^0)$.

Recall that $\mathcal{I}_r = \{q\lfloor T/q\rfloor, q\lfloor T/q\rfloor + 1, \cdots, T-1\}$, we have

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/q\rfloor} f(Z_{tq+j}^0)\right| + \sup_{f\in\mathcal{F}}\left|\sum_{t\in\mathcal{I}_r} f(Z_t^0)\right|.$$

Under the boundedness assumption on F, the second term on the right-hand-side is bounded from above by Mq. Without loss of generality, suppose $\lfloor T/q\rfloor$ is an even number. The first term on the right-hand-side can be bounded from above by $\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)|$. Therefore,

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| + Mq.$$

This, together with (32), yields that,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \leq \mathbb{P}\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) + \frac{T\beta(q)}{q}, \tag{33}$$

for any $\tau > 0$. By Bonferroni's inequality, we obtain that,

$$\mathbb{P}\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) \leq \sum_{j=0}^{2q-1}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > \tau\right),$$

for any $\tau > 0$. Since the process is stationary, we obtain that,

$$\mathbb{P}\left(\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| > 2\tau q\right) \leq 2q\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| > \tau\right).$$

Combining this with (33) yields that,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \leq 2q\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| > \tau\right) + \frac{T\beta(q)}{q}. \tag{34}$$

By construction, $\{Z^0_{2tq} : t \geq 0\}$ are i.i.d. This completes Step 1.

Step 2. Next, we relate the empirical process $\sup_{f\in\mathcal{F}} |\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})|$ to its expectation. Without loss of generality, suppose $T = kq$ for some integer $k > 0$. Set the constants η and δ in Lemma 1 of Adamczak (2008) to 1, we have that,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| > 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| + \tau\right) \leq 4\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + \exp\left(-\frac{\tau}{CM}\right),$$

for some constant $C > 0$. Combining this with (34), we obtain that,

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > 4q\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| + 2\tau q + Mq\right) \\ \leq 8q\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + 2q\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned} \tag{35}$$

for any $\tau > 0$. This completes Step 2.

Step 3. It remains to bound $\mathbb{E}\sup_{f\in\mathcal{F}} |\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})|$. By Corollary 5.1 of Chernozhukov et al. (2014), we have that,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| \preceq \sqrt{\frac{\nu\sigma^2 T}{q}\log\left(\frac{AM}{\sigma}\right)} + \nu M\log\left(\frac{AM}{\sigma}\right).$$

Combining this with (35), we obtain that,

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| > c\sqrt{\nu q\sigma^2 T\log\left(\frac{AM}{\sigma}\right)} + c\nu M\log\left(\frac{AM}{\sigma}\right) + cq\tau + Mq\right) \\ \leq Cq\exp\left(-\frac{\tau^2 q}{CT\sigma^2}\right) + Cq\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned}$$

for some constants $c, C > 0$, and any $\tau > 0, 1 \leq q < T/2$. This completes the proof of Lemma 1. □