On the Recoverability of Causal Relations from Temporally Aggregated I.I.D. Data

Shunxing Fan [1] **Mingming Gong** [2 1] **Kun Zhang** [1 3]

Abstract

We consider the effect of temporal aggregation on instantaneous (non-temporal) causal discovery in general setting. This is motivated by the observation that the true causal time lag is often considerably shorter than the observational interval. This discrepancy leads to high aggregation, causing time-delay causality to vanish and instantaneous dependence to manifest. Although we expect such instantaneous dependence has consistency with the true causal relation in certain sense to make the discovery results meaningful, it remains unclear what type of consistency we need and when will such consistency be satisfied. We proposed functional consistency and conditional independence consistency in formal way correspond functional causal model-based methods and conditional independence-based methods respectively and provide the conditions under which these consistencies will hold. We show theoretically and experimentally that causal discovery results may be seriously distorted by aggregation especially in complete nonlinear case and we also find causal relationship still recoverable from aggregated data if we have partial linearity or appropriate prior. Our findings suggest community should take a cautious and meticulous approach when interpreting causal discovery results from such data and show why and when aggregation will distort the performance of causal discovery methods.

1. Introduction

Causal discovery methods, which aim to uncover causal relationships from observational data, have been extensively researched and utilized across multiple disciplines including computer science, economics, and social sciences (Pearl, 2009; Spirtes et al., 2000). These methods can be broadly categorized into two types based on whether they explicitly utilize temporal information of the data. The first type is temporal causal discovery, which is specifically designed for analyzing time series data. Examples of this type include the Granger causality test (Granger, 1969) and its variants. The second type is non-temporal causal discovery referring to methods that do not explicitly utilize time index information, which is applicable to independent and identically distributed (i.i.d.) data. This category encompasses various approaches such as constraint-based, score-based, and functional causal model (FCM)-based methods like PC (Spirtes et al., 2000), GES (Chickering, 2002), and LiNGAM (Shimizu et al., 2006).

Temporal causal discovery often relies on the assumption that the causal frequency aligns with the observation frequency, but in real-world scenarios the causal frequency is often unknown, which means that the available observations may have a lower resolution than the underlying causal process. An instance of this is annual income, which is an aggregate of monthly or quarterly incomes (Drost & Nijman, 1993). Moreover, it is widely believed that causal interactions occur at high frequencies in fields such as economics (Ghysels et al., 2016) and neuroscience (Zhou et al., 2014). Some research has been conducted to explore the effects of temporal aggregation on time series modeling (Ghysels et al., 2016; Marcellino, 1999; Silvestrini & Veredas, 2008; Granger & Lee, 1999; Rajaguru & Abeysinghe, 2008; Gong et al., 2017). These works typically consider small aggregation factor k and still treat the temporal aggregation from causal processes as a time series [1].

However, in many real-world scenarios, the temporal aggregation factor can be quite large (we only consider aggrega-

---

[1]Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates [2]School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia [3]Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, United States. Correspondence to: Kun Zhang <kunz1@cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] The "aggregation factor k" refers to the number of data points from the underlying causal process that are combined to form each observed data point. It is also called the aggregation level or aggregation period.

tion without overlapping windows). In these cases, what was originally a time-delayed dependence can appear as an instantaneous dependence when observed. An example commonly used in statistics and econometrics textbooks to illustrate correlation, causation, and regression analysis is the influence of temperature on ice cream sales. Intuitively, one might think that the average daily temperature has an instantaneous causal effect on the total daily ice cream sales. However, in reality, the causal process involves a time lag: a high temperature at a specific past moment influences people's decision to purchase ice cream, which then leads to a sales transaction at a subsequent moment. But we typically work with the aggregation of this causal interaction, such as the average daily temperature and the total daily ice cream sales, which represent the sum of all individual sales transactions over the day and that is the reason we observed instantaneous causality.

Interestingly, the causality community has long acknowledged the significance of temporal aggregation as a common real-world explanation for instantaneous causal models like the structural equation model. Fisher (1970) argued that simultaneous equation models serve as approximations of true time-delayed causal relationships driven by temporal aggregation in the limit. He emphasized that while causation inherently involves a temporal aspect, as the reaction interval tends to zero, the aggregation factor k tends to infinity. Granger (1988) shared a similar view and claimed that "temporal aggregation is a realistic, plausible, and well-known reason for observing apparent instantaneous causation". This explanation has been consistently used in recent causal discovery papers, especially those discussing cyclic models (Rubenstein et al., 2017; Lacerda et al., 2012; Hyttinen et al., 2012).

When applying non-temporal causal discovery methods to uncover instantaneous causal relationships resulting from temporal aggregation, a fundamental question arises: Are these estimated instantaneous causal relationships consistent with the true time-delayed causal relationships? This issue is crucial because our primary concern lies in discerning the true causal relations. If the results obtained by the instantaneous causal discovery methods do not align with the true causal relationship, the results may not suggest anything like how to apply intervention. Regrettably, few studies have examined the alignment of these estimated instantaneous causal relationships stemming from temporal aggregation with the true time-delayed causal relationships. To the best of our knowledge, the only theoretical analysis related to this question is conducted by Fisher (1970) and Gong et al. (2017). We will delve into a comprehensive discussion of their contributions in subsection 1.2.

1.1. Contributions

We propose and investigate two different sense of principle consistencies: functional consistency and conditional independence consistency to formally define what does the recoverability means when we perform FCM-based and conditional independence-based methods on aggregated data respectively. To the best of our knowledge, our paper is the first to specifically discuss the risks and feasibility of performing non-temporal causal discovery methods on aggregated data in general(nonlinear) cases.

1. Functional consistency requires that the aggregated model maintain the consistent functional form of the original causal model. We discuss functional consistency in the context of additive noise models and non-stationary scenarios. For the additive noise model, we present the construction form of the causal function after aggregation in Theorem 3.3 and also provide the necessary and sufficient conditions for the Structural Causal Model (SCM) to hold. For non-stationary scenarios, we discuss different cases and provide results for the general case in Theorem 3.6.
2. Conditional independence consistency expects aggregated model maintain the consistent conditional independence properties as the original causal model. We investigate the three fundamental trivariate structure, chain, fork and collider. We show the collider structure naturally has conditional independence consistency with the causal Markov condition, even for the nonlinear case, but chain and fork do not (see Figure 2 and Remark 4.3). And we provide the necessary and sufficient condition(Theorem 4.4) for the conditional independence consistency for chain and fork structure which can be presented as a equation involving conditional density functions. Each conditional density functions only involving two components of the time series, which motivate us to prove that partial linearity is sufficient to ensure such consistency for chain and fork(Corollary 4.5 and 4.6).
3. We conducted five simulation experiments to support our findings, focusing on theorems for functional and conditional consistency, examining the effect of the k value, and proposing a trivial solution for the aggregation issue.

1.2. Related work

Fisher (1970) established the corresponding relationship between the simultaneous equation model and the true time-lagged model, providing the conditions to ensure such correspondence. His analysis encompassed both linear and general cases. Roughly speaking, he conducted theoretical analysis to show that this correspondence can be ensured when the function of the equation has a fixed point. However, the assumptions he employed were quite restrictive, assuming that the value of noise is fixed for all the causal reactions during the observed interval. Some subsequent studies have also adopted this assumption (Rubenstein et al., 2017; Lacerda et al., 2012). This assumption is too strong, as it implies that noise in the causal reaction is only related to our observation like measurement error. Actually, the noise defined in structural causal models or functional causal models also represents unobserved or unmeasured factors that contribute to the variation in a variable of interest, rather than being merely observational noise.

Gong et al. (2017) adopted a more practical and flexible assumption in their work. They defined the original causal process as a vector auto-regressive model (VAR) $X_t = AX_{t-1} + e_t$ and allowed for the randomness of the noise e_t in the observation interval. They gave a theoretical analysis showing that, in the linear case and as the aggregation factor tends to infinity, the temporally aggregated data $\overline{X}$ becomes i.i.d. and compatible with a structural causal model $\overline{X} = A\overline{X} + \overline{e}$. In this model, matrix A is consistent with the matrix A in the original VAR. This suggests that high levels of temporal aggregation preserve functional consistency in the linear case. However, their study only considers the linear case and lacks analysis for general cases.

2. Preliminary

In this section, we first introduce the Vector Autoregressive (VAR) model and its temporal aggregation. Then, we briefly describe how, as the aggregation level k increases, the time-delay dependence vanishes and how it can be approximated as the aggregation of an instantaneous causal model. Based on this fact, we ignore the unaligned variables caused by time-delay causal effects and define the underlying process with instantaneous causal effects across different components, while allowing potential time-delay causal effects between variables within the same component. The analysis in subsequent sections is based on the aggregation of instantaneous causal effects.

2.1. VAR Model and its Temporal Aggregation

Aligning with the settings in Gong et al. (2017), we assume the underlying causal process can be described by a VAR(1) (autoregressive model with a maximum lag of 1):

$$X_t = f(X_{t-1}, N_t), \quad t \geq 2,$$

where $X_t = (X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(s)})^T$ is the observed data vector at time t, s is the dimension of the random vector. f is a vector-valued function $\mathbb{R}^{2s} \to \mathbb{R}^s$, and $N_t = (N_t^{(1)}, \ldots, N_t^{(s)})^T$ denotes a temporally and contemporaneously independent noise process. When mentioning VAR in our paper, we refer to the general VAR model defined above, which includes both linear and nonlinear functions.

The temporally aggregated time series of this process, denoted by $\overline{X_t}$, is defined as:

$$\overline{X_t} = \frac{\sum_{i=1}^{k} X_{i+(t-1)k}}{g(k)}. \quad (1)$$

When we mention the aggregation of the time-delay model in the paper, we consider cases where k is large by default. In such cases, we treat the temporally aggregated data as i.i.d. data, and we will drop the subscripts t in Eq. 1 from now on. The normalization factor $g(k)$ is generally required to satisfy $\lim_{k\to\infty} g(k) = +\infty$, like $g(k) = k$. It should be chosen carefully to ensure that the limit of temporal aggregation has a finite, non-zero variance.

We also introduce the concept of a summary graph, which often serves as the objective for causal discovery from a random process. Figure 1 shows an example of a summary graph.

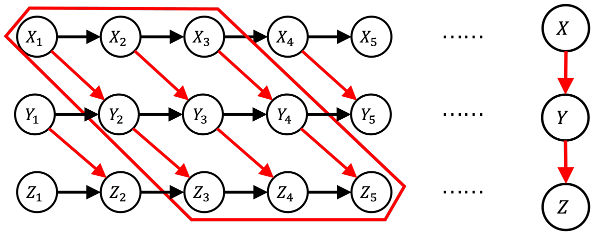

Figure 1. Left: Directed acyclic graph for the VAR model with chain-like cross lag effects. Right: The corresponding summary graph.

Definition 2.1 (Summary Graph)**.** Each time series component is collapsed into a node to form the summary causal graph. The summary graph represents causal relations between time series without referring to time lags. If there is an arrow from $X_{t-k}^{(i)}$ to $X_t^{(j)}$ in the original process for some $k \geq 0$, then it is represented in the summary graph.

2.2. Approximation

In this subsection, we introduce **how the aggregation of a time-delay model can be approximated as the aggregation of an instantaneous model.** This section merely provides a simple example to illustrate why the aggregation of instantaneous causation can be used to approximate the aggregation of time-delay causation. In the Appendix I, we provide more rigorous and general theorem and proof to justify this approach.

See Figure 1 as an example. The summary graph of this VAR is a chain structure involving triple variables. But the chain structure actually occurs in a lagged form:

$X_t \to Y_{t+1} \to Z_{t+2}$ [2]. For analytical convenience, we will perform an alignment $X'_t := X_t, Y'_t = Y_{t+1}, Z'_t := Z_{t+2}$ to make the causal effect instantaneous. We refer to the instantaneous model over X'_t, Y'_t, Z'_t : as the aligned model. In the example of chain-like VAR, the aligned model is in the red box in the Figure 1.

When k is large, the temporal aggregation from the original VAR and the aligned model is exactly the same: as k approaches infinity, the temporally aggregated data $(\overline{X'}, \overline{Y'}, \overline{Z'})$ tends towards $(\overline{X}, \overline{Y}, \overline{Z})$. This can be demonstrated by the following equalities: Since $g(k) \to \infty$, we have

$$\begin{aligned}\overline{Y'} - \overline{Y} &= \frac{\sum_{i=2}^{k+1} Y_i}{g(k)} - \frac{\sum_{i=1}^{k} Y_i}{g(k)} = \frac{Y_{k+1} - Y_1}{g(k)} \to 0,\\ \overline{Z'} - \overline{Z} &= \frac{\sum_{i=3}^{k+2} Z_i}{g(k)} - \frac{\sum_{i=1}^{k} Z_i}{g(k)}\\ &= \frac{Z_{k+2} + Z_{k+1} - Z_1 - Z_2}{g(k)} \to 0,\end{aligned}$$

as $k \to \infty$.

This approximation is natural, as the similar assertion that high aggregation causes time-delay dependence to become instantaneous has appeared and been discussed in many papers (Pierce & Haugh, 1977; Granger, 1980; 1988; Renault et al., 1998; Breitung & Swanson, 2002; Gong et al., 2017). Please note that the number of unaligned variables we rule out (outside the red block in Figure 1) is quite small compared to the variables within the block. Therefore, as k increases, the aggregation of the time-delay model will rapidly approach the aggregation of the instantaneous model in practice (see Appendix F for an experiment investigating the effect of k value). If the k value is quite small and the unaligned variables significantly contribute to causal discovery, it indicates that the data exhibit an obvious time-delay property. In such cases, one should employ a temporal causal discovery method, which is beyond the scope of this paper.

Therefore, all the theoretical results in this paper consider the aggregation of the instantaneous underlying model, also referred to as the aligned model. They are formally defined in Definition 3.1 in Section 3 and Definition 4.1 in Section 4. From this perspective, our paper can be seen as exploring when it is possible to recover causal relations from the aggregation of instantaneous causal models and **all the results can be applied to them with any finite k value and any choice of** $g(k)$ (thus we set $g(k) = 1$ for simplicity in the subsequent sections). Only when we want to apply the results to the aggregation of the time-delay model do we need k to be large and $g(k)$ to be appropriate, i.e., $\lim_{k\to\infty} g(k) = +\infty$.

3. Functional Consistency: Recoverability with FCM-based Methods

FCM-based methods make stronger assumptions and utilize more information beyond just conditional independence. Thus, they can distinguish cause from effect from observational data under the functional assumptions. If we want to ensure the reliability of the results from FCM-based causal discovery on temporally aggregated data, we need some functional consistency between the process of temporally aggregated data and the true causal process.

3.1. Definitions

The FCM-based methods primarily focus on the causal direction and consider the bivariate case in their studies. Aligned with (Shimizu et al., 2011; Hoyer et al., 2008; Zhang & Hyvarinen, 2012), we also consider the process involving two components.

Definition 3.1 (Bivariate Aligned Model with Instant Structures)**.** The aligned model for VAR model with two components $X \to Y$ is defined below:

X_0, Y_0 are independent and follow the initial distribution. when $t \geq 1$,

$X_t = f_X(X_{t-1}, N_{X,t}), Y_t = f_Y(X_t, N_{Y,t})$

where f_X, f_Y are general functions. $N_{X,t}$, $N_{Y,t}$ are independent random variables with non-zero variance, which are independent of each other and they are identically distributed and independent across time t.

The temporal aggregation (for simplicity we set $g(k) = 1$) are denoted as $\overline{X} := \sum_{i=1}^{k} X_i$, and similarly for $\overline{Y}$, $\overline{Z}$.

Intuitively, functional consistency implies that the functional causal model still exists after aggregation and adheres to the correct causal direction. We attempt to formulate a general definition; however, the problem becomes trivial because a causal function and independent noise can always be constructed to create a Structural Causal Model between any two random variables if no additional constraints are applied (Darmois, 1953; Hyvärinen & Pajunen, 1999; Zhang et al., 2015). Therefore, in line with the requirements of commonly used FCM-based methods, we introduce some reasonable constraints, such as requiring the model to incorporate additive noise.

---

[2] Here X_t represents a one-dimensional variable. Starting from here, X_t and $\overline{X}$ represent scalar random variable, and we will use X, Y, Z,... to represent the different components of multivariate time series, instead of using $X^{(1)}$,...,$X^{(s)}$ as defined in 3.1.

3.2. Discussion for Two Types of Functional Consistency

Definition 3.2 (Functional Consistency Regarding Additive Noise)**.** Consider the bivariate aligned model defined in 3.1 incorporates additive noise: $Y_t = f(X_t) + N_{Y,t}$. This process exhibits functional consistency regarding additive noise if there exists a function $\hat{f}$ such that the aggregated variables can be represented as $\overline{Y} = \hat{f}(\overline{X}) + N$, where N is independent of $\overline{X}$, and such $\hat{f}$ exists only in the correct causal direction.

When the original causal function f is linear, then $\overline{Y} = \sum_{i=1}^{k}(AX_i + N_{Y,t}) = A\overline{X} + \sum_{i=1}^{k} N_{Y,i}$ (Gong et al., 2017), $\sum_{i=1}^{k} N_{Y,i}$ is independent of $\overline{X}$. However, because such an SCM exists on both sides, it is unidentifiable; we require non-Gaussian (Shimizu et al., 2006; 2011) or nonlinearity (Hoyer et al., 2008) to achieve asymmetry. Nonetheless, due to the central limit theorem, even if $N_{Y,i}$ is non-Gaussian, $\sum_{i=1}^{k} N_{Y,i}$ will quickly exhibit Gaussian characteristics, making it unidentifiable. See Appendix D for experiments showing how rapidly non-Gaussianity vanishes as k increases.

In the nonlinear case, we derive the following theorem:

Theorem 3.3 (Construction of $\hat{f}$)**.** *If such $\hat{f}$, as defined in Definition 3.2, exists, then $\hat{f}$ must take the form:*

$$\hat{f}(T) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + c, \tag{2}$$

where c is any constant (which can be incorporated into the noise term) and the expression $\mathbb{E}(\cdot \mid \overline{X} = T)$ denotes the conditional expectation. For simplicity, we set $c = 0$. Consequently, this implies:

$$\mathbb{E}(\hat{f}(\overline{X})) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i)\right). \tag{3}$$

See Appendix A for the proof. This construction of $\hat{f}$ is very interesting and bears a strong resemblance to the Rao-Blackwell Theorem (Lehmann & Casella, 2006), which aims to find an estimator (or a function in the causal inference framework) that accurately represents the underlying relationship between variables while accounting for noise or other confounding factors. The principle that the new estimator should be a function of some statistics, and using conditional expectations to enforce this, is very similar to what is described by the Theorem 3.3.

Then we can present $\overline{Y}$ in two ways: $\overline{Y} = \sum_{i=1}^{k} f(X_i) + \sum_{i=1}^{k} N_{Y,i} = \hat{f}(\overline{X}) + N$, where $\hat{f}$ is known and, according to the definition, we expect N to be independent of $\overline{X}$. Therefore, we directly imply the results below.

Theorem 3.4 (Necessary and Sufficient Condition)**.** *The necessary and sufficient condition for the existence of the additive noise causal model defined in Definition 3.2 is that $N = \sum_{i=1}^{k} N_{Y,i} + \left(\sum_{i=1}^{k} f(X_i) - \hat{f}(\overline{X})\right)$ is independent of $\overline{X}$, where $\hat{f}$ is defined by Eq. 2.*

The proof can be found in Appendix A.

It is worth noting that the construction of $\hat{f}$ naturally ensures that N is uncorrelated with $\overline{X}$. However, it is still challenging to achieve independence. For the condition, we find that the term $\sum_{i=1}^{k} N_{Y,i}$ is independent of every X_i and $\overline{X}$, as well as the construction of $\hat{f}$. Thus, we cannot rely on this term to enforce independence. Furthermore, the term $\sum_{i=1}^{k} f(X_i) - \hat{f}(\overline{X})$ is likely dependent on $\overline{X}$. If we calculate the conditional variance, conditional on $\overline{X} = T$, for $\sum_{i=1}^{k} f(X_i) + \sum_{i=1}^{k} N_{Y,i}$ and $\hat{f}(\overline{X}) + N$ respectively, then we will get $\text{var}(N) = \text{var}\left[\sum_{i=1}^{k} N_{Y,i}\right] + \text{var}\left[\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right]$. Please note that the LHS is not related to T, and the first term in RHS is not related to T as well. This requires $\text{var}\left[\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right]$ also to be unrelated to T, which is a non-trivial requirement for the distribution of X_i and f.

Sometimes, data is collected from different regions/modules. The causal mechanism remains unchanged across these regions, but the distribution of data may vary. The identifiability of some causal discovery methods depends on the variance in distribution and the consistency of the causal function across different regions (Huang et al., 2020). We are now investigating whether the same causal mechanism persists in different regions when considering aggregated data.

Definition 3.5 (Functional Consistency with respect to different regions)**.** Consider causal models $X_{t,A} = f_X(X_{t-1,A}, N_{X,t,A})$, $Y_{t,A} = f(X_{t,A}, N_{t,A})$ and $X_{t,B} = f_X(X_{t-1,B}, N_{X,t,B})$, $Y_{t,B} = f(X_{t,B}, N_{t,B})$ for regions A and B, respectively, where $N_{X,t,A}$, $N_{X,t,B}$, $N_{t,A}$, and $N_{t,B}$ represent independent noises with region-specific distributions and are i.i.d. within each region. $\overline{X}_{A/B}$ and $\overline{Y}_{A/B}$ are the aggregated data across individuals within each region. This process exhibits functional consistency with respect to different regions if, only for the correct causal direction $X \to Y$, there exists a function $\hat{f}$ such that the aggregated variables satisfy $\overline{Y}_A = \hat{f}(\overline{X}_A, N_A)$ and $\overline{Y}_B = \hat{f}(\overline{X}_B, N_B)$.

Obviously, if f is a linear function, then it will always be consistent across different regions. For the nonlinear additive noise model, according to the construction formula Eq. 2, which is rely on the distribution of X_i, it is feasible to provide examples of $\hat{f}$ being inconsistent across different regions (see Appendix H). In general cases, due to a lack of constraint, it is always possible to find a consistent function; however, such a function can exist in both directions, rendering the system still unidentifiable. See Appendix A for the proof.

Theorem 3.6 (General Case for Functional Consistency Across Different Regions)**.** *For the causal model defined in Definition 3.5, assume the aggregated variables* $\overline{X}_A$*,* $\overline{X}_B$ *and* $\overline{Y}_A$*,* $\overline{Y}_B$ *have continuous support. Then, there exists functions* $\hat{f}$ *and* $\hat{g}$ *such that* $\overline{Y}_A = \hat{f}(\overline{X}_A, N_A)$*,* $\overline{Y}_B = \hat{f}(\overline{X}_B, N_B)$*,* $\overline{X}_A = \hat{g}(\overline{Y}_A, N'_A)$*, and* $\overline{X}_B = \hat{g}(\overline{Y}_B, N'_B)$*, where* N_A *is independent of* $\overline{X}_A$*,* N_B *is independent of* $\overline{X}_B$*,* N'_A *is independent of* $\overline{Y}_A$*, and* N'_B *is independent of* $\overline{Y}_B$*.*

4. Conditional Independence Consistency: Recoverability with Constraint-based Method

Constraint-based causal discovery methods utilize conditional independence to identify the Markov equivalence classes of causal structures. These methods heavily rely on the faithfulness assumption, which posits that every conditional independence in the data corresponds to a d-separation in the underlying causal graph.

The information utilized by constraint-based causal discovery methods is less than that used by FCM-based methods. This implies that the consistency we require for the recoverability of constraint-based methods on temporally aggregated data is less stringent than functional consistency. In essence, we only need the temporally aggregated data to preserve the conditional independence of the summary graph of the underlying causal process. If the temporal aggregation maintains such conditional independence consistency, then the constraint-based causal discovery method can recover the Markov equivalence class of the summary graph entailed by the underlying true causal process.

4.1. Definitions and Problem Formulation

To examine whether the temporally aggregated data preserves the conditional independence consistency with the summary graph of the underlying causal process, we will discuss the three fundamental causal structures of the summary graph: the chain, the fork, and the collider. We will provide theoretical analysis for each of these three fundamental cases respectively.

In subsection 3.1, we assume the original causal process is VAR(1) and we work with the temporal aggregation of it. But in this section, for analytical convenience we will assume the original model is an aligned version of VAR(1), and work with the temporal aggregation of it. We will show this alignment is reasonable because the temporal aggregation of these two original processes is the same when k is large.

4.1.1. ALIGNED MODEL

Definition 4.1 (Trivariant Aligned Model with Instant Structures)**.** The aligned model for VAR model with structure function f_X, f_Y, f_Z incorporating chain-like cross lag effect is given by:

X_0, Y_0, Z_0 are independent and follow the initial distribution. when $t \geq 1$,

Chain-like Model: $X_t = f_X(X_{t-1}, N_{X,t})$, $Y_t = f_Y(X_t, Y_{t-1}, N_{Y,t})$, $Z_t = f_Z(Y_t, Z_{t-1}, N_{Z,t})$,

Fork-like Model: $X_t = f_X(X_{t-1}, Y_t, N_{X,t})$, $Y_t = f_Y(Y_{t-1}, N_{Y,t})$, $Z_t = f_Z(Y_t, Z_{t-1}, N_{Z,t})$,

Collider-like Model: $X_t = f_X(X_{t-1}, N_{X,t})$, $Y_t = f_Y(X_t, Y_{t-1}, Z_t, N_{Y,t})$, $Z_t = f_Z(Z_{t-1}, N_{Z,t})$,

where f_X, f_Y, f_Z are general functions. $N_{X,t}, N_{Y,t}, N_{Z,t}$ are independent random variables with non-zero variance, which are independent of each other and they are identically distributed and independent across time t.

The temporal summation and aggregation are denoted as $S_X := \sum_{i=1}^{k} X_i$, $\overline{X} := \frac{S_X}{g(k)}$, and similarly for S_Y, $\overline{Y}$, S_Z, and $\overline{Z}$. When k is finite, S_X, S_Y, and S_Z have the same conditionally independent relationship with $\overline{X}$, $\overline{Y}$, and $\overline{Z}$, respectively. Therefore, when discussing conditional independence, we can treat S_X, S_Y, and S_Z as equivalent to $\overline{X}$, $\overline{Y}$, and $\overline{Z}$.

The figures of the aligned models of three fundamental structure involving temporal aggregation variables are presented in Figure 2.

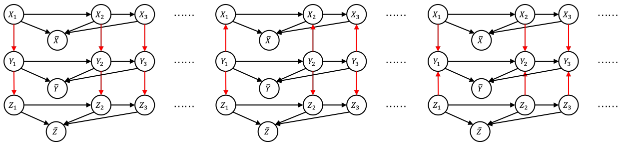

Figure 2. Left: Chain-like aligned model. Center: Fork-like aligned model. Right: Collider-like aligned model.

4.1.2. PROBLEM FORMULATION

Definition 4.2 (Conditional Independence Consistency)**.** Consider an underlying causal process generating temporally aggregated data. This process is said to exhibit conditional independence consistency if the distribution of temporally aggregated data satisfies the Markov condition with respect to the summary graph of original process.

We will address the problem of determining the conditions under which temporal aggregation preserves conditional independence consistency in three fundamental causal structures: chain, fork, and collider.

4.2. Necessary and Sufficient Conditions for Consistency

For the figure of a chain or fork structure, we expect the middle node can d-separate the nodes on both sides. However, from the structure of Figure 2 Left and Center, we can find that all adjacent nodes of $\overline{Y}$ point to $\overline{Y}$. Therefore, when we condition on $\overline{Y}$, we cannot block any path.

For the figure of a collider structure, we expect the nodes on both sides are unconditionally independent and when conditioned on the middle node, the nodes on both sides will be dependent. Fortunately, from the structure of Figure 2 Right, we can find that all the Y_t are collider for X_t and Z_t so $\overline{X} \perp\!\!\!\perp \overline{Z}$ unconditionally. And because $\overline{Y}$ is a descendant of these colliders, when we condition on $\overline{Y}$, the path involving X_t and Z_t will be open. As a result, $\overline{X}$ is dependent with $\overline{Z}$ conditional on $\overline{Y}$.

Remark 4.3 (Conditional Independence Consistency under Faithfulness Condition). Assume the aligned models satisfy the causal Markov condition and causal faithfulness condition with respect to the causal graphs of aligned models in Figure 2.

- The conditional independent sets of temporal aggregation of chain-like/fork-like aligned model is $\emptyset$, which is **not** consistent with the chain/fork structure.
- The conditional independent sets of temporal aggregation of collider-like aligned model is $\overline{X} \perp\!\!\!\perp \overline{Z}$, which is consistent with the collider structure.

This remark emphasizes that in general cases, the temporal aggregation of a model with chain-/fork-like structure does not exhibit the same conditional independence as a genuine chain or fork structure under the faithfulness assumption. As a result, we will explore the conditions required to ensure the validity of the conditional independence $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$ in the context of temporal aggregation.

Theorem 4.4 (Necessary and Sufficient Condition for Conditional Independence Consistency of Chain and Fork Structure)**.** *Consider the distribution of* $(\overline{X}, \overline{Y}, \overline{Z}, Y_1, ..., Y_k)$ *entailed from the aligned model, the following statements are equivalent when* $2 \leq k < \infty$*:*

(i) Conditional Independence: $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$
(ii) Conditional Probability Relation: $\forall s_X, s_Y, s_Z \in \mathbb{R}$
$\iint_{\mathbb{R}^k} \alpha(s_Z, s_Y, y_{1:k}) \left(\beta(y_{1:k}, s_Y, s_X) - \gamma(y_{1:k}, s_Y)\right)$
$dy_1 ... dy_k = 0$
(iii) Alternative Conditional Probability Relation:
$\forall s_X, s_Y, s_Z \in \mathbb{R}$
$\iint_{\mathbb{R}^k} \alpha^*(s_X, s_Y, y_{1:k}) \left(\beta^*(y_{1:k}, s_Y, s_Z) - \gamma(y_{1:k}, s_Y)\right)$
$dy_1 ... dy_k = 0$

where $\alpha := p_{S_Z|S_Y,Y_{1:k}}$, $\beta := p_{Y_{1:k}|S_Y,S_X}$, $\alpha^* := p_{S_X|S_Y,Y_{1:k}}$, $\beta^* := p_{Y_{1:k}|S_Y,S_Z}$, $\gamma := p_{Y_{1:k}|S_Y}$.

See Appendix B for the proof. From this sufficient and necessary condition, we find that the integrand can be divided into two parts. For example, the integrand in formula 4.4 can be divided into two parts. The first part is $p_{S_Z|S_Y,Y_{1:k}}(s_Z|s_Y, y_{1:k})$. Because $Y_1, ...Y_k$ d-separate S_Z from $X_1, ..., X_k$ perfectly, this part is related to the causal mechanism between Y and Z. The second part $\left(p_{Y_{1:k}|S_Y,S_X}(y_{1:k}|s_Y, s_X) - p_{Y_{1:k}|S_Y}(y_{1:k}|s_Y)\right)$ is related to the causal mechanism between Y and Z. This inspires us to consider different parts of the model individually.

Corollary 4.5 (Sufficient Conditions for Conditional Independence)**.** *If* $\{\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ ***or*** $\{\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ *holds, then* $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$ *holds.*

Proof can be found in Appendix B. This corollary has a very intuitive interpretation: when the information needed to infer $\overline{X}$ from $Y_{1:k}$ is completely included in $\overline{Y}$ ($\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$), then conditioning on $\overline{Y}$ is equivalent to conditioning on $Y_{1:k}$. In this case, $Y_{1:k}$ d-separate $\overline{X}$ from $\overline{Z}$. The same principle applies to $\overline{Z}$. When does the information to infer $\overline{X}/\overline{Z}$ from $Y_{1:k}$ is completely included in $\overline{Y}$?

Corollary 4.6 (Partial Linear Conditions)**.**

1. *For a fork-like aligned model:*
    - *If* $f_Z(Y_t, Z_{t-1}, N_{Z,t})$ *is of the form* $\alpha * Y_t + N_t$*, where* α *can be any real number, then* $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.
    - *If* $f_X(X_{t-1}, Y_t, N_{X,t})$ *is of the form* $\alpha * Y_t + N_t$*, where* α *can be any real number, then* $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.
2. *For a chain-like aligned model:*
    - *If* $f_Z(Y_t, Z_{t-1}, N_{Z,t})$ *is of the form* $\alpha * Y_t + N_t$*, where* α *can be any real number, then* $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.
    - *If the time series is stationary and Gaussian, and* $f_Y(X_t, Y_{t-1}, N_{Y,t})$ *is of the form* $\alpha * X_t + N_t$*, where* α *can be any real number, then* $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.

Proof can be found in Appendix B. Roughly speaking, this corollary suggests that if the causal relationship between X/Z and Y is linear, then the information needed to infer $\overline{X}/\overline{Z}$ from $Y_{1:k}$ is completely included in $\overline{Y}$. Further, based on the sufficient condition for conditional independence (refer to Corollary 4.5), we can see that it is not necessary for the entire system to be linear.

5. Simulation Experiments

We conducted five experiments to comprehensively address the various aspects of the aggregation problem. Firstly, we applied widely-used causal discovery methods(PC(Spirtes et al., 2000), FCI(Spirtes et al., 2013), GES(Chickering, 2002)) to aggregation data with 4 variables, enabling readers to grasp the motivation and core issue discussed in this paper. Secondly and thirdly, we conducted experiments on functional consistency(apply Direct LiNGAM(Shimizu et al., 2011)/ANM(Hoyer et al., 2008) to linear/nonlinear data with different aggregation levels) and conditional independence consistency(perform Kernel Conditional Independence test(Zhang et al., 2012) on aggregated data) to bolster the theorems presented in the main text. Fourthly, we carried out an experiment to investigate the impact of the k value and to justify the approximations made in this paper. Fifthly, we performed the PC algorithm with a skeleton prior on aggregated data and consistently obtained correct results, offering a preliminary solution to the aggregation problem and laying the groundwork for future research in this area. Due to the page limit, we present only a limited number of results in the main text. Detailed settings and results of the five experiments can be found respectively in the Appendices: C, D, E, F, and G.

5.1. FCM-based Causal discovery in Linear non-Gaussian case

Here we examine the use of a FCM-based causal discovery method on bivariate temporally aggregated data in the linear non-Gaussian case to distinguish between cause and effect. Specifically, we employ the Direct LiNGAM method to represent FCM-based causal discovery methods.

We perform a simulation experiment on the model $Y_i = 2X_i + N_{Y,i}$, where the noise follow uniform distribution. We then generating the dataset $\overline{X}, \overline{Y}$ with a sample size of 10,000. We apply the Direct LiNGAM method on this data. To investigate the relationship between the aggregation factor k and the performance of the Direct LiNGAM method, we vary the values of k from 1 to 100 to see the correct rate in 100 repetitions.

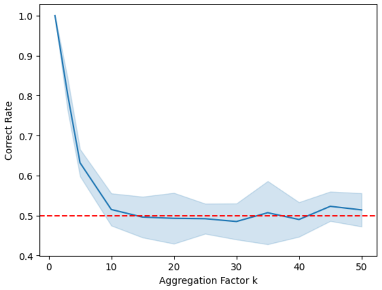

Figure 3. Relationship between the aggregation factor k and the performance of the Direct LiNGAM method. The blue area represents the standard deviation. The red line represents the random guess baseline.

Our results indicate that when k is small, the correct rate is near 100%, implying a good performance of the Direct LiNGAM method. However, as k increases from 3 to 30, the correct rate drops rapidly to 50% as random guess. This experiment demonstrates that even in the linear non-Gaussian case, temporal aggregation can significantly impair the identifiability of functional-based methods relying on non-Gaussianity.

5.2. Conditional Independence Test in Gaussian case

We perform experiments using three different structures: chain, fork, and collider to validate our theoretical results.

In all structures, each noise term N follows an independent and identically distributed (i.i.d) standard Gaussian distribution. For the causal relationship between X and Y, we use the function $f(\cdot, N)$. In the linear case, $f(\cdot) = (\cdot)$. In the nonlinear case, we use the post-nonlinear model $f(\cdot, N) = G(F(\cdot) + N)$ (Zhang & Hyvarinen, 2012) and uniformly randomly pick F and G from $(\cdot)^2$, $(\cdot)^3$, and $tanh(\cdot)$ for each repetition. This is to ensure that our experiment covers a wide range of nonlinear cases. Similarly, the same approach is applied for the relationship between Y and Z with the corresponding function $g(\cdot)$.

We set $t = 1, 2$, thus $\overline{X} = X_1 + X_2$, $\overline{Y} = Y_1 + Y_2$, $\overline{Z} = Z_1 + Z_2$. And we generate 1000 i.i.d. data points for $(X_1, X_2, Y_1, Y_2, \overline{X}, \overline{Y})$ and feed them into the approximate kernel-based conditional independence test (Strobl et al., 2019). We test the null hypothesis(conditional independence) for (I) $\overline{X} \perp\!\!\!\perp \overline{Y}$, (II) $\overline{Y} \perp\!\!\!\perp \overline{Z}$, (III) $\overline{X} \perp\!\!\!\perp \overline{Z}$, (IV) $\overline{X} \perp\!\!\!\perp \overline{Y} \mid \overline{Z}$, (V) $\overline{Y} \perp\!\!\!\perp \overline{Z} \mid \overline{X}$, (VI) $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$. We also test for the conditional independence in corollary 4.5: (A) $\overline{X} \perp\!\!\!\perp Y_1 \mid \overline{Y}$, and (B) $\overline{Z} \perp\!\!\!\perp Y_1 \mid \overline{Y}$. We report the rejection rate for fork structure at a 5% significance level in 100 repeated experiments in Table 3b. The results for chain structure and collider structure can be found in Appendix 3a.

The experiment shows that as long as there is some linearity, we can find a consistent conditional independence set, which aligns with Corollary 4.6. However, in completely nonlinear situations, we still cannot find a consistent conditional independence set, which aligns with Remark 4.3.

6. Conclusion and Limitation

This paper points out that although many people use the occurrence of instantaneous causal relationships due to temporal aggregation as a real-world explanation for instantaneous models, few people pay attention to whether these instantaneous causal relationships are consistent with the underlying time-delayed causal relationships when this situation occurs. This paper mainly discusses whether the causal models generated by temporal aggregation maintain functional consistency and conditional independence consistency in general (nonlinear) situations. Through theoretical analysis in the case of finite k, we show that functional consistency is difficult to achieve in non-linear situations. Furthermore, through theoretical analysis and experimen-

Table 1. Rejection rates for CI tests with different combinations of linear and nonlinear relationships. The index in the box represents the conditional independence that the structure should ideally have. Simply speaking, for the column VI, the closer the rejection rate is to 5%, the better. For all other columns from I to V, a higher rate is better.

$X_t \to Y_t$	$Y_t \to Z_t$	I	II	III	IV	V	$\boxed{\text{VI}}$	A	B
Linear	Linear	100%	100%	100%	100%	100%	5%	6%	5%
Nonlinear	Linear	92%	100%	84%	92%	100%	5%	76%	5%
Linear	Nonlinear	100%	93%	85%	100%	93%	5%	5%	71%
Nonlinear	Nonlinear	92%	93%	72%	86%	87%	58%	72%	74%

tal verification in the case of infinite k, we show that even in linear non-Gaussian situations, the instantaneous model generated by temporal aggregation is still unidentifiable. For conditional independence consistency, we show through sufficient and necessary conditions and experiments that it can be satisfied as long as the causal process has some linearity. However, it is still difficult to achieve in completely non-linear situations.

Limitations: Although the negative impact of temporal aggregation on instantaneous causal discovery has been pointed out, a solution has not been provided.

Acknowledgements

We would like to acknowledge the support from NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Florin Court Capital. MG was supported by ARC DE210101624 and DP240102088.

Impact Statement

This study offers a theoretical analysis of potential challenges when applying non-temporal causal discovery algorithms to temporally-aggregated time series data. It provides valuable insights for researchers involved in algorithm design and serves as guidance for practitioners employing causal discovery algorithms. Importantly, the research operates at an abstract level, avoiding any involvement with sensitive data and raising no ethical concerns.

References

Breitung, J. and Swanson, N. R. Temporal aggregation and spurious instantaneous causality in multiple time series models. *Journal of Time Series Analysis*, 23(6):651–665, 2002.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Darmois, G. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pp. 2–8, 1953.

Drost, F. C. and Nijman, T. E. Temporal aggregation of garch processes. *Econometrica: Journal of the Econometric Society*, pp. 909–927, 1993.

Fisher, F. M. A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pp. 73–92, 1970.

Ghysels, E., Hill, J. B., and Motegi, K. Testing for granger causality with mixed frequency data. *Journal of Econometrics*, 192(1):207–230, 2016.

Gong, M., Zhang, K., Schölkopf, B., Glymour, C., and Tao, D. Causal discovery from temporally aggregated time series. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 2017. NIH Public Access, 2017.

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

Granger, C. W. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

Granger, C. W. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2):199–211, 1988.

Granger, C. W. and Lee, T.-H. The effect of aggregation on nonlinearity. *Econometric reviews*, 18(3):259–269, 1999.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Lacerda, G., Spirtes, P. L., Ramsey, J., and Hoyer, P. O. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Marcellino, M. Some consequences of temporal aggregation in empirical analysis. *Journal of Business & Economic Statistics*, 17(1):129–136, 1999.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pierce, D. A. and Haugh, L. D. Causality in temporal systems: Characterization and a survey. *Journal of econometrics*, 5(3):265–293, 1977.

Rajaguru, G. and Abeysinghe, T. Temporal aggregation, cointegration and causality inference. *Economics Letters*, 101(3):223–226, 2008.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Renault, E., Sekkat, K., and Szafarz, A. Testing for spurious causality in exchange rates. *Journal of Empirical Finance*, 5(1):47–66, 1998.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., Bollen, K., and Hoyer, P. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Silvestrini, A. and Veredas, D. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497, 2008.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Spirtes, P. L., Meek, C., and Richardson, T. S. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Strobl, E. V., Zhang, K., and Visweswaran, S. Approximate kernel-based conditional independence tests for fast nonparametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

Zhang, K. and Hyvarinen, A. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Zhou, D., Zhang, Y., Xiao, Y., and Cai, D. Analysis of sampling artifacts on the granger causality analysis for topology extraction of neuronal dynamics. *Frontiers in computational neuroscience*, 8:75, 2014.

A. Proof for Necessary Condition of Functional Consistency

Theorem A.1 (Construction of $\hat{f}$)**.** *If such $\hat{f}$, as defined in Definition 3.2, exists, then $\hat{f}$ must take the form:*

$$\hat{f}(T) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + c, \tag{4}$$

where c is any constant (which can be incorporated into the noise term) and the expression $\mathbb{E}(\cdot \mid \overline{X} = T)$ denotes the conditional expectation. For simplicity, we set $c = 0$. Consequently, this implies:

$$\mathbb{E}(\hat{f}(\overline{X})) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i)\right). \tag{5}$$

Proof. According to the definition of functional consistency regarding additive noise, we can express $\overline{Y}$ in two ways: $\overline{Y} = \sum_{i=1}^{k} f(X_i) + \sum_{i=1}^{k} N_{Y,i} = \hat{f}(\overline{X}) + N$.

Taking the conditional expectation conditioned on $\overline{X} = T$ for both sides of the equation (noting that both $N_{Y,i}$ and N are independent of $\overline{X}$):

$$\mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + \mathbb{E}\left(\sum_{i=1}^{k} N_{Y,i}\right) = \hat{f}(T) + \mathbb{E}(N). \tag{6}$$

Then, we have:

$$\hat{f}(T) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + \left(\mathbb{E}\left(\sum_{i=1}^{k} N_{Y,i}\right) - \mathbb{E}(N)\right). \tag{7}$$

Please note that $\mathbb{E}\left(\sum_{i=1}^{k} N_{Y,i}\right) - \mathbb{E}(N)$ is a constant, not related to T.

Therefore,

$$\hat{f}(T) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + c. \tag{8}$$

Finally, applying the law of total expectation, we obtain the expectation of $\hat{f}$. □

Theorem A.2 (Necessary and Sufficient Condition)**.** *The necessary and sufficient condition for the existence of the additive noise causal model defined in Definition 3.2 is that $N = \sum_{i=1}^{k} N_{Y,i} + \left(\sum_{i=1}^{k} f(X_i) - \hat{f}(\overline{X})\right)$ is independent of $\overline{X}$, where $\hat{f}$ is defined by Eq. 2.*

Proof. According to Definition 3.1, $\overline{Y}$ is described as

$$\overline{Y} = \sum_{i=1}^{k} f(X_i) + \sum_{i=1}^{k} N_{Y,i}. \tag{9}$$

Furthermore, according to Theorem 3.3, $\overline{Y}$ can be expressed as

$$\overline{Y} = \hat{f}(\overline{X}) + N. \tag{10}$$

By combining Equations 9 and 10, we derive the expression for N as

$$N = \sum_{i=1}^{k} N_{Y,i} + \left(\sum_{i=1}^{k} f(X_i) - \hat{f}(\overline{X})\right). \tag{11}$$

Given the requirement in Definition 3.1 that N should be independent of $\overline{X}$, we thereby establish the theorem.

□

Theorem A.3 (General Case for Functional Consistency Across Different Regions)**.** *For the causal model defined in Definition 3.5, assume the aggregated variables* $\overline{X}_A$*,* $\overline{X}_B$ *and* $\overline{Y}_A$*,* $\overline{Y}_B$ *have continuous support. Then, there exists functions* $\hat{f}$ *and* $\hat{g}$ *such that* $\overline{Y}_A = \hat{f}(\overline{X}_A, N_A)$*,* $\overline{Y}_B = \hat{f}(\overline{X}_B, N_B)$*,* $\overline{X}_A = \hat{g}(\overline{Y}_A, N'_A)$*, and* $\overline{X}_B = \hat{g}(\overline{Y}_B, N'_B)$*, where* N_A *is independent of* $\overline{X}_A$*,* N_B *is independent of* $\overline{X}_B$*,* N'_A *is independent of* $\overline{Y}_A$*, and* N'_B *is independent of* $\overline{Y}_B$*.*

Proof. We only provide the proof for $\overline{Y}_A = \hat{f}(\overline{X}_A, N_A), \overline{Y}_B = \hat{f}(\overline{X}_B, N_B)$ here. And the construction of $\hat{g}$ is very similar to $\hat{f}$.

Aligned with Lemma 1 in (Zhang et al., 2015), for any random variables, we can construct independent noise with $N_A = q \circ F_{\overline{Y}_A|\overline{X}_A}(\overline{Y}_A)$ and $N_B = F_{\overline{Y}_B|\overline{X}_B}(\overline{Y}_B)$. Since q is an arbitrary, strictly increasing function and the support of the random variable is continuous, the CDF is invertible. We can set $q = F_{\overline{Y}_B|\overline{X}_B} \circ F^{-1}_{\overline{Y}_A|\overline{X}_A}$, making N_A constructed by the same function as N_B. Given the CDF's invertibility, we can define $\hat{f} := F^{-1}_{\overline{Y}_B|\overline{X}_B}$ to derive $\overline{Y}_A = \hat{f}(\overline{X}_A, N_A)$, $\overline{Y}_B = \hat{f}(\overline{X}_B, N_B)$.

□

B. Proofs for Conditions of Conditional Independence Consistency

Theorem B.1 (Necessary and Sufficient Condition for Conditional Independence Consistency)**.** *Consider the distribution of* $(\overline{X}, \overline{Y}, \overline{Z}, Y_1, ..., Y_k)$ *entailed from the aligned model, the following statements are equivalent when* $2 \leq k < \infty$*:*

(i) Conditional Independence: $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$

(ii) Conditional Probability Relation: $\forall s_X, s_Y, s_Z \in \mathbb{R}$

$$\iint_{\mathbb{R}^k} \alpha(s_Z, s_Y, y_{1:k}) \left(\beta(y_{1:k}, s_Y, s_X) - \gamma(y_{1:k}, s_Y)\right) dy_1 ... dy_k = 0 \tag{12}$$

(iii) Alternative Conditional Probability Relation: $\forall s_X, s_Y, s_Z \in \mathbb{R}$

$$\iint_{\mathbb{R}^k} \alpha^*(s_X, s_Y, y_{1:k}) \left(\beta^*(y_{1:k}, s_Y, s_Z) - \gamma(y_{1:k}, s_Y)\right) dy_1 ... dy_k = 0 \tag{13}$$

where

- $\alpha(s_Z, s_Y, y_{1:k}) := p_{S_Z|S_Y, Y_{1:k}}(s_Z|s_Y, y_{1:k})$
- $\beta(y_{1:k}, s_Y, s_X) := p_{Y_{1:k}|S_Y, S_X}(y_{1:k}|s_Y, s_X)$
- $\alpha^*(s_X, s_Y, y_{1:k}) := p_{S_X|S_Y, Y_{1:k}}(s_X|s_Y, y_{1:k})$
- $\beta^*(y_{1:k}, s_Y, s_Z) := p_{Y_{1:k}|S_Y, S_Z}(y_{1:k}|s_Y, s_Z)$
- $\gamma(y_{1:k}, s_Y) := p_{Y_{1:k}|S_Y}(y_{1:k}|s_Y)$

Proof. The proof will proceed by showing that statements (i), (ii), and (iii) are mutually equivalent.

Proof that (i) is equivalent to (ii): Suppose that $S_X \perp\!\!\!\perp S_Z \mid S_Y$ holds. By the definition of conditional independence, this is equivalent to the statement that for all s_X, s_Y, and s_Z in $\mathbb{R}$, we have $p_{S_Z|S_Y, S_X}(s_Z|s_Y, s_X) = p_{S_Z|S_Y}(s_Z|s_Y)$.

We can now derive both sides of this equation as follows.

On the left hand side(LHS):

$$\begin{aligned}
&p_{S_Z|S_Y,S_X}(s_Z|s_Y,s_X)\\
&=\frac{p_{S_Z,S_Y,S_X}(s_Z,s_Y,s_X)}{p_{S_Y,S_X}(s_Y,s_X)}\\
&=\frac{\iint_{\mathbb{R}^k} p_{S_Z|S_Y,S_X,Y_{1:k}}(s_Z|s_Y,s_X,y_{1:k})p_{S_Y,S_X,Y_{1:k}}(s_Y,s_X,y_{1:k})dy_1...dy_k}{p_{S_Y,S_X}(s_Y,s_X)}\\
&=\iint_{\mathbb{R}^k} p_{S_Z|S_Y,Y_{1:k}}(s_Z|s_Y,y_{1:k})p_{Y_{1:k}|S_Y,S_X}(y_{1:k}|s_Y,s_X)dy_1...dy_k\\
&=\iint_{\mathbb{R}^k} \alpha(s_Z,s_Y,y_{1:k})\beta(y_{1:k},s_Y,s_X)dy_1...dy_k
\end{aligned}$$

The first steps is based on the definition of conditional probability and the second step uses the law of total probability. The third step is using the d-separation: $\{Y_1,\ldots,Y_k\}$ d-separates S_Z from S_X.

Meanwhile, RHS:

$$\begin{aligned}
&p_{S_Z|S_Y}(s_Z|s_Y)\\
&=\frac{p_{S_Z,S_Y}(s_Z,s_Y)}{p_{S_Y}(s_Y)}\\
&=\frac{\iint_{\mathbb{R}^k} p_{S_Z|S_Y,Y_{1:k}}(s_Z|s_Y,y_{1:k})p_{S_Y,Y_{1:k}}(s_Y,y_{1:k})dy_1...dy_k}{p_{S_Y}(s_Y)}\\
&=\iint_{\mathbb{R}^k} \alpha(s_Z|s_Y,y_{1:k})\gamma(y_{1:k},s_Y)dy_1...dy_k
\end{aligned}$$

Finally, substitute both to the original equality:

$$\begin{aligned}
&p_{S_Z|S_Y,S_X}(s_Z|s_Y,s_X)-p_{S_Z|S_Y}(s_Z|s_Y)\\
&=\iint_{\mathbb{R}^k} \alpha(s_X,s_Y,y_{1:k})\left(\beta(y_{1:k},s_Y,s_Z)-\gamma(y_{1:k},s_Y)\right)\,dy_1\ldots dy_k\\
&=0
\end{aligned}$$

Hence, we arrive at the condition specified in (ii).

Proof that (i) is equivalent to (iii): The proof that (i) and (iii) are equivalent is analogous to the above arguments. We therefore omit the details for brevity.

□

Corollary (Sufficient Conditions for Conditional Independence). If $\{\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ **or** $\{\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ hold, then $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$ holds.

Proof. This corollary introduces two sufficient conditions for conditional independence. While the proofs for each are analogous, we demonstrate the proof for the first condition to avoid redundancy.

Proof that $\{\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ *implies* $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$:

Assume that $\{\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}\}$ is true. By definition, this is equivalent to

$$p_{Y_{1:k}|S_Y,S_X}(y_{1:k}|s_Y,s_X)=p_{Y_{1:k}|S_Y}(y_{1:k}|s_Y)$$

for all s_X, $y_{1:k}$, and s_Y. Utilizing the notation from Theorem 4.4, we can rewrite this as

$$\beta(y_{1:k}, s_Y, s_X) = \gamma(y_{1:k}, s_Y).$$

This simplification makes it clear that the equality conforms to the second condition of Theorem 4.4, which is

$$\iint_{\mathbb{R}^k} \alpha(s_Z, s_Y, y_{1:k}) \left(\beta(y_{1:k}, s_Y, s_X) - \gamma(y_{1:k}, s_Y)\right)\, dy_1 ... \, dy_k = 0.$$

Because this holds for all s_X, s_Y, and s_Z in $\mathbb{R}$, it follows that $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$ is true, completing our proof.

□

Corollary. 1. For a fork-like aligned model:

- If $f_Z(Y_t, Z_{t-1}, N_{Z,t})$ is of the form $\alpha * Y_t + N_t$, where α can be any real number, then $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.
- If $f_X(X_{t-1}, Y_t, N_{X,t})$ is of the form $\alpha * Y_t + N_t$, where α can be any real number, then $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.

2. For a chain-like aligned model:

- If $f_Z(Y_t, Z_{t-1}, N_{Z,t})$ is of the form $\alpha * Y_t + N_t$, where α can be any real number, then $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.
- If the time series is stationary and Gaussian, and $f_Y(X_t, Y_{t-1}, N_{Y,t})$ is of the form $\alpha * X_t + N_t$, where α can be any real number, then $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.

Proof. The proof for these four sufficient conditions is tied to the bivariate substructure within the fork and chain models.

We categorize the bivariate substructures within these trivariate structures into two types. The first type is where the middle node directs the side nodes, such as in the fork structure where the middle node Y directs X and Z. There are two such substructures in the fork model and one in the chain model where Y directs Z. The second type is where the side node directs the middle node, seen in the chain model with X directing Y.

Due to the causal direction in the bivariate structure, the sufficient conditions for $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ and $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in the fork, and $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in the chain are similar and share a similar proof. The sufficient condition for $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in the chain is different from the other three. To avoid redundancy, we provide a proof for the sufficient condition for $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in the chain; the proof for the two conditions in the fork model is similar to this. We also provide the proof for the sufficient condition for $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in the chain.

proof for sufficient condition of $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in chain model:

Suppose $f_Z(Y_t, N_{Z,t}) = \alpha * Y_t + N_{Z,t}$ for some real number $alpha$. Then, by substitution, we get

$$\begin{aligned} S_Z &= \sum_{t=1}^{k} Z_t \\ &= \sum_{t=1}^{k} (\alpha Y_t + N_{Z,t}) \\ &= \alpha S_Y + \sum_{t=1}^{k} N_{Z,t} \end{aligned}$$

Given S_Y, the random part of S_Z is only $\sum_{t=1}^{k} N_{Z,t}$, which is independent of $Y_{1:k}$. Therefore, it follows that $\overline{Z} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$.

proof for sufficient condition of $\overline{X} \perp\!\!\!\perp Y_{1:k} \mid \overline{Y}$ in chain model:

We will prove the case for $k = 2$, and it can be easily generalized to $k \geq 3$. In the linear, Gaussian, stationary model for $k = 2$, we have:

$$
\begin{aligned}
X_1 &\sim \mathcal{N}(0, \sigma_{X_1}^2) \\
Y_1 &= \alpha X_1 + N_{Y,1} \\
X_2 &= \beta X_1 + N_{X,2} \\
Y_2 &= \alpha X_2 + N_{Y,2}
\end{aligned}
$$

where $N_{X,2} \sim \mathcal{N}(0, \sigma_{N_X}^2)$, $N_{Y,1}, N_{Y,2}$ i.i.d. $\sim \mathcal{N}(0, \sigma_{N_Y}^2)$. And due to stationarity, $\sigma_{X_1}^2 = \frac{\sigma_{N_X}^2}{1-\beta^2}$.

In the linear Gaussian case, conditional independence implies that the partial correlation equals 0. We have:

$$
\begin{aligned}
\text{cov}_{S_X,Y_1|S_Y} &= \text{cov}_{S_X,Y_1} - \frac{\text{cov}_{S_X,S_Y}\text{cov}_{Y_1,S_Y}}{\text{var}_{S_Y}} \\
\text{cov}_{S_X,Y_2|S_Y} &= \text{cov}_{S_X,Y_2} - \frac{\text{cov}_{S_X,S_Y}\text{cov}_{Y_2,S_Y}}{\text{var}_{S_Y}}
\end{aligned}
$$

where

$$
\begin{aligned}
\text{cov}(S_X, Y_1) &= \text{cov}(X_1, Y_1) + \text{cov}(X_2, Y_1), \\
\text{cov}(S_X, Y_2) &= \text{cov}(X_1, Y_2) + \text{cov}(X_2, Y_2), \\
\text{cov}(S_X, S_Y) &= \text{cov}(S_X, Y_1) + \text{cov}(S_X, Y_2), \\
\text{cov}(Y_1, S_Y) &= \text{var}(Y_1) + \text{cov}(Y_1, Y_2), \\
\text{cov}(Y_2, S_Y) &= \text{var}(Y_2) + \text{cov}(Y_1, Y_2).
\end{aligned}
$$

Substitute these into the partial covariance equations to get

$\text{cov}_{S_X,Y_1 \cdot Z} = \text{cov}_{S_X,Y_2 \cdot Z} = 0$

$\square$

C. Causal Discovery from Aggregated Data

To investigate the direct effects of aggregation on causal discovery, we applied three widely-used causal discovery methods on both the original and aggregated data, comparing the results in both linear and nonlinear scenarios. The performance of these methods is measured using the correct rate over 100 repetitions.

For data generation, in each repetition, the original data is generated based on the causal graph shown in Figure 4. The causal relationships are defined as:

$$
\begin{aligned}
Z &= X + Y + N_Z, \\
H &= Z + N_H \quad \text{(for linear)}; \\
Z &= X^2 + Y^2 + N_Z, \\
H &= Z^2 + N_H \quad \text{(for nonlinear)}.
\end{aligned}
$$

The aggregated data is the result of aggregation with a factor of $k = 2$ based on the aligned model 4.1, having an instant structure resembling the original data. All datasets have a sample size of 500.

Regarding the method parameters, we used the Fisher-Z test for PC and FCI, and the BIC score for GES in the linear scenario. In the nonlinear scenario, we set the conditional independence test for PC and FCI as the Kernel Conditional Independence Test (KCI) with the default kernel and chose the “local score CV general” score function for GES. All other settings are kept at their default values.

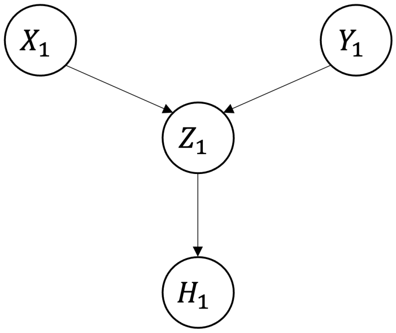

Figure 4. Causal graph of original data

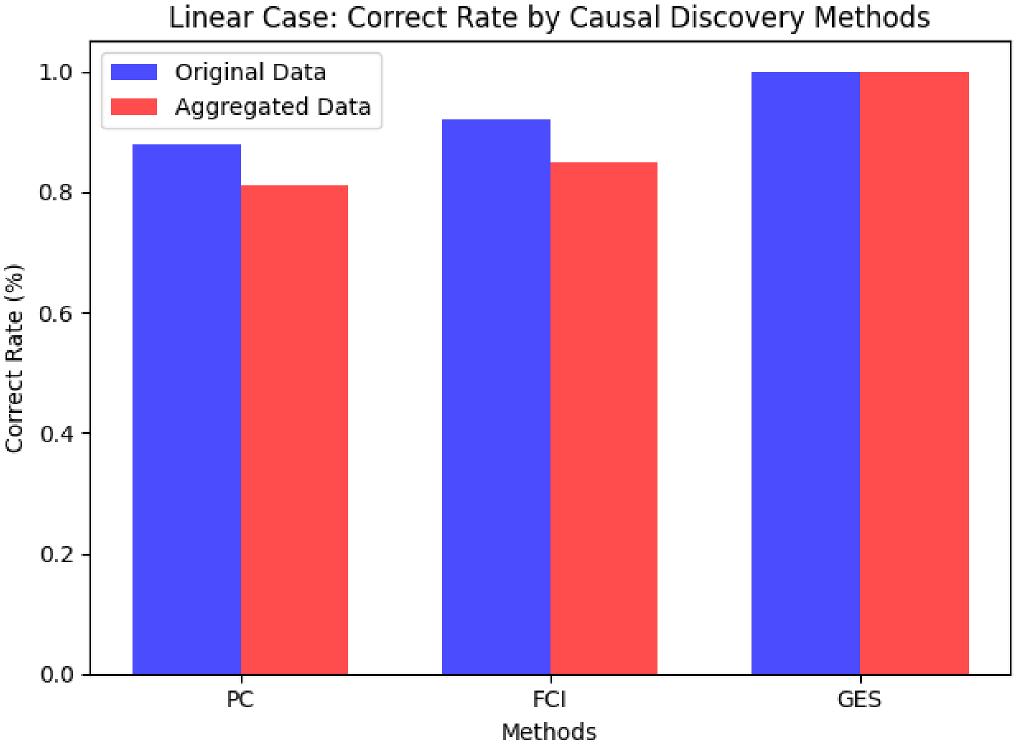

Figure 5. Linear Case: Correction Rate by Causal Discovery Method

From Figure 5, we observe that, in the linear scenario, aggregation does not adversely affect the performance of the causal discovery methods. This might explain why the causal community has not prioritized the aggregation issue in instantaneous causal discovery for a long time.

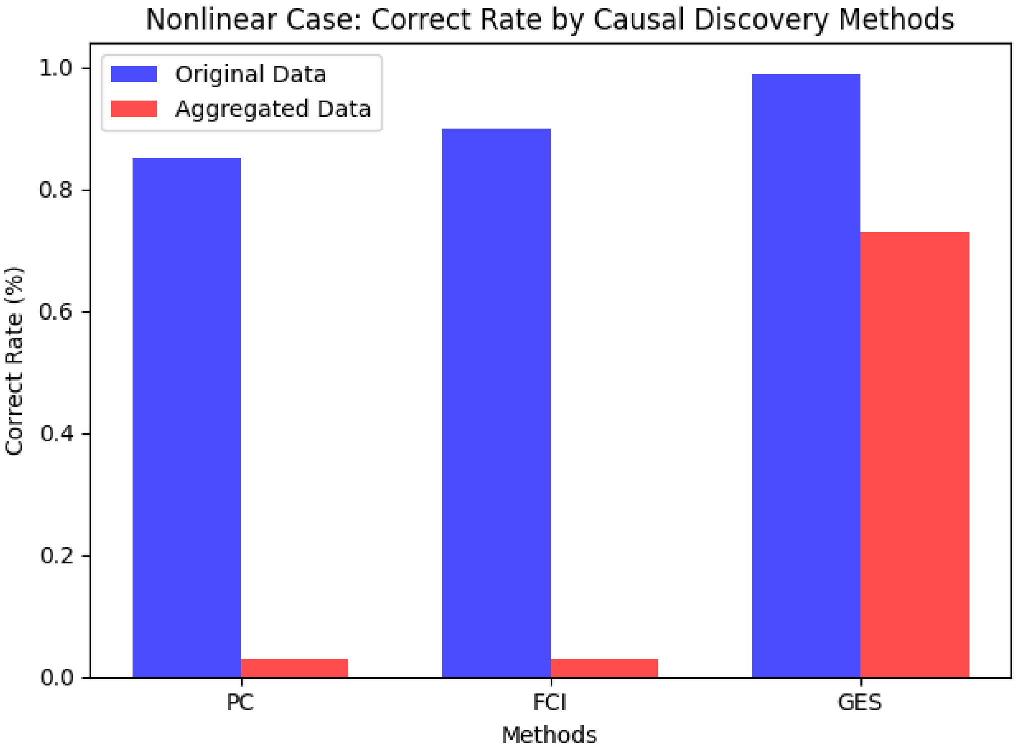

Figure 6. Nonlinear Case: Correction Rate by Causal Discovery Method

Contrastingly, the nonlinear scenario paints a completely different picture, with aggregation causing a significant drop in the performance of all three methods. It is crucial to rigorously investigate this issue to understand its causes and potential solutions.

Additionally, Figure 6 shows the comparison of experimental results of PC, FCI, and GES on nonlinear disaggregated/aggregated data. All three methods have near 90% accuracy for disaggregated data but experience a significant collapse when they come to aggregated data. The GES method is least affected, still maintaining an accuracy of around 70%. Therefore, we conducted more experiments to test whether this is due to coincidence or because GES truly has the ability to handle aggregated data.

Motivated by Reisach et al. (2021), some score-based methods are sensitive to variance, which may result in falsely high performance. So, we changed the variance of some independent noise in the underlying model:

Underlying model: $X_t = N_{X,t},\ Y_t = N_{Y,t},\ Z_t = X_t^2 + Y_t^2 + N_{Z,t},\ H_t = Z_t^2 + N_{H,t}.$
All the independent noises N iid follow $N(0, 1).$
We respectively changed $N_{X,t}$ and $N_{Z,t}$ to $N(0, 4),$ keeping other settings unchanged to perform experiments.

Experimental Change	Disaggregated Accuracy (50 reps)	Aggregated Accuracy
Increase VAR($N_{X,t}$)	0.98 ± 0.03	0.13 ± 0.04
Increase VAR($N_{Z,t}$)	0.94 ± 0.04	0.28 ± 0.09

Table 2. The accuracy was calculated by repeating the experiment 50 times, and in the outside loop, we repeated this process 5 times to calculate the standard deviation for the accuracy. The results were rounded to two decimal places.

From the results, we find that even when we increase the variance of one of the independent noises, GES performs very well on disaggregated data. However, when it comes to aggregated data, the accuracy collapses to lower than 30%.

Therefore, there is no solid evidence to confirm that GES has the ability to deal with aggregated data, but it does have less collapse than other methods in some cases, and it may need further investigation in future work.

D. Experiment for Functional Consistency

Determining the causal direction between two variables is an essential task in causal discovery. To understand the impact of aggregation on this task, we employed two renowned FCM-based causal discovery methods: Direct LiNGAM for the linear scenario and Additive Nonlinear Model (ANM) for the nonlinear one. We assessed how the correct rate in 100 repetitions varies with the aggregation factor k.

For data generation, it's straightforward. The data is generated based on an aligned model with an instantaneous causal relationship, given by:

$$\begin{aligned} Y &= 2X + N_Y \quad \text{(for linear);} \\ Y &= X^2 + N_Y \quad \text{(for nonlinear),} \end{aligned}$$

where the independent noise follows a standard uniform distribution. The sample size is 500.

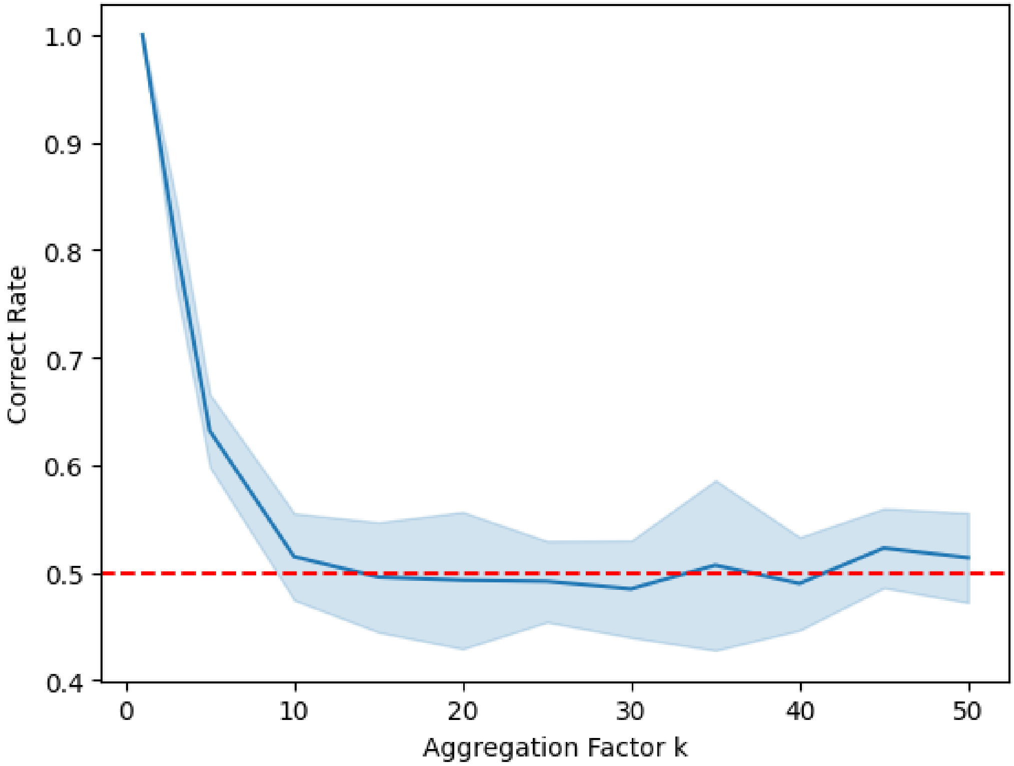

Figure 7. Linear Case: Direct LiNGAM Correction Rate with Different Aggregation Factors k. The blue area represents the standard deviation. The red line represents the random guess baseline.

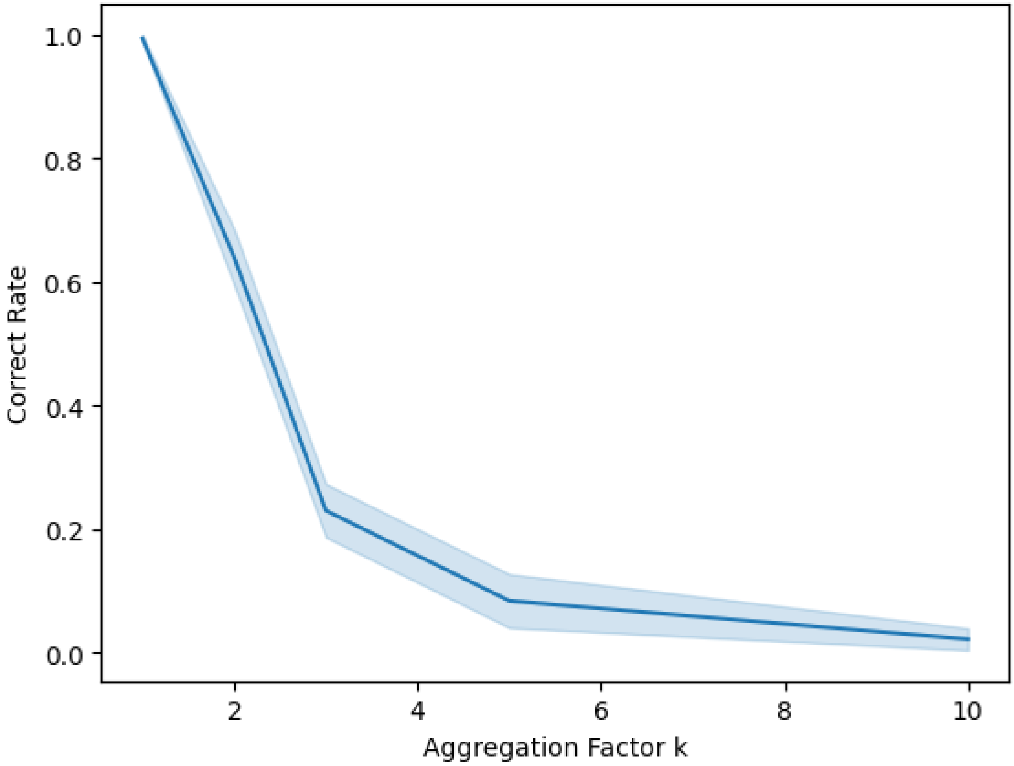

Figure 8. Nonlinear Case: ANM Correction Rate with Different Aggregation Factors k. The blue area represents the standard deviation.

From the presented figures, it's evident that in the linear non-Gaussian case, the non-Gaussian distribution increasingly approaches a Gaussian one as k grows. Eventually, Direct LiNGAM resembles a random guess. It's noteworthy that the x-axis range for the linear scenario spans from 0 to 50, while for the nonlinear case, it's from 0 to 10. This difference suggests that the performance of ANM deteriorates faster than Direct LiNGAM. ANM is not reliant on non-Gaussian properties but on additive noise. Aggregated data lacks functional consistency as the additive noise function is disrupted by the aggregation, rendering ANM ineffective.

E. Experiment for Conditional Independence Consistency

We conduct experiments using three different structures: chain, fork, and collider, to validate our theoretical results in Section 4. However, due to space constraints, we only report the results for the fork-like model in the main text. In this section, we will reiterate the experiment details and report the complete results for chain, fork, and collider structures, along with a more detailed analysis.

The specific settings for these structures are as follows:

Chain-like Model: $X_t = N_{X,t}$, $Y_t = f(X_t, N_{Y,t})$, $Z_t = g(Y_t, N_{Z,t})$

Fork-like Model: $X_t = f(Y_t, N_{X,t})$, $Y_t = N_{Y,t}$, $Z_t = g(Y_t, N_{Z,t})$

Collider-like Model: $X_t = N_{X,t}$, $Y_t = f(X_t, N_{Y,t}) + g(Z_t, N_{Y,t})$, $Z_t = N_{Z,t}$

In all structures, each noise term N follows an independent and identically distributed (i.i.d) standard Gaussian distribution. For the causal relationship between X and Y, we use the function $f(\cdot, N)$. In the linear case, $f(\cdot) = (\cdot)$. In the nonlinear case, we use the post-nonlinear model $f(\cdot, N) = G(F(\cdot) + N)$ (Zhang & Hyvarinen, 2012) and uniformly randomly pick F and G from $(\cdot)^2$, $(\cdot)^3$, and $tanh(\cdot)$ for each repetition. This is to ensure that our experiment covers a wide range of nonlinear cases. Similarly, the same approach is applied for the relationship between Y and Z with the corresponding function $g(\cdot)$.

And $t = 1, 2$, $\overline{X} = X_1 + X_2$, $\overline{Y} = Y_1 + Y_2$, $\overline{Z} = Z_1 + Z_2$. And we generate 1000 i.i.d. data points for $(X_1, X_2, Y_1, Y_2, \overline{X}, \overline{Y})$ and feed them into the approximate kernel-based conditional independence test (Strobl et al., 2019). We test the null hypothesis(conditional independence) for (I) $\overline{X} \perp\!\!\!\perp \overline{Y}$, (II) $\overline{Y} \perp\!\!\!\perp \overline{Z}$, (III) $\overline{X} \perp\!\!\!\perp \overline{Z}$, (IV) $\overline{X} \perp\!\!\!\perp \overline{Y} \mid \overline{Z}$, (V) $\overline{Y} \perp\!\!\!\perp \overline{Z} \mid \overline{X}$, (VI) $\overline{X} \perp\!\!\!\perp \overline{Z} \mid \overline{Y}$. And we also test for the conditional independence in corollary 4.5: (A) $\overline{X} \perp\!\!\!\perp Y_1 \mid \overline{Y}$, and (B) $\overline{Z} \perp\!\!\!\perp Y_1 \mid \overline{Y}$. We report the rejection rate, rounded to the nearest percent, at a 5% significance level in 1000 repeated experiments in Table 3a, 3b, 3c.

Table 3. Rejection rates for CI tests with different combinations of linear and nonlinear relationships. The index in the box represents the conditional independence that the structure should ideally have. Simply speaking, for the column with the index in the box, the closer the rejection rate is to 5%, the better. For all other columns from I to VI, a higher rate is better.

(a) Chain Structure

$X_t \to Y_t$	$Y_t \to Z_t$	I	II	III	IV	V	$\boxed{\text{VI}}$	A	B
Linear	Linear	100%	100%	100%	100%	100%	6%	4%	7%
Nonlinear	Linear	92%	100%	89%	63%	100%	9%	27%	10%
Linear	Nonlinear	100%	95%	87%	100%	94%	5%	6%	82%
Nonlinear	Nonlinear	92%	86%	56%	89%	85%	18%	27%	39%

(b) Fork Structure

$X_t \to Y_t$	$Y_t \to Z_t$	I	II	III	IV	V	$\boxed{\text{VI}}$	A	B
Linear	Linear	100%	100%	100%	100%	100%	5%	6%	5%
Nonlinear	Linear	92%	100%	84%	92%	100%	5%	76%	5%
Linear	Nonlinear	100%	93%	85%	100%	93%	5%	5%	71%
Nonlinear	Nonlinear	92%	93%	72%	86%	87%	58%	72%	74%

(c) Collider Structure

$X_t \to Y_t$	$Y_t \to Z_t$	I	II	$\boxed{\text{III}}$	IV	V	VI	A	B
Linear	Linear	100%	100%	5%	100%	100%	99%	4%	5%
Nonlinear	Linear	95%	89%	5%	96%	91%	51%	17%	56%
Linear	Nonlinear	90%	95%	5%	91%	96%	48%	54%	15%
Nonlinear	Nonlinear	81%	81%	6%	83%	81%	29%	26%	26%

This experiment support our theoretical results, suggesting that conditional independence consistency can be ensured even with some nonlinearity in the model.

Specifically, let's examine the results for chain and fork. We anticipate the tested conditional independence set to contain only $S_X \perp\!\!\!\perp S_Z \mid S_Y$. If so, we can assert that temporal aggregation maintains conditional independence consistency.

From the first and second tables for chain and fork, it's evident that when the model is entirely nonlinear, the results for conditional independence can be erroneous. For instance, the rejection rate for conditional independence that should have been rejected is not high. In the chain structure, the rejection rate for the conditional independence III is zero, implying that every conditional independence test wrongly accepted this conditional independence (type II error). Conversely, the conditional independence that should have been accepted, VI $S_X \perp\!\!\!\perp S_Z \mid S_Y$, has rejection rates of 62% (chain) and 36% (fork), significantly exceeding the significance level of 5%. This aligns with our conclusion in remark 4.3, stating that chain and fork models cannot guarantee conditional independence consistency in general cases.

However, when half the model is linear, all conditional independence that should be rejected exhibit high rejection rates, indicating fewer type II errors. Moreover, the rejection rate for the acceptable conditional independence VI is quite low, closely approximating the significance level of 5%. This suggests that conditional independence-based causal discovery methods can still be applied to temporally aggregated data when the system is partially linear.

Columns A and B primarily aim to validate corollary 4.6 and corollary 4.5. The conditional independence represented by

A and B corresponds to the two sufficient conditions for conditional independence consistency in corollary 4.5. From the experimental results, we find that if one of these conditions holds, we can ensure conditional independence consistency. For example, in the fork results, under the nonlinear+linear case, B holds while A does not. Nonetheless, we still have conditional independence consistency. Moreover, our findings further verify corollary 4.6, indicating that when a certain part of the causal mechanism is linear, the corresponding sufficient condition in this part is satisfied, ultimately ensuring the conditional independence consistency of the entire system.

Finally, looking at the collider results, conditional independence consistency is maintained under all nonlinear and linear combinations, which agrees with our conclusion in remark 4.3, stating that collider can ensure conditional independence consistency under general conditions.

F. Effect of k Value

Gong et al. (2017) have proven that the time-delay causal model (time series data) will transform into an instantaneous causal model (i.i.d. data). In our paper, we utilize large k values to approximate the aggregation of the time-delay model as the aggregation of the instantaneous model (aligned model as defined in the main text). We aim to demonstrate the reasonableness of this approximation and to show how quickly the time-delay model transitions to an instantaneous model.

We apply the GES method on a linear fork-like time-delay model (VAR) and a fork-like instantaneous model (aligned model) across different values of k.

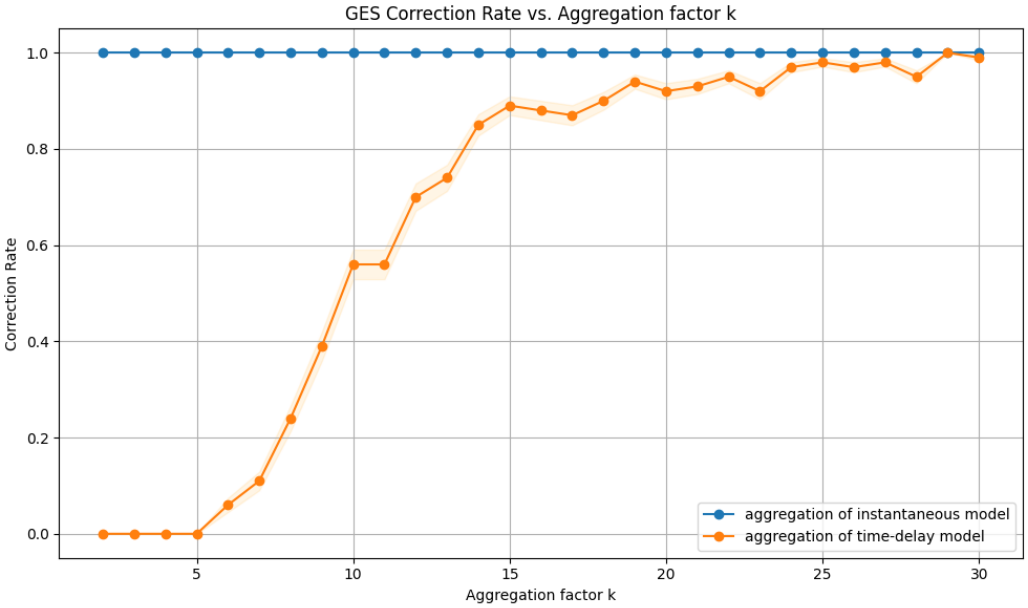

Figure 9. GES Correct Rate vs. Factor k

From Figure 9, we observe that k does not need to be infinite. A sufficiently large k, such as 20, ensures that the time-delay model becomes an instantaneous model detectable by the instantaneous causal discovery method.

G. Causal Discovery on Aggregated Data with Prior Knowledge

Inspired by Remark 4.3, we propose a straightforward solution to ensure that the PC algorithm identifies the correct Markov equivalence. The remark indicates that the collider structure retains conditional independence consistency. This implies that the PC algorithm can determine the correct v-structure if it has the correct skeleton. We therefore conducted experiments to assess whether the PC with a given skeleton as prior knowledge can discern the correct Markov equivalence.

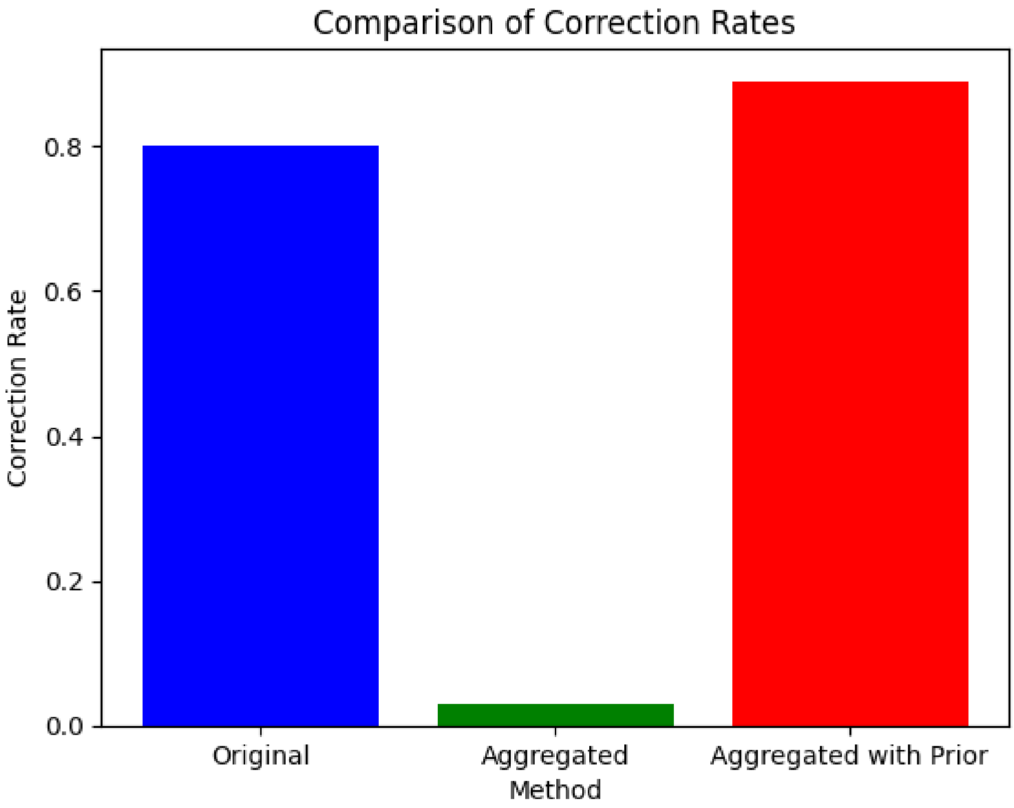

Figure 10. Correct Rate: PC on Original Data vs. PC on Aggregated Data vs. PC with Prior on Aggregated Data

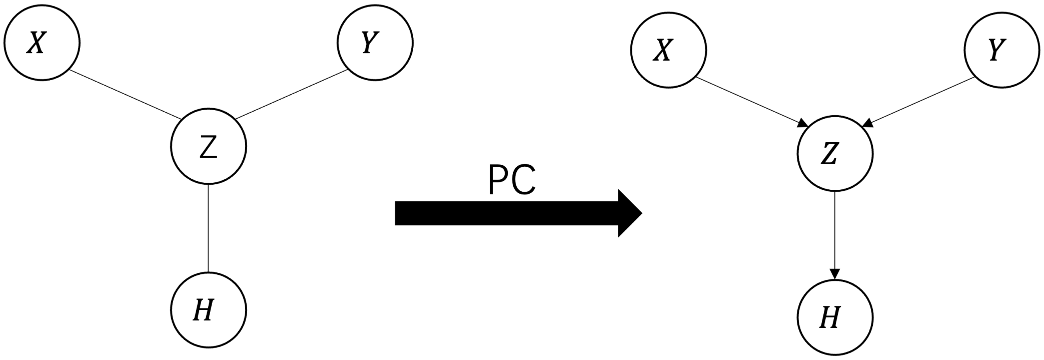

Figure 11. PC can find correct v-structure on aggregated data

From Figure 10, it becomes evident that when the PC is applied directly to aggregated data, its performance is subpar. However, when provided with the skeleton as prior knowledge, the PC consistently identifies the correct v-structure, yielding accurate results(see Figure 11). This discovery suggests that future work should concentrate on addressing the aggregation problem during the skeleton discovery process.

H. Example for Inconsistency in the Nonlinear Additive Noise Model

Consider the nonlinear additive noise models $X_{t,A} = N_{X,t,A}, Y_{t,A} = X_{t,A}^3 + N_{t,A}$ and $X_{t,B} = N_{X,t,B}, Y_{t,B} = X_{t,B}^3 + N_{t,B}$ for regions A and B, where $t = 1, 2$, $N_{X,t,A} \sim N(0, 1)$, $N_{X,t,B} \sim N(0, 2)$, g(k)=1.

According to Theorem 3.3, if we require

$$\overline{Y}_A = \hat{f}(\overline{X}_A) + N_A$$

, then it must be

$$\hat{f}(T) = \mathbb{E}\left(\sum_{i=1}^{k} f(X_i) \mid \overline{X} = T\right) + c$$

here we incorporate c into the noise term, and substituting the definition of the causal function in this example, then we have:

$$\hat{f}(T) = \mathbb{E}\left(X_1^3 + X_2^3 \mid X_1 + X_2 = T\right)$$

And according to the formula for conditional density, we can derive when X_1,X_2 i.i.d. follow $N(0, \sigma^2)$, and conditioning on $X_1 + X_2 = T$, the conditional distribution of X_1 or X_2 is $N(\frac{T}{2}, \frac{\sigma^2}{2})$.

And the 3rd raw moment for $N(\mu', \sigma'^2)$ is $\mu'^3 + 3\mu'\sigma'^2$. Thus, we substitute the conditional expectation and conditional variance to derive (excluding the constant term):

$$\hat{f}(T) = \frac{3T\sigma^2}{4}$$

Therefore,

$$\hat{f}(T) = \frac{3T}{4}$$

for region A, and

$$\hat{f}(T) = 3T$$

for region B, which are inconsistent.

I. Theorem and Proof for Approximating Time-delay Causal Effect by Instantaneous Causal Effect

Theorem I.1 (Asymptotic Equivalence of Time-Delayed and Instantaneous Models)**.** *Given an s-dimensional independent noise random vector sequence $N_0, N_1, N_2, \ldots$ that are i.i.d. and follow a distribution with zero mean and finite variance, and given that matrix B is an $s \times s$ square matrix with $\|B\| < 1$ (all norms here are l^2 norms), and $g(k)$ is always positive and satisfies $\lim_{k\to\infty} g(k) = \infty$.*

Consider the time-delayed model:

$$X_0 = N_0$$

$$X_t = BX_{t-1} + N_t$$

and the corresponding instantaneous model:

$$Y_t = BY_t + N_t$$

Then, when $k \to \infty$,

$$\frac{1}{g(k)}\sum_{t=1}^{k} Y_t - \frac{1}{g(k)}\sum_{t=1}^{k} X_t \xrightarrow{P} 0$$

Proof. Since $\|B\| < 1$, then $I - B$ is invertible.

Thus,

$$Y_t = (I - B)^{-1} N_t$$

and

$$\sum_{t=1}^{k} Y_t = \sum_{t=1}^{k} (I - B)^{-1} N_t \tag{14}$$

$$= \sum_{t=1}^{k} (I + B + B^2 + B^3 + \ldots) N_t \tag{15}$$

According to

$$X_1 = BX_0 + N_1$$

$$X_2 = BX_1 + N_2 = B^2 N_0 + BN_1 + N_2$$

$$\ldots$$

we have

$$X_t = \sum_{i=0}^{t} B^{t-i} N_i$$

then

$$\sum_{t=1}^{k} X_t = \sum_{t=1}^{k}\sum_{i=0}^{t} B^{t-i} N_i \tag{16}$$

$$= \sum_{i=1}^{k}\sum_{t=i}^{k} B^{t-i} N_i + \sum_{t=1}^{k} B^t N_0 \tag{17}$$

$$= \sum_{i=1}^{k}\left(\sum_{h=0}^{k-i} B^h\right) N_i + \sum_{t=1}^{k} B^t N_0 \text{ (let } h = t - i) \tag{18}$$

Therefore, combining these equations,

$$\frac{1}{g(k)}\sum_{t=1}^{k} Y_t - \frac{1}{g(k)}\sum_{t=1}^{k} X_t = \frac{1}{g(k)}\left(\sum_{t=1}^{k}\left(\sum_{i=k-t+1}^{\infty} B^i\right) N_t - \left(\sum_{i=1}^{k} B^i\right) N_0\right)$$

$$\begin{aligned}
&\left\| \frac{1}{g(k)} \sum_{t=1}^{k} Y_t - \frac{1}{g(k)} \sum_{t=1}^{k} X_t \right\| \\
&\leq \frac{1}{g(k)} \left(\sum_{t=1}^{k} \left(\sum_{i=k-t+1}^{\infty} \|B\|^i \right) \|N_t\| + \left(\sum_{i=1}^{k} \|B\|^i \right) \|N_0\| \right) \\
&= \frac{1}{g(k)} \left(\sum_{t=1}^{k} \frac{\|B\|^{k-t+1}}{1-\|B\|} \|N_t\| + \frac{\|B\| - \|B\|^{k+1}}{1-\|B\|} \|N_0\| \right) \text{ (the sum of a geometric series)} \\
&= \frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \|N_t\| + (\|B\| - \|B\|^{k+1}) \|N_0\| \right)
\end{aligned}$$

Because $N_0, N_1, N_2, \ldots$ are i.i.d. and follow a distribution with zero mean and finite variance, the mean and variance for $\|N_t\|$ are also finite, denoted as μ_{norm} and σ^2_{norm}, respectively.

Let's prove the expectation and variance of $\frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \|N_t\| + (\|B\| - \|B\|^{k+1}) \|N_0\| \right)$ tend to 0 when $k \to \infty$.

Then,

$$\begin{aligned}
&\mathbb{E} \left(\frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \|N_t\| + (\|B\| - \|B\|^{k+1}) \|N_0\| \right) \right) \\
&= \frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \mu_{\text{norm}} + (\|B\| - \|B\|^{k+1}) \mu_{\text{norm}} \right) \\
&= \frac{1}{(1-\|B\|)g(k)} \left(\frac{\|B\| - \|B\|^{k+1}}{1-\|B\|} \mu_{\text{norm}} + (\|B\| - \|B\|^{k+1}) \mu_{\text{norm}} \right)
\end{aligned}$$

Because $\|B\| < 1$ and $g(k) \to \infty$ as $k \to \infty$, we have

$$\frac{1}{(1-\|B\|)g(k)} \left(\frac{\|B\| - \|B\|^{k+1}}{1-\|B\|} \mu_{\text{norm}} + (\|B\| - \|B\|^{k+1}) \mu_{\text{norm}} \right) \to 0$$

and

$$\begin{aligned}
&\text{VAR} \left(\frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \|N_t\| + (\|B\| - \|B\|^{k+1}) \|N_0\| \right) \right) \\
&= \frac{1}{((1-\|B\|)g(k))^2} \left(\sum_{t=1}^{k} \|B\|^{2(k-t+1)} \sigma^2_{\text{norm}} + (\|B\| - \|B\|^{k+1})^2 \sigma^2_{\text{norm}} \right) \\
&= \frac{1}{((1-\|B\|)g(k))^2} \left(\frac{\|B\|^2 - \|B\|^{2k+2}}{1-\|B\|^2} \sigma^2_{\text{norm}} + (\|B\| - \|B\|^{k+1})^2 \sigma^2_{\text{norm}} \right)
\end{aligned}$$

Because $\|B\| < 1$ and $g(k) \to \infty$ as $k \to \infty$, we can also conclude the variance tends to 0.

Therefore, according to Chebyshev's inequality, for any $\epsilon > 0$, when $k \to \infty$,

$$P \left(\frac{1}{(1-\|B\|)g(k)} \left(\sum_{t=1}^{k} \|B\|^{k-t+1} \|N_t\| + (\|B\| - \|B\|^{k+1}) \|N_0\| \right) > \epsilon \right) \to 0$$

therefore, for any $\epsilon > 0$, when $k \to \infty$,

$$
\begin{aligned}
&P\left(\left\|\frac{1}{g(k)}\sum_{t=1}^{k}Y_t-\frac{1}{g(k)}\sum_{t=1}^{k}X_t\right\|>\epsilon\right)\\
&\leq P\left(\frac{1}{(1-\|B\|)g(k)}\left(\sum_{t=1}^{k}\|B\|^{k-t+1}\|N_t\|+(\|B\|-\|B\|^{k+1})\|N_0\|\right)>\epsilon\right)\to 0
\end{aligned}
$$

Thus, we conclude convergence in probability.

□