Journal of Machine Learning Research 26 (2025) 1-62 Submitted 3/23; Revised 2/25; Published 2/25

Directed Cyclic Graphs for Simultaneous Discovery of Time-Lagged and Instantaneous Causality from Longitudinal Data Using Instrumental Variables

Wei Jin WJIN@JHU.EDU
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218, USA

Yang Ni YNI@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77843, USA

Amanda B. Spence ABS132@GEORGETOWN.EDU
Department of Medicine
Georgetown University
Washington, DC 20007, USA

Leah H. Rubin LRUBIN@JHU.EDU
Departments of Neurology and Psychiatry
Johns Hopkins University School of Medicine
Baltimore, MD 21287, USA

Yanxun Xu YANXUN.XU@JHU.EDU
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218, USA

Editor: Elias Bareinboim

Abstract

We consider the problem of causal discovery from longitudinal observational data. We develop a novel framework that simultaneously discovers the time-lagged causality and the possibly cyclic instantaneous causality. Under common causal discovery assumptions, combined with additional instrumental information typically available in longitudinal data, we prove the proposed model is generally identifiable. To the best of our knowledge, this is the first causal identification theory for directed graphs with general cyclic patterns that achieves unique causal identifiability. Structural learning is carried out in a fully Bayesian fashion. Through extensive simulations and an application to the Women's Interagency HIV Study, we demonstrate the identifiability, utility, and superiority of the proposed model against state-of-the-art alternative methods.

Keywords: Bayesian structural learning, Causal discovery, Directed cyclic graph, Instrumental variable, Longitudinal cohort study.

1. Introduction

Causal discovery, which investigates the underlying causal relationships among a set of variables, has become increasingly important in statistics and machine learning and has found a broad range of applications in, e.g., bioinformatics (Hill et al., 2016), neuroscience (Shen et al., 2020), and atmospheric science (Runge et al., 2019a). While controlled experiments are the gold standard for establishing causality, they can be expensive, unethical, or even infeasible to implement, especially in scientific fields that involve human subjects. Therefore, many causal discovery methods aim to establish causality from observational data alone.

In this paper, we focus on causal discovery for longitudinal data, which naturally arise from many scientific disciplines. In our motivating application, people with HIV are recommended to follow up with their physicians semi-annually by current HIV guidelines (US Department of Health & Human Services, 2020) in order to collect their health information such as viral load, depressive symptoms, and kidney function longitudinally for better disease management. Among these health outcomes, some may cause others. For example, cognitive impairment may cause depression (Murata et al., 2000), and obesity may increase the risk for cardiovascular diseases (Zalesin et al., 2008). Learning causation instead of correlation among these comorbid conditions from such complex longitudinal data not only provides more accurate and robust predictions on future observations but also lays a foundation for downstream studies such as early intervention and therapeutic development.

Discovering causality from longitudinal/time-series data possesses its own advantages and challenges. On the one hand, the blessing of temporal priority (i.e., the cause always precedes its effects) breaks the symmetry in time, which can be used for orientating the causal relationship between two variables measured at different time points. On the other hand, the time gaps could be too large (say, months/years) for inferring causality that occurs at a faster rate (say, days/weeks). Although numerous methods have been developed for discovering causal structure from longitudinal/time-series data, such as the Granger causality model (Granger, 1969), the vector autoregressive model (Swanson and Granger, 1997), and the constraint-based approach (Runge et al., 2019b), the vast majority of them fail to account for instantaneous causality, which could be problematic if the gap between measuring times is large relatively to the rate of change from causal influence. In many real-world applications, detecting instantaneous causality is important. For example, the follow-up visits in our motivating HIV application are 6-months apart but the physiological and psychological changes can occur in a much shorter period of time, e.g., the improvement of people's psychosocial functioning usually reduces subsequent depressive symptoms within a month (Dunn et al., 2012).

Causal discovery is commonly formalized as a structural learning task of a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which consists of a finite set of vertices $\mathcal{V}$ representing the random variables of interest and a set of directed edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ representing the direct causal relationships (e.g., for any $X, Y \in \mathcal{V}$, $(X \to Y) \in \mathcal{E}$ indicates that X has a direct causal effect on Y) (Spirtes et al., 2000; Pearl, 2009). A fundamental challenge using this framework is to determine under what circumstances one can uniquely identify the graphical representation of the underlying causal mechanism from purely observational data. Existing works often rely on the assumption of acyclicity (Hyvärinen et al., 2010; Peters et al., 2013; Pamfil et al., 2020), i.e., there exist no feedback loops/directed cycles in the directed graph. For example,

$X \rightleftarrows Y$ is not allowed. The assumption of acyclicity significantly simplifies both theoretical and computational analyses of directed graphs due to its convenient factorization, but many real-world causal relationships are cyclic/reciprocal. For instance, there may exist a cyclic causal relationship between viral load and depression for people with HIV: a high level of viral load is a crucial risk factor for developing depressive symptoms (Jain et al., 2021), while depressed people are more likely to engage in risk-taking behaviors, which may result in rapid HIV disease progression with a higher viral load (Brickman et al., 2017).

Despite the remarkable success of causal identification theories under the assumption of acyclicity (Shimizu et al., 2006; Hoyer et al., 2008a; Peters and Bühlmann, 2014), it remains an open question whether the unique causal identifiability for observational data alone can be achieved in general directed graphs that allow for cycles. It is well-known that a directed cyclic graph is generally only identifiable up to its Markov equivalence class (Spirtes, 1995; Koster, 1996; Lanne et al., 2017; Mooij and Claassen, 2020). The literature of unique causal identifiability is quite sparse. Lacerda et al. (2008) developed a linear additive model with non-Gaussian noises and provided a sufficient condition for its unique causal identifiability by assuming that the underlying directed graph only contains disjoint cycles, which may not hold and is hard to verify in practice. Mooij et al. (2011) proposed a non-linear additive model and proved its unique identifiability under the bivariate case. Hyttinen et al. (2012) proved the identifiability of a linear cyclic model using both observational and interventional data. In summary, existing work either makes the disjoint cycle assumption, requires interventions, or is restricted to the bivariate case.

In this paper, we develop a novel framework built upon directed cyclic graphs for discovering causal relationships from longitudinal data using instrumental variables. The key idea is to use time-lagged causes as instrumental variables for instantaneous causal discoveries. While instrumental variable approaches (Angrist et al., 1996) have been extensively used as powerful tools for inferring causal effects from observational data, our use of instrumental variables is to identify the causal structure without imposing the disjoint cycle assumption. By taking advantage of the identifiability results of independent component analysis (Comon, 1994), we prove the unique causal identifiability of the proposed model with a sufficient set of instrumental variables. For structural learning, we adopt a fully Bayesian approach through spike-and-slab priors for selecting a sparse set of causes, while adjusting for both time-varying and time-invariant covariates. Through extensive simulations and an application to a large-scale HIV longitudinal cohort study, i.e., the Women's Interagency HIV Study (WIHS, Adimora et al. 2018), we demonstrate the identifiability, utility, and robustness of the proposed method, and also prove its advantages over state-of-the-art competitors. For reproducibility and broader dissemination, we make the R code that implements the proposed model publicly available at `https://github.com/bluejw/BayesDCG`.

In summary, our major contributions are two-fold. First, we propose a novel framework for longitudinal causal discovery that accounts for both time-lagged and possibly cyclic instantaneous causal relationships. Second, we establish the causal identifiability theory for directed graphs with general cyclic patterns by using the instrumental variable approach and taking advantage of the longitudinal nature of the data.

The rest of this paper proceeds as follows. In Section 2, we present the data-generating model of the proposed framework. In Section 3, we prove the unique causal identifiability of the proposed model under a general directed cyclic graph setup. In Section 4, we describe the Bayesian structural learning procedure for estimating the proposed model. In Section 5, through extensive simulation studies, we empirically verify our causal identification theory and evaluate the performance of the proposed model against state-of-the-art alternatives. In Section 6, we demonstrate the utility and superiority of the proposed model by applying it to a real-world large-scale HIV database. Lastly, we conclude with a discussion in Section 7.

2. Data-Generating Model

Let $\boldsymbol{Y}_{ij} \in \mathbb{R}^Q$ denote longitudinal health outcomes (e.g., viral load and depression score) for individual i at visit $j = 1, \ldots, J_i$, for which we aim to discover their causal relationships. Let $\boldsymbol{X}_{ij} \in \mathbb{R}^S$ denote a set of exogenous variables/covariates (e.g., age and race) for individual i at visit j, which can be either time-varying or time-invariant. The causal direction between any covariate and any health outcome is fixed to be from the former to the latter *a priori*. The causal relationships/dependencies among covariates are not of primary interest and thus will not be modeled in this work.

For notational simplicity, the individual index i will be suppressed when understood from the context. To take into account both the time-lagged and the instantaneous causalities, we propose the following data-generating model,

$$\boldsymbol{Y}_j = \boldsymbol{\mu} + \sum_{\ell=0}^{L_y} \boldsymbol{B}_\ell \boldsymbol{Y}_{j-\ell} + \sum_{\ell=0}^{L_x} \boldsymbol{A}_\ell \boldsymbol{X}_{j-\ell} + \boldsymbol{E}_j, \tag{1}$$

where L_y and L_x are the numbers of time lags for longitudinal health outcomes and covariates, respectively, $\boldsymbol{B}_\ell$ is a $Q \times Q$ matrix whose (q,p)-th element is $\beta_{\ell qp}$, $\boldsymbol{A}_\ell$ is a $Q \times S$ matrix whose (q,s)-th element is $\alpha_{\ell qs}$, $\boldsymbol{\mu} \in \mathbb{R}^Q$ is the intercept, and $\boldsymbol{E}_j \in \mathbb{R}^Q$ is the exogenous error. The direct causes of $\boldsymbol{Y}_j$ consist of both instantaneous direct causes (i.e., a subset of $\boldsymbol{Y}_j$ and $\boldsymbol{X}_j$) and time-lagged direct causes (i.e., a subset of $\boldsymbol{Y}_{j-1}, \ldots, \boldsymbol{Y}_{j-L_y}, \boldsymbol{X}_{j-1}, \ldots, \boldsymbol{X}_{j-L_x}$).

The proposed model is paired with a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which consists of a finite set of vertices $\mathcal{V} = (\cup_{j=1,q=1}^{J,\ Q} Y_{jq}) \bigcup (\cup_{j=1,s=1}^{J,\ S} X_{js})$ and a set of directed edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. Under the causal Markov assumption (Richardson, 1996), i.e., the probability distribution respects the Markov property of the causal graph, $\mathcal{G}$ has a causal interpretation. Specifically, there exists a causal relationship between two health outcomes $(Y_{jq} \leftarrow Y_{j'p}) \in \mathcal{E}$ if $0 \leq j - j' \leq L_y$ and $\beta_{j-j',qp} \neq 0$. Note that we assume $\beta_{0qq} = 0$ for any q as we do not allow instantaneous *self-loops* (i.e., $Y_{jq} \leftarrow Y_{jq}$). For health outcomes and covariates, there exists a causal relationship between a health outcome and a covariate $(Y_{jq} \leftarrow X_{j's}) \in \mathcal{E}$ if $0 \leq j - j' \leq L_x$ and $\alpha_{j-j',qs} \neq 0$. By definition, cycles are allowed for the instantaneous causality, e.g., if $\beta_{0qp} \neq 0$ and $\beta_{0pq} \neq 0$, $p \neq q$, then there is a cycle $Y_{jq} \rightleftarrows Y_{jp}$ between Y_{jq} and Y_{jp}. Figure 1 illustrates the graphical representation of the proposed model.

Due to the existence of cycles in the instantaneous causal effects $\boldsymbol{B}_0$, the right-hand side of Equation (1) does not directly specify the conditional distribution of $\boldsymbol{Y}_j$. However, we can derive this conditional distribution from the distribution of the errors $\boldsymbol{E}_j$. Specifically, consider the mapping $f : \boldsymbol{Y}_j \to \boldsymbol{E}_j$, which has the form $(\boldsymbol{I} - \boldsymbol{B}_0)\boldsymbol{Y}_j - \boldsymbol{\mu} - \sum_{\ell=1}^{L_y} \boldsymbol{B}_\ell \boldsymbol{Y}_{j-\ell} -$

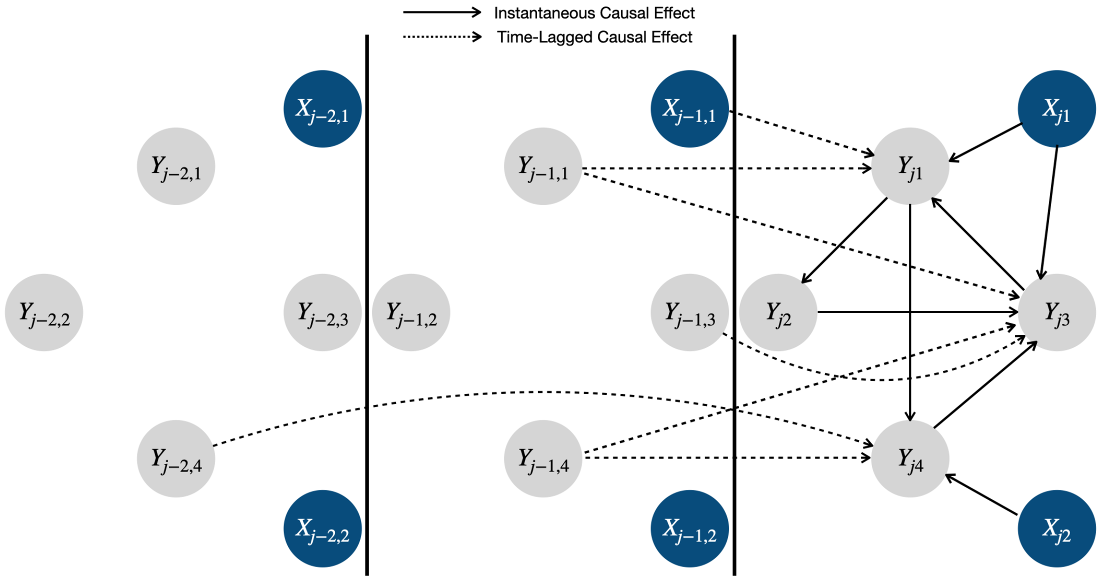


Figure 1: Graphical representation of the proposed model. The grey and blue circles represent the longitudinal health outcomes and covariates, respectively. The solid and dashed black lines indicate the instantaneous and time-lagged causal effects, respectively. Directed cycles are allowed for the instantaneous causality in the proposed model. The causal relationships among covariates are not accounted for in the proposed model. Note that due to the assumption of stationarity, this graphical representation applies to any time step j.

$\sum_{\ell=0}^{L_x} \boldsymbol{A}_\ell \boldsymbol{X}_{j-\ell} = \boldsymbol{E}_j$ induced by Equation (1). By the change-of-variable formula, we have

$$p_{\boldsymbol{Y}_j}(\boldsymbol{y}_j) = p_{\boldsymbol{E}_j}(f(\boldsymbol{y}_j))|\nabla f(\boldsymbol{y}_j)| = p_{\boldsymbol{E}_j}\left(\boldsymbol{y}_j - \boldsymbol{B}_0\boldsymbol{y}_j - \boldsymbol{\mu} - \sum_{\ell=1}^{L_y}\boldsymbol{B}_\ell\boldsymbol{y}_{j-\ell} - \sum_{\ell=0}^{L_x}\boldsymbol{A}_\ell\boldsymbol{x}_{j-\ell}\right)|\boldsymbol{I}-\boldsymbol{B}_0|, \tag{2}$$

where $|\cdot|$ denotes the absolute value of the determinant of a matrix.

For each time step j, Equation (1) can be written as $\boldsymbol{Y}_j = \boldsymbol{B}_0\boldsymbol{Y}_j + \boldsymbol{C}_j$, where $\boldsymbol{C}_j$ includes all non-cyclic components (i.e., the intercept $\boldsymbol{\mu}$, the time-lagged effects of health outcomes $\sum_{\ell=1}^{L_y} \boldsymbol{B}_\ell \boldsymbol{Y}_{j-\ell}$, the covariate effects $\sum_{\ell=0}^{L_x} \boldsymbol{A}_\ell \boldsymbol{X}_{j-\ell}$, and the exogenous error $\boldsymbol{E}_j$). To ensure that the proposed model is well-defined, we impose the following *stability* condition on $\boldsymbol{B}_0$. Specifically, we assume that the maximum modulus of $\boldsymbol{B}_0$'s eigenvalues is strictly less than 1, indicating that there exists a matrix norm $\|\cdot\|$ such that $\|\boldsymbol{B}_0\| < 1$ (Theorem 5.6.12 in Horn and Johnson (2012)). Consequently, the matrix $\boldsymbol{I} - \boldsymbol{B}_0$ is always invertible, ensuring that Equation (1) has a unique solution at each time step j, given by $\boldsymbol{Y}_j = (\boldsymbol{I} - \boldsymbol{B}_0)^{-1}\boldsymbol{C}_j$. This is because if there exists a matrix norm $\|\cdot\|$ such that $\|\boldsymbol{B}_0\| < 1$, then the series $\sum_{n=0}^{\infty} \boldsymbol{B}_0^n$ converges to some matrix. Since $(\boldsymbol{I} - \boldsymbol{B}_0)\sum_{n=0}^{N} \boldsymbol{B}_0^n = \boldsymbol{I} - \boldsymbol{B}_0^{N+1} \to \boldsymbol{I}$ as $N \to \infty$, we conclude that $(\boldsymbol{I} - \boldsymbol{B}_0)^{-1} = \sum_{n=0}^{\infty} \boldsymbol{B}_0^n$, and hence $\boldsymbol{I} - \boldsymbol{B}_0$ is invertible (Corollary 5.6.16 in Horn and Johnson (2012)).

There are two special cases of the proposed model, i.e., the (cross-sectional) structural causal model (SCM, Bollen 1989) and the vector autoregressive model (VAR, Swanson and

Granger 1997). In SCM, $L_y = L_x = 0$ in Equation (1); and in VAR, $\boldsymbol{Y}_j$ only appears on the left-hand side of Equation (1). The proposed model is advantageous over both of them. Compared to SCM, the proposed model, as we will show later, has stronger causal identification results by leveraging the longitudinal data. Compared to VAR, the proposed model can capture the instantaneous causality that occurs faster than the measuring gaps. Although there exist many causal discovery methods that also account for both time-lagged and instantaneous causal relationships (Entner and Hoyer, 2010; Kadowaki et al., 2013; Malinsky and Spirtes, 2018; Runge, 2020), they all rely on the assumption of acyclicity. While theoretically and computationally convenient, acyclic models fall short when attempting to represent the cyclic/reciprocal causal relationships, which are common in many real-world applications. In this work, we address this critical gap by offering a novel framework for longitudinal causal discovery, which not only establishes a theoretical guarantee of unique causal identifiability for directed graphs with general cyclic patterns but also provides a Bayesian structural learning algorithm that efficiently estimates the proposed model.

Here we briefly introduce the key ideas for proving the causal identifiability of the proposed model (i.e., Equation (1)), with more details to be discussed in the next section. We begin by establishing the causal identifiability theory for directed graphs with general cyclic patterns, leveraging the identifiability results of the independent component analysis (ICA, Comon 1994), and using instrumental variables. We then apply this causal identifiability theory to the causal graph $\mathcal{G}_j$ associated with the (cross-sectional) SCM, $\boldsymbol{Y}_j = \boldsymbol{B}_0\boldsymbol{Y}_j + \boldsymbol{E}_j$, defined by the instantaneous causal effects $\boldsymbol{B}_0$ that potentially involve cycles among variables in $\boldsymbol{Y}_j$ at each time step j of the proposed model. Lastly, we derive the causal identifiability of the proposed model from the causal identifiability of the instantaneous causal graph $\mathcal{G}_j$.

Specifically, we first follow the idea of Shimizu et al. (2006) and Lacerda et al. (2008) by making the following assumptions, which are common in the causal discovery literature.

Assumption 1 (Causal Sufficiency) *There are no unmeasured confounders (i.e., hidden common causes of two or more longitudinal health outcomes).*

Assumption 2 (Non-Gaussian Noise) *The jointly independent exogenous errors are all continuous-valued random variables with non-Gaussian distributions.*

Remark 1 *Causal sufficiency is assumed only for longitudinal health outcomes, while no assumptions are made regarding causal relationships/dependencies among covariates. The covariates in Equation (1) essentially serve the same role as secondary variables in a conditional directed acyclic graph (Oates et al., 2016). In other words, the proposed model is conditioned on the covariates.*

These two assumptions are critical to establishing our causal identifiability theory by leveraging the identifiability result of ICA. The identifiability theory of ICA defines an equivalence class that is much smaller and easier to handle compared to the Markov equivalence class of directed graphs with general cyclic patterns.

Next, to further determine a unique directed graph from the equivalence class defined by ICA, we use the instrumental variable approach (Angrist et al., 1996).

Definition 2 (Instrumental Variable) *For a variable Y in the directed graph $\mathcal{G}$, $I_Y \notin \mathcal{G}$ is an instrumental variable for Y if $I_Y \to Y$ is the only directed edge involving both I_Y and any variables in $\mathcal{G}$.*

Remark 3 *Note that the above definition slightly differs from the original concept of an instrumental variable. In our definition, we assume the absence of unmeasured confounders (i.e., Assumption 1), whereas the traditional instrumental variable is required to be independent of any unmeasured confounders.*

Remark 4 *Note that instrumental variables are known to be parents, but not children, of the variables in the directed graph $\mathcal{G}$, and they are not in $\mathcal{G}$. For example, at each time step j of the proposed model (1), the time-lagged variables $\mathbf{Y}_{j-\ell}$, $1 \leq \ell \leq L_y$ and the covariates $\boldsymbol{X}_{j-\ell}$, $0 \leq \ell \leq L_x$ can serve as the potential instrumental variables for $\mathbf{Y}_j \in \mathcal{G}_j$.*

Remark 5 *There are no constraints on the causal relationships/dependencies among instrumental variables. This is because, as we will see later, the causal relationships/dependencies among instrumental variables do not impact our causal identification results.*

The instrumental variable approach is a powerful tool commonly used in causal inference literature to identify causal effects under unmeasured confounding. However, its application in identifying causal structures using SCMs remains relatively limited in the literature. Both Oates et al. (2016) and Chen et al. (2023) developed methods for estimating cross-sectional SCMs using instrumental variables. Thams et al. (2024) proposed a framework for learning underlying causal structures among multivariate time series using instrumental time series. However, all these approaches were restricted to directed acyclic graphs. A recent work (Li et al., 2024) proposed to discover the causal direction in bivariate graphs using instrumental variables, allowing for both cycles and unmeasured confounders, but it remains unclear how to extend the approach to handle multivariate graphs.

In this work, we propose to use the instrumental variable approach to identify a unique directed graph within the equivalence class defined by ICA. In particular, each instrumental variable may have different children across different graphs in the equivalence class, enabling the separation of certain causal graphs from others. When a sufficient set of instrumental variables is available, unique identifiability of the causal graph can be achieved.

3. Causal Identification Theory

In this section, under common causal discovery assumptions, combined with additional instrumental information typically available in longitudinal data, we establish the first causal identification theory for directed graphs with general cyclic patterns that achieves unique causal identifiability. The main idea is to find a sufficient set of instrumental variables that guarantee the causal identifiability, by taking advantage of the longitudinal data.

We begin with a brief introduction to structural causal models, directed cyclic graphs, and the independent component analysis, which are important ingredients for establishing our causal identification theory. Then we present our causal identification results. Lastly, we prove the unique causal identifiability of the proposed model (1) by applying our causal identification theory.

3.1 Preliminaries

Let $\boldsymbol{Y} = (Y_1, \dots, Y_Q)$ denote a number of Q observed variables. A structural causal model (SCM, Bollen 1989; Pearl 2009) consists of Q structural equations,

$$Y_q = f_q(\mathrm{pa}(Y_q), e_q), \quad q = 1, \dots, Q, \tag{3}$$

where $\mathrm{pa}(Y_q) \subseteq \{Y_1, \dots, Y_Q\} \setminus Y_q$ is the set of parents (i.e., direct causes) of Y_q, $f_q(\cdot)$ is the structural causal function determining the value of the effect Y_q in terms of its direct causes $\mathrm{pa}(Y_q)$ and an exogenous error e_q. For simplicity and interpretability, we assume $f_q(\cdot)$ to be linear in this work. Then the above Q structural equations can be written in the following equivalent matrix form,

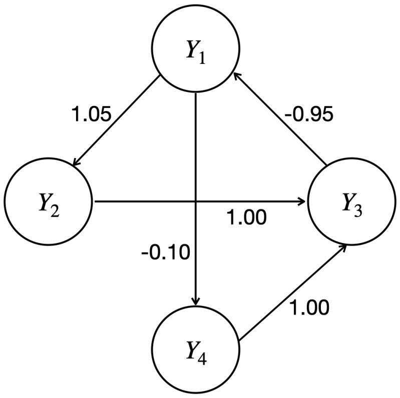
$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y} + \boldsymbol{E}, \tag{4}$$

where $\boldsymbol{B}$ denotes the $Q \times Q$ linear coefficient matrix whose (q, p)-th element is β_{qp}, and $\boldsymbol{E} = (e_1, \dots, e_Q)$ denotes the errors. If the maximum modulus of $\boldsymbol{B}$'s eigenvalues is strictly less than 1, then the SCM (4) is stable. The SCM (4) is associated with a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that represents its causal structure, where $Y_q \in \mathcal{V}$, $q = 1, \dots, Q$, and $(Y_q \leftarrow Y_p) \in \mathcal{E}$ if and only if $\beta_{qp} \neq 0$. Note that directed cycles are allowed in $\mathcal{G}$.

Definition 6 (Directed Cycle) *A (directed) cycle, denoted by $\mathcal{O}$, consists of a sequence of vertices $(Y_1, \dots, Y_M)$ along with exactly $M \geq 2$ directed edges $(Y_1 \to Y_2) \in \mathcal{E}, \dots, (Y_{M-1} \to Y_M) \in \mathcal{E}, (Y_M \to Y_1) \in \mathcal{E}$.*

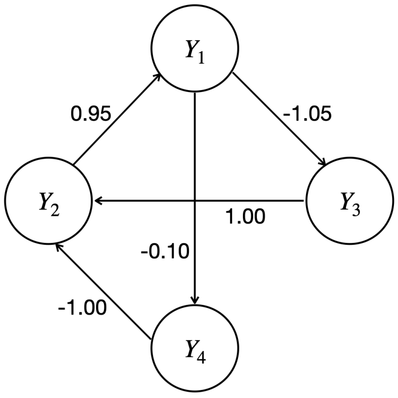
We assume that the number of vertices $M \geq 2$ as we exclude self-loops (e.g., $Y_1 \to Y_1$). For example, the sequence of vertices (Y_1, Y_2, Y_3) forms a cycle if $Y_1 \to Y_2$, $Y_2 \to Y_3$, and $Y_3 \to Y_1$; while it does not form a cycle if (i) $Y_1 \to Y_2$, $Y_2 \to Y_3$, and $Y_1 \to Y_3$, or (ii) $Y_1 \to Y_2$, $Y_2 \to Y_3$, $Y_3 \to Y_1$, and $Y_1 \to Y_3$. Definition 6 is general for defining a directed cycle, which is also consistent with the definitions introduced by Spirtes (1995) and Koster (1996). By definition, for any variable Y in a cycle $\mathcal{O}$, its parent in $\mathcal{O}$ denoted by $\mathrm{pa}_{\mathcal{O}}(Y)$ is unique. In addition, with a slight abuse of notation, we define the intersection of two cycles to be the vertices that are common to both cycles. Then two cycles $\mathcal{O}_1$ and $\mathcal{O}_2$ are *disjoint* if $\mathcal{O}_1 \cap \mathcal{O}_2 = \emptyset$, i.e., they don't share any vertices. Figure 2(a,b) illustrate two examples of directed graphs with joint cycles. Specifically, in $\mathcal{G}_1$, two cycles $Y_1 \to Y_2 \to Y_3 \to Y_1$ and $Y_1 \to Y_4 \to Y_3 \to Y_1$ intersect at $\{Y_1, Y_3\}$; in $\mathcal{G}_2$, two cycles $Y_1 \to Y_3 \to Y_2 \to Y_1$ and $Y_1 \to Y_4 \to Y_2 \to Y_1$ intersect at $\{Y_1, Y_2\}$. More examples of joint/disjoint cycles can be found in Appendix Figure S1.

An SCM associated with a directed cyclic graph is generally only identifiable up to its Markov equivalence class (Spirtes, 1995; Koster, 1996). To achieve unique causal identifiability, we adopt standard causal discovery assumptions (i.e., Assumptions 1 and 2 introduced in Section 2), which are essential for establishing our causal identification theory by leveraging the identifiability result of ICA. Specifically, the goal of the ICA is to obtain a unique *unmixing* matrix $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}$ in Equation (4) (so that $\boldsymbol{Y}$ can be uniquely expressed as a linear combination of the errors, i.e., $\boldsymbol{Y} = \boldsymbol{W}^{-1}\boldsymbol{E}$, provided $\boldsymbol{W}$ is invertible), which is equivalent to a unique graphical representation $\mathcal{G}$. Under Assumptions 1 and 2, the solution of ICA is guaranteed to be identifiable up to row-permuted row-scaled versions of $\boldsymbol{W}$, which define an equivalence class of $\mathcal{G}$. Note that if there are no self-loops in $\mathcal{G}$ (e.g., a directed edge



$$\begin{pmatrix} 0 & 0 & -0.95 & 0 \\ 1.05 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ -0.1 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0.95 & 0 \\ -1.05 & 1 & 0 & 0 \\ 0 & -1 & 1 & -1 \\ 0.1 & 0 & 0 & 1 \end{pmatrix}$$

(a) $\mathcal{G}_1$ (left), $\boldsymbol{B}_1$ (middle), $\boldsymbol{W}_1$ (right)



$$\begin{pmatrix} 0 & 0.95 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ -1.05 & 0 & 0 & 0 \\ -0.1 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & -0.95 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 1.05 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 1 \end{pmatrix}$$

(b) $\mathcal{G}_2$ (left), $\boldsymbol{B}_2$ (middle), $\boldsymbol{W}_2$ (right)

Figure 2: Two directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ (both contain joint cycles) with their corresponding linear coefficient matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, and unmixing matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. $\mathcal{G}_1$ and $\mathcal{G}_2$ are in the same ICA equivalence class. In $\mathcal{G}_2$, the values 0.95 and -1.05 are approximations of 1/1.05 and -1/0.95, respectively.

$(Y_q \leftarrow Y_q) \in \mathcal{E}$), then all the diagonal elements of $\boldsymbol{B}$ are zero (i.e., $\beta_{qq} = 0$, $q = 1, \ldots, Q$). This implies that all the diagonal elements of $\boldsymbol{W}$ are non-zero (up to a scaling factor which can be determined by normalizing the rows of $\boldsymbol{W}$ such that all diagonal elements equal to one). Therefore, to avoid self-loops in $\mathcal{G}$, we only consider row-permutations that are *admissible*, the definition of which is provided below.

Definition 7 (Row-Permutation) *A row-permutation ϕ (applied to a subset of the rows of an unmixing matrix $\boldsymbol{W}$) is a bijective mapping from R to R, where $R = \{r_1, \ldots, r_K\} \subseteq \{1, \ldots, Q\}$, $K \geq 2$, such that $\phi(r_k) \neq r_k$ for $k = 1, \ldots, K$.*

Definition 8 (Admissible Row-Permutation) *A row-permutation ϕ (applied to an unmixing matrix) is said to be admissible if all the diagonal elements of the resulting unmixing matrix are non-zero.*

In summary, the *ICA equivalence class* consists of all directed graphs $\mathcal{G}$ whose unmixing matrices $\boldsymbol{W}$'s are admissible row-permutations of each other. For example, Figure 2 displays two directed graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ with their corresponding unmixing matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$. Note that $\boldsymbol{W}_2$ can be obtained by applying the admissible row-permutation $\phi : \{1, 2, 3\} \to \{1, 2, 3\}$ such that $\phi(1) = 3$, $\phi(2) = 1$, $\phi(3) = 2$ to $\boldsymbol{W}_1$, and then normalizing each row by dividing its diagonal element. Therefore, $\mathcal{G}_1$ and $\mathcal{G}_2$ are in the same ICA equivalence class.

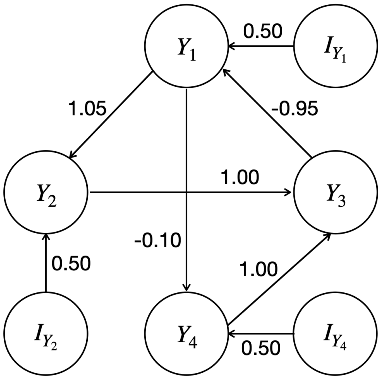
Using this theoretical framework, Shimizu et al. (2006) proved that there exists exactly one directed acyclic graph in its ICA equivalence class. Lacerda et al. (2008) further proved that there exists exactly one directed graph associated with a stable SCM among all directed graphs that only contain disjoint cycles. However, the theory for unique causal identifiability still remains open for directed graphs that possibly contain joint cycles, since it is easy to find examples such that there are multiple directed graphs with joint cycles in the same ICA equivalence class, each of which corresponds to a stable SCM. For example, $\mathcal{G}_1$ and $\mathcal{G}_2$ in Figure 2 are in the same ICA equivalence class, and both of them are associated with stable SCMs (since the modulus of all eigenvalues of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are strictly less than 1). Therefore, $\mathcal{G}_1$ and $\mathcal{G}_2$ are not identifiable from observational data, although their graphical representations and causal interpretations are quite different. To the best of our knowledge, there are no existing unique causal identification results that consider such a setup where directed graphs may contain joint cycles. In the rest of this section, we will fill the gap and all detailed proofs are provided in Appendix A.

3.2 Main Results

We now summarize our main causal identification result for SCMs under the general directed cyclic graph setup using instrumental variables in Theorem 9.

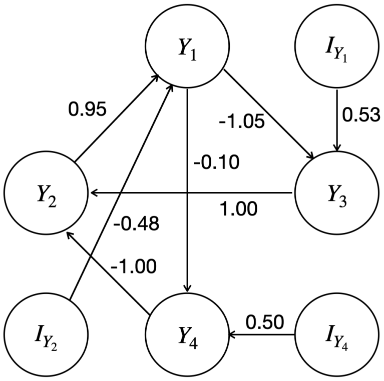
Theorem 9 (Causal Identification) *Suppose that Assumptions 1-2 hold, then a directed graph $\mathcal{G}$ with N directed cycles $\mathcal{O}_1, \ldots, \mathcal{O}_N$ (possibly joint with each other) can be uniquely identified in its ICA equivalence class if there are N (not necessarily distinct) variables $Y_1 \in \mathcal{O}_1, \ldots, Y_N \in \mathcal{O}_N$, each of which has its own instrumental variable I_{Y_n}, for $n = 1, \ldots, N$.*

Theorem 9 entails that the unique causal identifiability of any directed cyclic graph is guaranteed as long as each directed cycle within the graph consists of a variable that has its own instrumental variable. For example, $\mathcal{G}_1$ in Figure 2 consists of two directed cycles $\mathcal{O}_1 : Y_1 \to Y_2 \to Y_3 \to Y_1$ and $\mathcal{O}_2 : Y_1 \to Y_4 \to Y_3 \to Y_1$, which are joint with each other. Suppose that I_{Y_1}, I_{Y_2}, and I_{Y_4} are the instrumental variables for Y_1, Y_2, and Y_4, respectively. Let $\mathcal{G}_1^+$ denote the directed graph formed by incorporating these instrumental variables into $\mathcal{G}_1$ (i.e., Figure 3(a)). By Theorem 9, there are two ways to achieve the unique identification of $\mathcal{G}_1^+$ within its ICA equivalence class $\{\mathcal{G}_1^+, \mathcal{G}_2^+, \mathcal{G}_3^+\}$ (shown in Figure 3). First, we can uniquely identify $\mathcal{G}_1^+$ by utilizing the instrumental variable I_{Y_1} for $Y_1 \in \mathcal{O}_1 \cap \mathcal{O}_2$. This is because Y_1 is not the child of I_{Y_1} in both $\mathcal{G}_2^+$ and $\mathcal{G}_3^+$. Second, since $Y_2 \in \mathcal{O}_1$ is not the child of I_{Y_2} in $\mathcal{G}_2^+$ and $Y_4 \in \mathcal{O}_2$ is not the child of I_{Y_4} in $\mathcal{G}_3^+$, $\mathcal{G}_1^+$ can also be uniquely identified with the help of the collection of instrumental variables $\{I_{Y_2}, I_{Y_4}\}$. However, either I_{Y_2} or I_{Y_4} alone is not sufficient for the unique identification of $\mathcal{G}_1^+$. Note that the unique identification of $\mathcal{G}_1$ is then derived from the unique identification of $\mathcal{G}_1^+$.



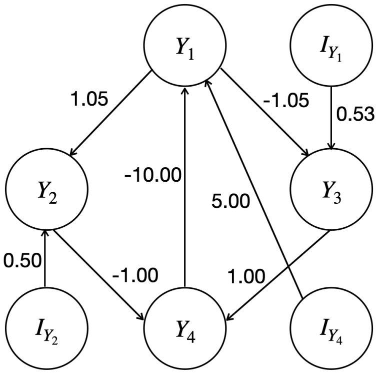
$$\begin{pmatrix} 0 & 0 & -0.95 & 0 & 0.5 & 0 & 0 \\ 1.05 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ -0.1 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0.95 & 0 & -0.5 & 0 & 0 \\ -1.05 & 1 & 0 & 0 & 0 & -0.5 & 0 \\ 0 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0 & 0 & -0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(a) $\mathcal{G}_1^+$ (left), $\boldsymbol{B}_1^+$ (middle), $\boldsymbol{W}_1^+$ (right)



$$\begin{pmatrix} 0 & 0.95 & 0 & 0 & 0 & -0.48 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ -1.05 & 0 & 0 & 0 & 0.53 & 0 & 0 \\ -0.1 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & -0.95 & 0 & 0 & 0 & 0.48 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 & 0 \\ 1.05 & 0 & 1 & 0 & -0.53 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0 & 0 & -0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(b) $\mathcal{G}_2^+$ (left), $\boldsymbol{B}_2^+$ (middle), $\boldsymbol{W}_2^+$ (right)



$$\begin{pmatrix} 0 & 0 & 0 & -10 & 0 & 0 & 5 \\ 1.05 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ -1.05 & 0 & 0 & 0 & 0.53 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 10 & 0 & 0 & -5 \\ -1.05 & 1 & 0 & 0 & 0 & -0.5 & 0 \\ 1.05 & 0 & 1 & 0 & -0.53 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(c) $\mathcal{G}_3^+$ (left), $\boldsymbol{B}_3^+$ (middle), $\boldsymbol{W}_3^+$ (right)

Figure 3: The ICA equivalence class of $\mathcal{G}_1^+$: $\{\mathcal{G}_1^+, \mathcal{G}_2^+, \mathcal{G}_3^+\}$, with their corresponding linear coefficient matrices $\boldsymbol{B}_1^+$, $\boldsymbol{B}_2^+$, and $\boldsymbol{B}_3^+$, and unmixing matrices $\boldsymbol{W}_1^+$, $\boldsymbol{W}_2^+$, and $\boldsymbol{W}_3^+$. The last three rows and columns in both the linear coefficient matrices and the unmixing matrices correspond to the instrumental variables I_{Y_1}, I_{Y_2}, and I_{Y_4}. In $\mathcal{G}_2^+$ and $\mathcal{G}_3^+$, the values 0.95, -1.05, 0.53, -0.48 are approximations of 1/1.05, -1/0.95, 0.5/0.95, -0.5/1.05, respectively.

Remark 10 *To achieve unique causal identification, as stated in Theorem 9, prior knowledge of instrumental variables is essential. This information allows us to exclude certain* *members from the ICA equivalence class, which is a critical step that goes beyond the existing causal identification result in Lacerda et al. (2008). For instance, consider the ICA equivalence class shown in Figure 3, denoted as* $\{\mathcal{G}_1^+, \mathcal{G}_2^+, \mathcal{G}_3^+\}$*. If we know that* I_{Y_1} *is an instrumental variable for* Y_1*, we can uniquely identify* $\mathcal{G}_1^+$*. In contrast, if we ascertain that* I_{Y_1} *is an instrumental variable for* Y_3*, we can exclude* $\mathcal{G}_1^+$ *from consideration, narrowing down the ICA equivalence class to* $\{\mathcal{G}_2^+, \mathcal{G}_3^+\}$*. In summary, the key point to uniquely identifying a directed graph from its ICA equivalence class lies in determining a sufficient set of instrumental variables based on prior knowledge.*

Remark 11 *Theorem 9 provides a sufficient condition for unique causal identifiability, which may require a large number of instrumental variables if there are many directed cycles. However, even with an insufficient number of instrumental variables, we can still reduce the number of equivalent graphs. For example, consider the ICA equivalence class shown in Figure 3, denoted as* $\{\mathcal{G}_1^+, \mathcal{G}_2^+, \mathcal{G}_3^+\}$*. Suppose we only know that* I_{Y_2} *is an instrumental variable for* Y_2*. Although this information is insufficient for achieving unique causal identifiability, we can exclude* $\mathcal{G}_2^+$ *from consideration, thereby narrowing the ICA equivalence class to* $\{\mathcal{G}_1^+, \mathcal{G}_3^+\}$*. Similarly, if we only know that* I_{Y_4} *is an instrumental variable for* Y_4*, we can exclude* $\mathcal{G}_3^+$ *from consideration, thereby reducing the ICA equivalence class to* $\{\mathcal{G}_1^+, \mathcal{G}_2^+\}$*.*

Remark 12 *It is possible to reduce the number of instrumental variables required for unique identification in Theorem 9 by leveraging the stability condition in certain cases. For example, since the SCM associated with* $\mathcal{G}_3^+$ *in Figure 3 is unstable, we can uniquely identify* $\mathcal{G}_1^+$ *from its ICA equivalence class by only using* Y_2*'s instrumental variable* I_{Y_2}*, which is impossible without the stability condition. In contrast, even with the stability condition, we are still not able to uniquely identify* $\mathcal{G}_1^+$ *by only using the instrumental variable* I_{Y_4} *for* Y_4*.*

We now outline our identification strategy leading to Theorem 9. First, we introduce the definition of *irreducible* row-permutations, which plays a central role in the proof. Then we summarize two key ingredients in the proof of the identification theory, both built upon irreducible row-permutations, in Lemma 14.

Definition 13 (Irreducible Row-Permutation) *A row-permutation* $\phi : R \to R$ *is said to be irreducible if there does not exist another row-permutation* $\psi : T \to T$*,* $T \subset R$*, such that* $\psi(T) = \phi(T)$*, where* $\phi(T)$ *is the image of the map* ϕ *restricted to domain* T*.*

For example, the row-permutation $\phi : \{1, 2, 3\} \to \{1, 2, 3\}$, where $\phi(1) = 3$, $\phi(2) = 1$, and $\phi(3) = 2$, is irreducible. However, the row-permutation $\phi : \{1, 2, 3, 4, 5\} \to \{1, 2, 3, 4, 5\}$, where $\phi(1) = 3$, $\phi(2) = 4$, $\phi(3) = 5$, $\phi(4) = 2$, and $\phi(5) = 1$, is not irreducible, since there exist two other row-permutations $\psi_1 : \{1, 3, 5\} \to \{1, 3, 5\}$, where $\psi_1(1) = 3$, $\psi_1(3) = 5$, and $\psi_1(5) = 1$, and $\psi_2 : \{2, 4\} \to \{2, 4\}$, where $\psi_2(2) = 4$ and $\psi_2(4) = 2$.

Lemma 14 *If an irreducible row-permutation* $\phi : R \to R$*,* $R = \{r_1, \ldots, r_K\}$*, applied to the unmixing matrix* $\boldsymbol{W}$ *associated with a directed graph* $\mathcal{G}$ *is admissible, then*

(i) there is a directed cycle $\mathcal{O}$ *in* $\mathcal{G}$ *formed by* $\{Y_{r_1}, \ldots, Y_{r_K}\}$*, and* ϕ *reverses the direction of* $\mathcal{O}$*. Without loss of generality, assume that* $\mathcal{O} : Y_{r_1} \leftarrow Y_{r_2} \leftarrow \cdots \leftarrow Y_{r_K} \leftarrow Y_{r_1}$*, then*

in the directed graph $\mathcal{G}'$ *associated with the resulting unmixing matrix* $\boldsymbol{W}'$, *we have* $\mathcal{O}' : Y_{r_1} \rightarrow Y_{r_2} \rightarrow \cdots \rightarrow Y_{r_K} \rightarrow Y_{r_1}$;

(ii) ϕ *changes the edge* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k}) \rightarrow Y_{r_k}$ *in* $\mathcal{G}$ *to the edge* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k}) \rightarrow pa_{\mathcal{O}}(Y_{r_k})$ *in* $\mathcal{G}'$, *for* $k = 1, \ldots, K$, *where* $pa_{\mathcal{O}}(Y_{r_k})$ *denotes the unique parent of* Y_{r_k} *that lies inside the directed cycle* $\mathcal{O}$, *and* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k}) = \{Y \in \mathcal{V} | (Y \rightarrow Y_{r_k}) \in \mathcal{E}\setminus\mathcal{E}_{\mathcal{O}}\}$ *where* $\mathcal{E}_{\mathcal{O}}$ *is the set of edges in* $\mathcal{O}$, *i.e.,* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k})$ *denotes all parents of* Y_{r_k} *such that the edge* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k}) \rightarrow Y_{r_k}$ *lies outside the directed cycle* $\mathcal{O}$.

Remark 15 *Note that* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k})$ *defined in Lemma 14(ii) may be empty and if not, there may be more than one such parent. For example,* $\mathcal{G}_2$ *in Figure 2 can be obtained by reversing* $\mathcal{O}_1 : Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_1$ *in* $\mathcal{G}_1$, *and then changing the edge* $Y_4 \rightarrow Y_3$ *to* $Y_4 \rightarrow Y_2$. *In this case, we have* $pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_3) = \{Y_4\}$, *whereas* $pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_1) = pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_2) = \emptyset$. *Now suppose that we add an instrumental variable* I_{Y_3} *for* Y_3 *in* $\mathcal{G}_1$, *then* $pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_3) = \{Y_4, I_{Y_3}\}$. *Additionally, the definition of* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k})$ *only requires that the edge* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k}) \rightarrow Y_{r_k}$ *lies outside the directed cycle* $\mathcal{O}$, *meaning that the vertices* $pa_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k})$ *themselves can be either inside or outside the cycle. For example, in Figure 2 we have* $pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_3) = \{Y_4\}$, *where vertex* Y_4 *lies outside directed cycle* $\mathcal{O}_1$. *Now suppose that we add an edge* $Y_1 \rightarrow Y_3$ *in* $\mathcal{G}_1$, *then we have* $pa_{\mathcal{G}_1\setminus\mathcal{O}_1}(Y_3) = \{Y_1, Y_4\}$ *since edge* $Y_1 \rightarrow Y_3$ *lies outside* $\mathcal{O}_1$, *even though vertex* Y_1 *itself lies inside* $\mathcal{O}_1$.

Lemma 14 indicates that applying an admissible irreducible row-permutation to the unmixing matrix is equivalent to performing the following two steps to the corresponding directed graph $\mathcal{G}$: (i) reversing a directed cycle $\mathcal{O} : \{Y_{r_1}, \ldots, Y_{r_K}\}$ in $\mathcal{G}$ and (ii) changing the child of $\text{pa}_{\mathcal{G}\setminus\mathcal{O}}(Y_{r_k})$ from Y_{r_k} to $\text{pa}_{\mathcal{O}}(Y_{r_k})$. Note that the proof of Theorem 4 in Lacerda et al. (2008) employed an argument that is identical to our Lemma 14(i). This argument leads to the unique identification of any directed graph that only consists of disjoint cycles under the stability condition. However, this argument alone is insufficient to uniquely identify a directed graph that possibly contains joint cycles, even when considering the stability condition. In fact, this is precisely the theoretical gap that we aim to address in this work.

Next, we will show in Proposition 16 that any admissible row-permutation can be decomposed into a collection of disjoint admissible irreducible row-permutations.

Proposition 16 *For any admissible row-permutation* $\phi : R \rightarrow R$, *there exists a collection of* $D \geq 1$ *admissible irreducible row-permutations* $\{\phi_d : R_d \rightarrow R_d\}_{d=1}^{D}$, *where* $\cup_{d=1}^{D} R_d = R$, $R_d \cap R_{d'} = \emptyset$ *for* $1 \leq d \neq d' \leq D$, *such that applying* ϕ *or sequentially applying* $\phi_1, \ldots, \phi_D$ *to any unmixing matrix yields the same row-permuted unmixing matrix.*

Therefore, by Proposition 16, we can generalize the results from Lemma 14 regarding admissible irreducible row-permutations to any admissible row-permutations. This extension leads to the characterization of the ICA equivalence class of any directed graph $\mathcal{G}$ as outlined in the following Lemma 17.

Lemma 17 (Characterization of the ICA Equivalence Class) *For any directed graph* $\mathcal{G}$ *with* N *directed cycles* $\mathcal{O}_1, \ldots, \mathcal{O}_N$ *(possibly joint with each other), all the directed graphs* $\mathcal{G}'$ *in its ICA equivalence class can be obtained by performing the following two steps:*

(i) reversing $N^\star$ of its disjoint cycles $\widetilde{\mathcal{O}}_1, \ldots, \widetilde{\mathcal{O}}_{N^\star}$, where $\widetilde{\mathcal{O}}_n \cap \widetilde{\mathcal{O}}_{n'} = \emptyset$ for $1 \leq n \neq n' \leq N^\star$, $\{\widetilde{\mathcal{O}}_1, \ldots, \widetilde{\mathcal{O}}_{N^\star}\} \subseteq \{\mathcal{O}_1, \ldots, \mathcal{O}_N\}$, and $1 \leq N^\star \leq N$;

(ii) changing the edge $\mathrm{pa}_{\mathcal{G}\setminus\widetilde{\mathcal{O}}_{n^\star}}(Y_{r_{k_{n^\star}}}) \to Y_{r_{k_{n^\star}}}$ in $\mathcal{G}$ to the edge $\mathrm{pa}_{\mathcal{G}\setminus\widetilde{\mathcal{O}}_{n^\star}}(Y_{r_{k_{n^\star}}}) \to \mathrm{pa}_{\widetilde{\mathcal{O}}_{n^\star}}(Y_{r_{k_{n^\star}}})$ in $\mathcal{G}'$, for $k_{n^\star} = 1, \ldots, K_{n^\star}$, where $\widetilde{\mathcal{O}}_{n^\star} = \{Y_{r_1}, \ldots, Y_{r_{K_{n^\star}}}\}$ and $n^\star = 1, \ldots, N^\star$.

Additionally, applying steps (i) and (ii) to any disjoint cycles in any directed graph $\mathcal{G}$ will result in a directed graph $\mathcal{G}'$ that remains in the same ICA equivalence class as $\mathcal{G}$.

For example, $\mathcal{G}_2^+$ in Figure 3 can be obtained by reversing $\mathcal{O}_1 : Y_1 \to Y_2 \to Y_3 \to Y_1$ in $\mathcal{G}_1^+$, and then changing the edges $Y_4 \to Y_3$, $I_{Y_1} \to Y_1$, and $I_{Y_2} \to Y_2$ to $Y_4 \to Y_2$, $I_{Y_1} \to Y_3$, and $I_{Y_2} \to Y_1$, respectively. Similarly, $\mathcal{G}_3^+$ in Figure 3 can be obtained by reversing $\mathcal{O}_2 : Y_1 \to Y_4 \to Y_3 \to Y_1$ in $\mathcal{G}_1^+$, and then changing the edges $Y_2 \to Y_3$, $I_{Y_1} \to Y_1$, and $I_{Y_4} \to Y_4$ to $Y_2 \to Y_4$, $I_{Y_1} \to Y_3$, and $I_{Y_4} \to Y_1$, respectively. Note that since $\mathcal{O}_1 \cap \mathcal{O}_2 = \{Y_1, Y_3\} \neq \emptyset$ (i.e., $\mathcal{O}_1$ and $\mathcal{O}_2$ are joint with each other), there are no admissible row-permutations that can reverse both $\mathcal{O}_1$ and $\mathcal{O}_2$ in $\mathcal{G}_1^+$. Consequently, $\mathcal{G}_2^+$ and $\mathcal{G}_3^+$ are the only two other directed graphs in the ICA equivalence class of $\mathcal{G}_1^+$.

We now use the instrumental variable approach to achieve the unique identification of any directed graph $\mathcal{G}$ with N directed cycles $\mathcal{O}_1, \ldots, \mathcal{O}_N$ (possibly joint with each other) in its ICA equivalence class. Suppose that for each directed cycle $\mathcal{O}_n$ in $\mathcal{G}$, $1 \leq n \leq N$, there exists a variable Y_n has its own instrumental variable I_{Y_n}. Let $\mathcal{G}^+$ denote the directed graph formed by incorporating these instrumental variables $\mathcal{I} = \{I_{Y_1}, \ldots, I_{Y_N}\}$ into $\mathcal{G}$. By the definition of the instrumental variable (i.e., Definition 2), if $\mathcal{I}$ introduces additional cycles in $\mathcal{G}^+$, then these cycles will *only* involve variables within $\mathcal{I}$ and will remain disjoint from $\mathcal{O}_1, \ldots, \mathcal{O}_N$. By Lemma 17, the reversal of disjoint cycles is a necessary condition for two directed graphs to belong to the same ICA equivalence class. Therefore, there exists a one-to-many mapping from the ICA equivalence class of $\mathcal{G}$ to the ICA equivalence class of $\mathcal{G}^+$ as they share common cycles. In particular, if $\mathcal{I}$ introduces additional cycles, each graph in the ICA equivalence class of $\mathcal{G}$ will correspond to multiple directed graphs in the ICA equivalence class of $\mathcal{G}^+$, and the latter only differ from each other in the part that *only* involves $\mathcal{I}$. On the other hand, the unique identification of $\mathcal{G}$ is independent of the part that *only* involves $\mathcal{I}$ (see details in Appendix A.1). Therefore, the unique identification of $\mathcal{G}$ can be derived from the unique identification of $\mathcal{G}^+$.

Consider any directed cyclic graph $(\mathcal{G}^+)'$ in the ICA equivalence class of $\mathcal{G}^+$ obtained by performing the two steps described in Lemma 17. Note that by the definition of instrumental variable (i.e., Definition 2), $I_{Y_{n^\star}} \in \mathcal{G}^+ \setminus \widetilde{\mathcal{O}}_{n^\star}$ is a special case of $\mathrm{pa}_{\mathcal{G}^+\setminus\widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star})$, where $Y_{n^\star} \in \widetilde{\mathcal{O}}_{n^\star}$ is the only child of $I_{Y_{n^\star}}$ in $\mathcal{G}$, for $n^\star = 1, \ldots, N^\star$. According to Lemma 17(ii), the edge $I_{Y_{n^\star}} \to Y_{n^\star}$ in $\mathcal{G}^+$ will be changed to $I_{Y_{n^\star}} \to \mathrm{pa}_{\widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star})$ in $(\mathcal{G}^+)'$. In other words, the only child of $I_{Y_{n^\star}}$ will be different in $\mathcal{G}$ and $\mathcal{G}'$ (i.e., $Y_{n^\star} \neq \mathrm{pa}_{\widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star})$ due to no self-loops), and thus $I_{Y_{n^\star}}$ will not be an instrumental variable for $Y_{n^\star}$ in $(\mathcal{G}^+)'$. Therefore, we can identify $\mathcal{G}^+$ from $(\mathcal{G}^+)'$ by utilizing a collection of instrumental variables $I_{Y_{n^\star}}$ for $Y_{n^\star} \in \widetilde{\mathcal{O}}_{n^\star}$, where $n^\star = 1, \ldots, N^\star$. We finish the proof of Theorem 9 by noting that the above argument can be applied to any $(\mathcal{G}^+)'$ within the ICA equivalence class of $\mathcal{G}^+$.

In summary, in the proof of Theorem 9, we first establish an equivalent form of applying an admissible irreducible row-permutation to an unmixing matrix and conducting two operations on the directed graph associated with the unmixing matrix in Lemma 14, then extend this result from admissible irreducible row-permutations to admissible row-permutations by illustrating their connection in Proposition 16. This leads to the characterization of the ICA equivalence class, as outlined in Lemma 17. We then establish a one-to-many mapping from the ICA equivalence class of any directed cyclic graph $\mathcal{G}$ to the ICA equivalence class of $\mathcal{G}^+$, which is the augmented graph obtained by incorporating all instrumental variables into $\mathcal{G}$. The proof is concluded by demonstrating the unique identification of $\mathcal{G}$ based on the unique identification of $\mathcal{G}^+$, utilizing a collection of instrumental variables.

Lastly, in the following corollary of Theorem 9, we prove the unique causal identifiability of the proposed model (1).

Corollary 18 *Let $\mathcal{G}_j = (\mathcal{V}_j, \mathcal{E}_j)$ denote the directed cyclic graph such that $\mathcal{V}_j = \{Y_{j1}, \dots, Y_{jQ}\}$, and $(Y_{jq} \leftarrow Y_{jp}) \in \mathcal{E}_j$ if and only if $\beta_{0qp} \neq 0$, $1 \leq p, q \leq Q$. Suppose that Assumptions 1-2 hold, then the causal structure of the proposed model* (1) *is uniquely identifiable if for any cycle $\mathcal{O}_j$ in $\mathcal{G}_j$, there exists a variable $Y_{jq} \in \mathcal{O}_j$ has its own instrumental variable $I_{Y_{jq}}$.*

There are two possible sources of the instrumental variable $I_{Y_{jq}}$ for the longitudinal health outcome Y_{jq} at each time step j in the proposed model (1). The first source is the covariate $X_{j-\ell,s}$, $\ell \geq 0$, e.g., $I_{Y_{j1}} = X_{j-1,1}$ and $I_{Y_{j4}} = X_{j2}$ as shown in Figure 1. Note that X_{j1} can not be used as an instrumental variable for Y_{j1} since it has another child Y_{j3} in $\mathcal{G}_j$. In practice, suitable covariates that satisfy the condition of instrumental variables may or may not exist. Therefore, for models that only consider instantaneous causal relationships (e.g., the SCM (4)), verifying unique causal identifiability becomes challenging when such covariates do not exist. Fortunately, we can find instrumental variables in the proposed model (1) by taking advantage of the longitudinal data. Specifically, the second source is the previous measurement of the longitudinal health outcome $Y_{j-\ell,q}$, $\ell > 0$, e.g., $I_{Y_{j3}} = Y_{j-1,3}$ and $I_{Y_{j4}} = Y_{j-2,4}$ as shown in Figure 1. This is because in longitudinal data, the causal influence of a variable Y_{jq} from its previous measurement $Y_{j-\ell,q}$, $\ell > 0$, typically exhibits a slower rate of decay compared to the causal effects originating from previous measurements of other variables $Y_{j-\ell,p}$, $\ell > 0$, $p \neq q$. As a result, it is highly plausible that an instrumental variable can be identified at some time lag $\ell > 0$ for each variable, i.e., $I_{Y_{jq}} = Y_{j-\ell,q}$.

4. Bayesian Structural Learning

We have shown that the proposed model is identifiable from observational data. Therefore, structural learning can be carried out with any appropriate estimation procedure. In this work, we adopt a Bayesian approach due to two reasons. First, it yields a computationally efficient inference procedure through posterior computation with uncertainty quantification. Second, it provides a flexible framework to incorporate prior knowledge and imposes sparsity through prior distributions for better interpretability of the inferred causal graph.

Specifically, we assign spike-and-slab priors (Ishwaran and Rao, 2005) on both $\beta_{\ell qp}$ and $\alpha_{\ell qs}$ in Equation (1) to select a sparse set of the causes. We describe the prior for $\beta_{\ell qp}$, and the prior for $\alpha_{\ell qs}$ is analogously defined. We assume that $\beta_{\ell qp} \sim \mathcal{N}(0, \gamma_{\ell qp}\nu_{\ell qp})$ with $\nu_{\ell qp} \sim$ Inverse-Gamma(a_ν, b_ν) and $\gamma_{\ell qp} \sim \rho\delta_1(\gamma_{\ell qp}) + (1-\rho)\delta_{\nu_0}(\gamma_{\ell qp})$, where $\delta_x(\cdot)$ denotes the Dirac measure at x. The hyper-parameter ν_0 is a very small, pre-specified value. In particular, if $\gamma_{\ell qp} = 1$ (slab), $\beta_{\ell qp}$ is non-zero, which suggests that there exists a causal effect of $Y_{j-\ell,p}$ on Y_{jq}; if $\gamma_{\ell qp} = \nu_0$ (spike), $\beta_{\ell qp}$ is almost negligible and can be safely treated as zero (i.e.,

$\beta_{\ell qp} \approx 0$), which in turn implies that there is no significant causal effect of $Y_{j-\ell,p}$ on Y_{jq}. We assume $\rho \sim \text{Beta}(a_\rho, b_\rho)$ following the idea of Scott and Berger (2010).

Furthermore, we assume that the non-Gaussian error e_{jq} follows a Laplace distribution (Choi and Hobert, 2013), i.e., $e_{jq} \sim \text{Laplace}(0, 2\sigma_q)$, which can be viewed as a continuous scale mixture of normal distributions. Specifically, let e_{jq} and τ_{jq} be pairs of independent random variables such that $e_{jq} \mid \tau_{jq} \sim \mathcal{N}(0, \sigma_q^2/\tau_{jq})$, and $\tau_{jq} \sim \text{Inverse-Gamma}(1, 1/8)$, then marginally $e_{jq} \sim \text{Laplace}(0, 2\sigma_q)$. In addition, we assign $\mu_q \sim \mathcal{N}(0, \sigma_\mu^2)$ and $\sigma_q^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma)$, which result in closed-form full conditionals for ease of posterior computation. We carry out posterior inference using a standard Markov chain Monte Carlo (MCMC) algorithm, the details of which are included in Appendix B.

Lastly, by Corollary 18, the unique identifiability of the proposed model (1) relies on the presence of a sufficient set of instrumental variables. Therefore, a post-hoc validation step becomes essential to ensure the learned causal graph $\mathcal{G}$ achieves unique identifiability. When such a sufficient set cannot be identified, the output of Bayesian structural learning consists of all directed graphs in the ICA equivalence class of $\mathcal{G}$ associated with stable SCMs (since we impose the stability condition on the instantaneous causal effects $\boldsymbol{B}_0$ in the MCMC algorithm, as described in Step 1.1 in Appendix B). In contrast, when a sufficient set of instrumental variables is available, the unique causal identifiability of $\mathcal{G}$ within its ICA equivalence class is guaranteed.

5. Simulation Study

In this section, we conducted a series of simulation studies to empirically verify our causal identification theory established in Section 3, and evaluate the performance of the Bayesian structural learning algorithm proposed in Section 4.

5.1 Simulation Scenario I

This scenario was designed to empirically verify our causal identification theory. The simulated true causal graph was set to be $\mathcal{G}$ in the left panel of Figure 4 with one instrumental variable $I_{Y_{j1}}$ for Y_{j1}. The right panel of Figure 4 plots the only graph $\mathcal{G}'$ in the ICA equivalence class of $\mathcal{G}$ for which the corresponding SCM is stable. We considered two types of instrumental variables: $I_{Y_{j1}} = X_{j1} \sim N(0, 1)$, and $I_{Y_{j1}} = Y_{j-1,1}$. We set the simulated true values for the intercepts to be $\mu_q = 0$, and the variances of the Laplace errors to be 1, yielding $\sigma_q^2 = 1/8$, for $q = 1, \ldots, Q$, where $Q = 4$. Assume that there were 200, 500, and $1,000$ individuals, each of which had $J_i = 5$ longitudinal observations, yielding a total sample size of $1,000$, $2,500$, and $5,000$. Then we generated the simulated true Y_{ijq} using the following data-generating process,

$$\begin{cases} Y_{ij1} = \mu_1 - 0.95 \times Y_{ij3} + 0.5 \times I_{Y_{ij1}} + e_{ij1}, \\ Y_{ij2} = \mu_2 + 1.05 \times Y_{ij1} + e_{ij2}, \\ Y_{ij3} = \mu_3 + 1 \times Y_{ij2} + 1 \times Y_{ij4} + e_{ij3}, \\ Y_{ij4} = \mu_4 - 0.1 \times Y_{ij1} + e_{ij4}. \end{cases} \tag{5}$$

The data-generating process in this study is specific to scenarios where certain conditions are met, such as the existence of multiple directed graphs with joint cycles within the same

ICA equivalence class, each corresponding to a stable SCM. We believe that this focus is necessary to fully demonstrate the strengths of the proposed model.

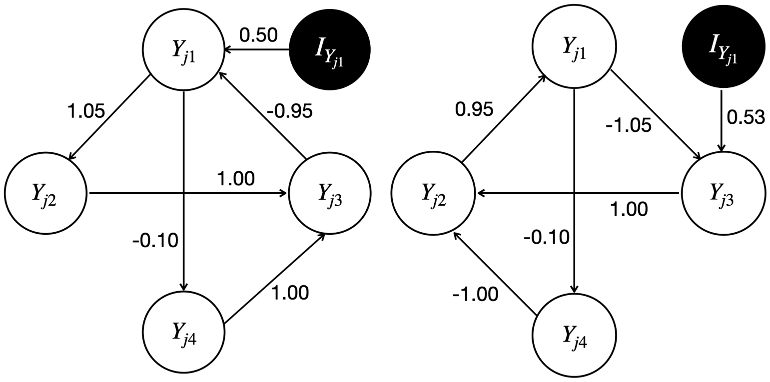


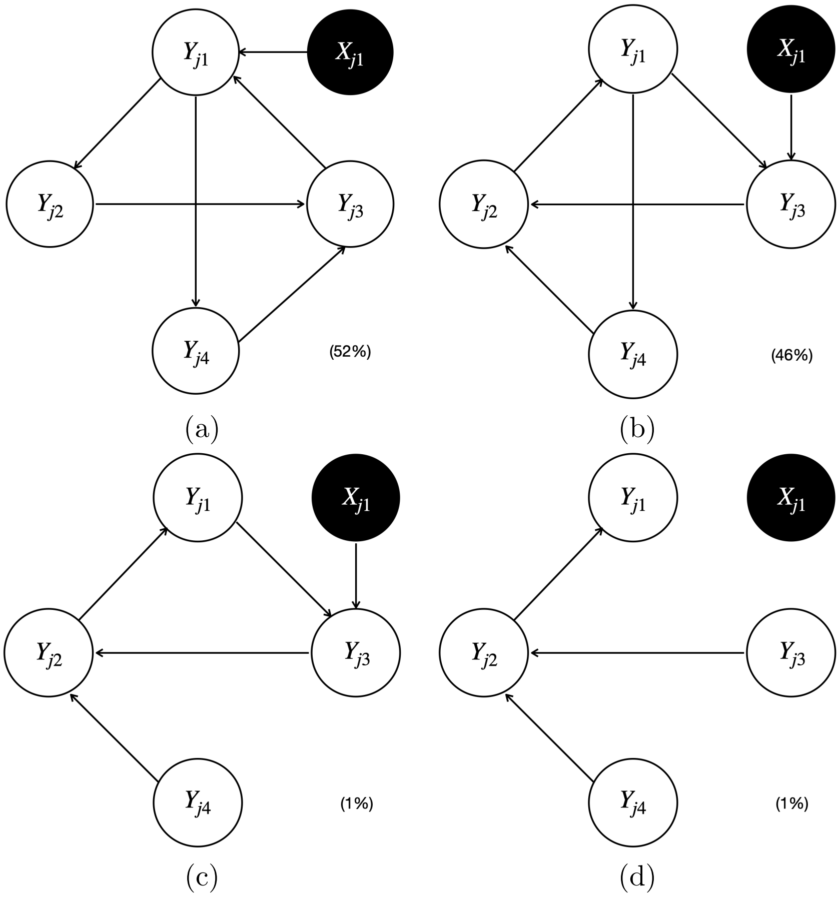
Figure 4: The simulated true causal graph $\mathcal{G}$ (left) and the only graph $\mathcal{G}'$ (right) associated with a stable SCM in its ICA equivalence class in simulation scenario I. The instrumental variable $I_{Y_{j1}}$ for Y_{j1} is highlighted by the black circle.

We applied the proposed Bayesian structural learning algorithm to the simulated datasets with the following hyper-parameter values: $\nu_0 = 2.5e\text{-}4$, $a_\nu = 5$, $b_\nu = 50$, $a_\rho = b_\rho = 0.5$, $\sigma_\mu^2 = 100$, and $a_\sigma = b_\sigma = 1$. For each configuration of the simulated dataset, we repeated the experiment with 100 replications. We ran 5,000 MCMC iterations with an initial burn-in of 2,500 iterations and a thinning factor of 5 for each analysis. For the number of time lags in the proposed model (1), we fixed $L_y = L_x = 0$ when $I_{Y_{j1}} = X_{j1}$, and set $L_y = 1$ and $L_x = 0$ when $I_{Y_{j1}} = Y_{j-1,1}$. To determine whether the estimated coefficients are zeros or non-zeros, we used the median probability model criteria (Barbieri and Berger, 2004). If $P(\gamma_{\ell qp} = 1) > 0.5$ calculated from the post-burn-in MCMC samples, we included an edge $Y_{j-\ell,p} \to Y_{jq}$. The same rule was used for determining the existence of $X_{j-\ell,s} \to Y_{jq}$.

For comparison, we considered three alternative methods. The first method is LiNG-D (Lacerda et al., 2008), which is an ICA-based causal discovery approach for cross-sectional data. While LiNG-D accounts for directed cycles, it does not incorporate covariates into its analysis. The other two methods are state-of-the-art time-series causal discovery techniques called VAR-LiNGAM (Hyvärinen et al., 2010) and PCMCI$^+$ (Runge, 2020). VAR-LiNGAM is a two-step method that first estimates time-lagged causalities using VAR (Swanson and Granger, 1997) and then estimates instantaneous causalities by applying LiNGAM algorithm (Shimizu et al., 2006) on the residuals of the first step. PCMCI$^+$ is a constraint-based approach that exploits conditional independencies to first build a skeleton of the causal graph, and then orient the skeleton according to a set of rules that define constraints on admissible orientations. Although VAR-LiNGAM and PCMCI$^+$ are capable of detecting both time-lagged and instantaneous causal relationships, they do not handle directed cycles. VAR-LiNGAM also does not consider covariates.

We implemented LiNG-D, VAR-LiNGAM, and PCMCI$^+$ using python packages py-tetrad, causal-learn, and tigramite, respectively. We set the hyper-parameters threshold_b = 0.1 and threshold_w = 0.1 as their default values for LiNG-D, and set the maximal time lag to lags = 2 for VAR-LiNGAM. For PCMCI$^+$, we used RobustParCorr to test the conditional independence for all continuous data, and RegressionCI for mixed data. In addition, we used the J-PCMCI$^+$ algorithm (Günther et al., 2023) to accommodate covariates, set the maximal time lag to $\tau_{\max} = 2$, and determined the significance level pc_α to be the optimal one in the set $\{0.01, 0.025, 0.05\}$. Since all the alternative methods do not account for individual-level information (carried by index i in the proposed model), we adopted the following data-generating process for fair comparison. For each simulated dataset used for the proposed model, we excluded its individual pattern by generating a corresponding dataset with the same sample size of $\sum_i J_i$ for each variable Y_{jq}, $j = 1, \ldots, \sum_i J_i$, $q = 1, \ldots, Q$.

We now present the simulation results for scenario I. Figure 5 and Figure 6 plot all the individual causal graphs identified by the proposed model when X_{j1} and $Y_{j-1,1}$ were used as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles) respectively, with a sample size of 5,000. The percentage within the parenthesis indicates the relative frequency of detecting the corresponding causal graph across 100 replications.



(a) (b) (c) (d)

Figure 5: All four individual causal graphs identified by the proposed model in simulation scenario I with a sample size of 5,000, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles). The percentage within the parenthesis indicates the relative frequency of detecting the corresponding causal graph across 100 replications.

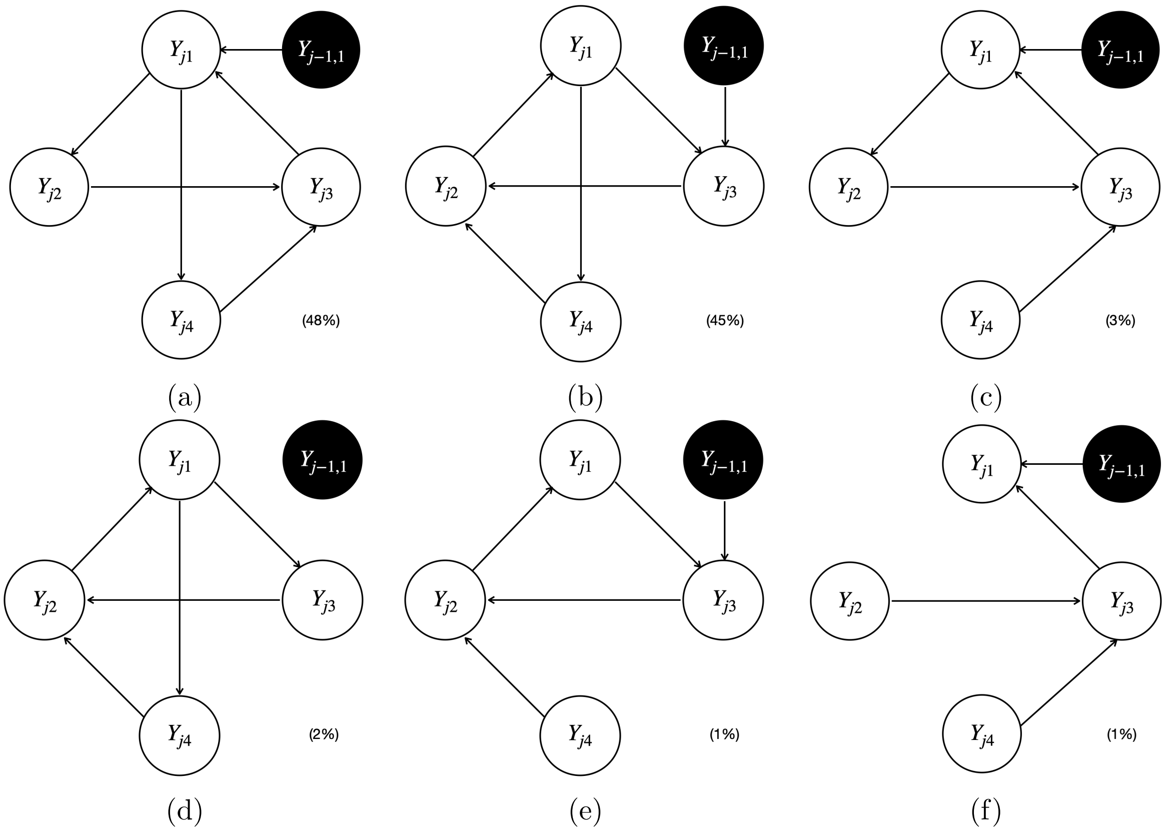
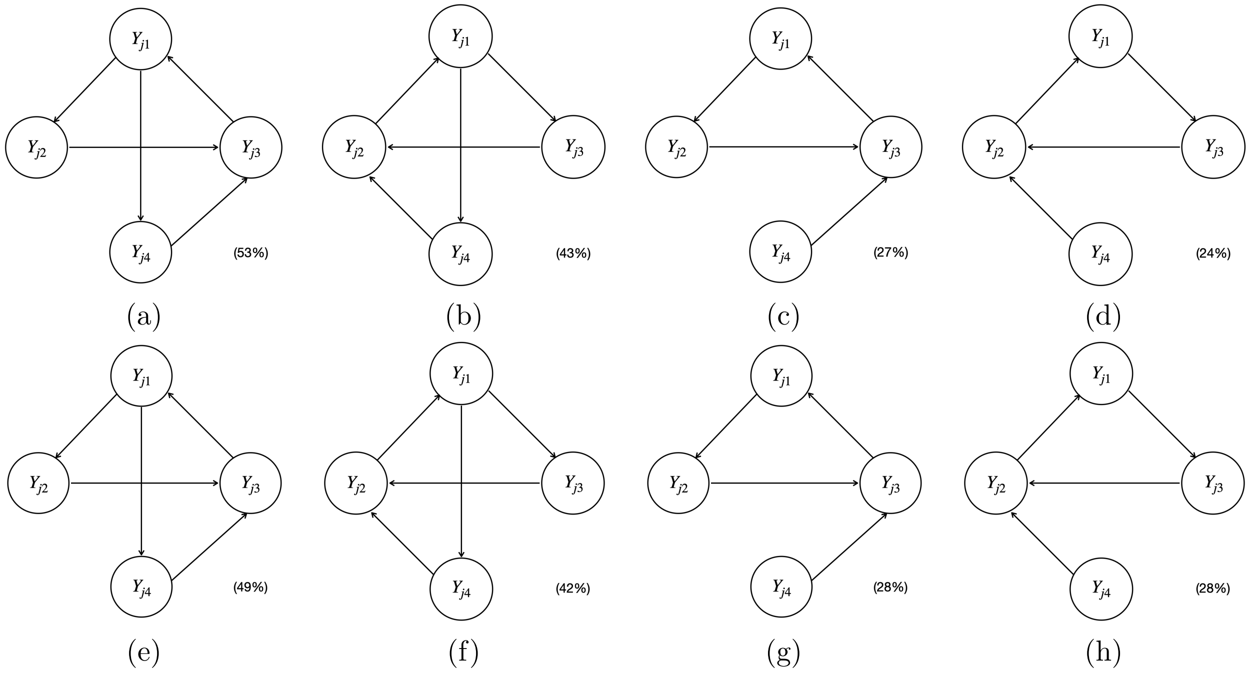


Figure 6: All six individual causal graphs identified by the proposed model in simulation scenario I with a sample size of 5,000, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles). The percentage within the parenthesis indicates the relative frequency of detecting the corresponding causal graph across 100 replications.

The estimated causal graphs under the proposed model have two distinct modes $\widehat{\mathcal{G}}^{(1)}_{\text{Proposed}}$ (i.e., Figure 5(a) and Figure 6(a)) and $\widehat{\mathcal{G}}^{(2)}_{\text{Proposed}}$ (i.e., Figure 5(b) and Figure 6(b)), where the first one corresponds to the simulated true graph $\mathcal{G}$, and the second one corresponds to the only graph $\mathcal{G}'$ associated with a stable SCM in the ICA equivalence class of $\mathcal{G}$. Note that Figure 5(c)&(d) are slight variations of Figure 5(b), Figure 6(c)&(f) are slight variations of Figure 6(a), and Figure 6(d)&(e) are slight variations of Figure 6(b). These slight variations in certain missing edges (e.g., $Y_{j1} \to Y_{j4}$ in Figure 5(c) compared to Figure 5(b)) are possibly due to finite sample size and MCMC iterations. The existence of two distinct modes $\widehat{\mathcal{G}}^{(1)}_{\text{Proposed}}$ and $\widehat{\mathcal{G}}^{(2)}_{\text{Proposed}}$ indicates the observation equivalence between $\mathcal{G}$ and $\mathcal{G}'$. Therefore, it is impossible to distinguish $\mathcal{G}$ from $\mathcal{G}'$ without the help of instrumental variables. If we can not identify a sufficient set of instrumental variables, the output of our proposed Bayesian structural learning algorithm will consist of the ICA equivalence class $\{\mathcal{G}, \mathcal{G}'\}$. However, by Corollary 18, the proposed model is uniquely identifiable, since there exists an instrumental variable $I_{Y_{j1}}$ for $Y_{j1} \in \mathcal{O}_{j1} \cap \mathcal{O}_{j2}$, where $\mathcal{O}_{j1} : Y_{j1} \to Y_{j2} \to Y_{j3} \to Y_{j1}$ and $\mathcal{O}_{j2} : Y_{j1} \to Y_{j4} \to Y_{j3} \to Y_{j1}$. Specifically, we can uniquely identify $\mathcal{G}$ within its ICA equivalence class $\{\mathcal{G}, \mathcal{G}'\}$ by noting that the child of $I_{Y_{j1}}$ is Y_{j1} in $\mathcal{G}$, whereas in $\mathcal{G}'$, it is Y_{j3}. Importantly, this unique identification holds for either $I_{Y_{j1}} = X_{j1}$ or $I_{Y_{j1}} = Y_{j-1,1}$, which empirically verified our causal identification theory.

Furthermore, Figure 7, Figure 8, and Figure 9 plot the estimated causal graphs under the alternative methods (i.e., LiNG-D, PCMCI$^+$, and VAR-LiNGAM) when X_{j1} and $Y_{j-1,1}$ were used as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles) with a sample size of 5,000. The solid red lines indicate the unoriented causal links. The solid and dashed red arrows indicate reversed and spurious causal links, respectively.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7: The estimated causal graphs under LiNG-D with a sample size of 5,000, using X_{j1} (Panel(a)-(d)) and $Y_{j-1,1}$ (Panel(e)-(h)) as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles) in simulation scenario I. The percentage within the parenthesis indicates the relative frequency of detecting the corresponding causal graph across 100 replications.

Depending on the specific estimation procedures, the equivalence classes of the estimated causal graphs reported by different methods displayed distinct patterns. Specifically, the proposed model, PCMCI$^+$, and VAR-LiNGAM output a single causal graph for each experiment, leading to the sum of the percentages across all graphs being equal to 100% (e.g., 52%+46%+1%+1%=100% for the proposed model shown in Figure 5, 83%+10%+7%=100% for PCMCI$^+$ shown in Figure 8(a)-(c), and 36%+23%+15%+14%+9%+2%+1%=100% for VAR-LiNGAM shown in Figure 9(a)-(g)). LiNG-D output either a single causal graph or two equivalent causal graphs for each experiment, resulting in a sum of percentages across both modes exceeding 100% (e.g., 53%+43%+27%+24%>100% shown in Figure 7(a)-(d)).

LiNG-D identified two distinct modes of the causal graphs $\widehat{\mathcal{G}}^{(1)}_{\text{LiNG-D}}$ (i.e., Figure 7(a)&(e)) and $\widehat{\mathcal{G}}^{(2)}_{\text{LiNG-D}}$ (i.e., Figure 7(b)&(f)) among 100 replications. Note that Figure 7(c), (d), (g), and (h) are slight variations of Figure 7(a), (b), (e), and (f), respectively, all of which missed

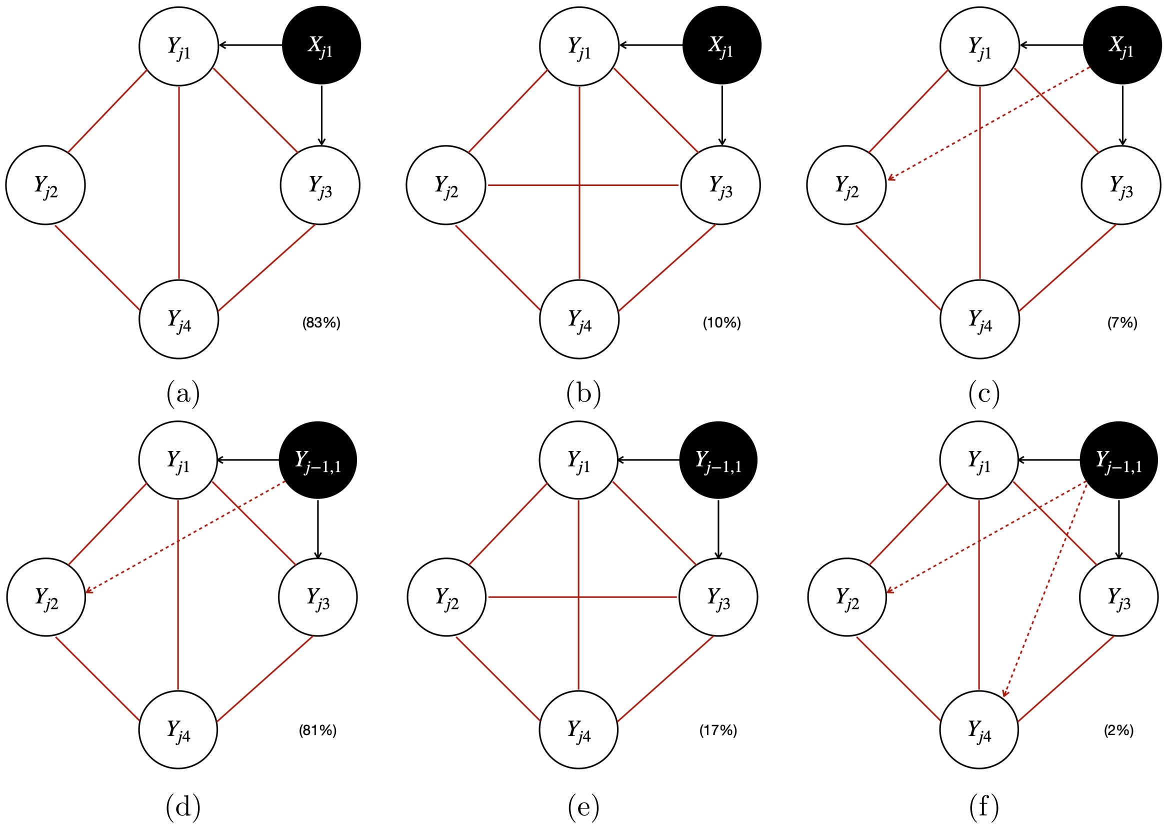
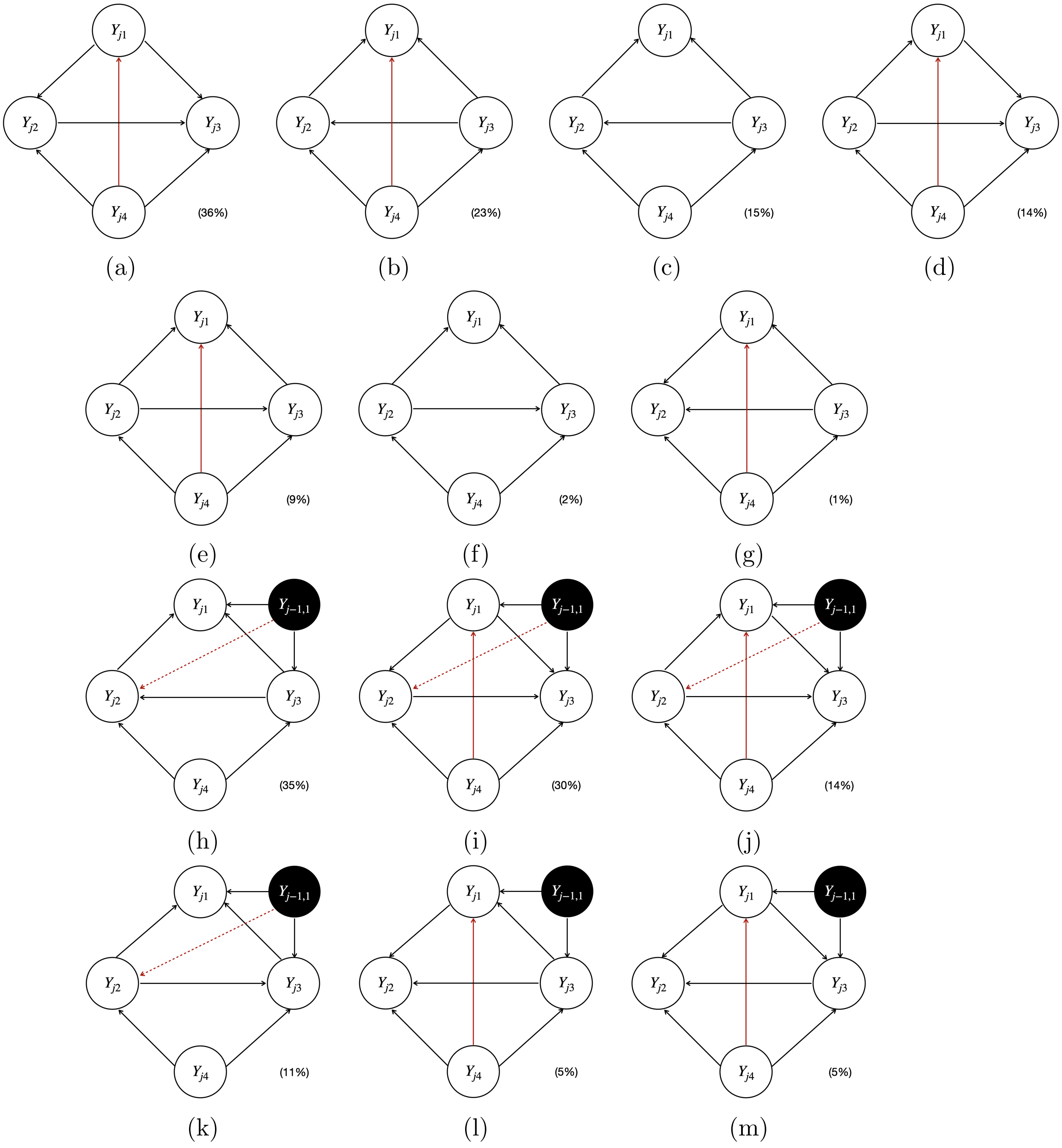


Figure 8: The estimated causal graphs under PCMCI$^+$ with a sample size of 5,000, using X_{j1} (Panel(a)-(c)) and $Y_{j-1,1}$ (Panel(d)-(f)) as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles) in scenario I. The solid red lines and dashed red arrows indicate the unoriented and spurious causal links, respectively. The percentage within the parenthesis indicates the relative frequency of detecting the corresponding graph across 100 replications.

the edge $Y_{j1} \to Y_{j4}$, possibly due to finite sample size. Despite recovering the ICA equivalence class $\{\mathcal{G}, \mathcal{G}'\}$, LiNG-D was unable to differentiate between $\mathcal{G}$ and $\mathcal{G}'$. This is because LiNG-D does not exploit covariates or time-lagged variables as instrumental variables for unique causal identification as in the proposed model. In contrast, both VAR-LiNGAM and PCMCI$^+$ did not identify the ICA equivalence class $\{\mathcal{G}, \mathcal{G}'\}$. Specifically, due to relying on the assumption of acyclicity, VAR-LiNGAM identified a Markov equivalence class comprising acyclic graphs that resemble the ICA equivalence class of cyclic graphs. For example, in order to prevent the formation of cycles, VAR-LiNGAM reversed the edge from $Y_{j1} \to Y_{j4}$ to $Y_{j1} \leftarrow Y_{j4}$. PCMCI$^+$ detected all the edges appearing in both $\mathcal{G}$ and $\mathcal{G}'$. Since PCMCI$^+$ relies on the assumption of acyclicity, all the edges among $\boldsymbol{Y}_j$'s remained unoriented. When PCMCI$^+$ detected an unoriented causal relationship between Y_{jq} and Y_{jp}, two possible scenarios arose: (i) the estimated graphs under PCMCI$^+$ with either $Y_{jq} \leftarrow Y_{jp}$ or $Y_{jq} \to Y_{jp}$ were observationally equivalent; (ii) conflicting orientations $Y_{jq} \leftarrow Y_{jp}$ and $Y_{jq} \to Y_{jp}$ were suggested by different orientation rules, possibly due to finite sample size or violations of



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m)

Figure 9: The estimated causal graphs under VAR-LiNGAM with a sample size of 5,000, using X_{j1} (Panel(a)-(g)) and $Y_{j-1,1}$ (Panel(h)-(m)) as the instrumental variable $I_{Y_{j1}}$ (highlighted by the black circles) in scenario I. The solid and dashed red arrows indicate the reversed and spurious causal links, respectively. The percentage within the parenthesis indicates the relative frequency of detecting the corresponding graph across 100 replications.

the assumption of acyclicity. VAR-LiNGAM and PCMCI$^+$ also detected a spurious causal link between $Y_{j-1,1}$ and Y_{j2} that does not appear in either $\mathcal{G}$ or $\mathcal{G}'$.

We included the simulation results under scenario I with the sample size being 2,500 and 1,000 in Appendix Figures S2-S17. We observed similar simulation results for identifying the ICA equivalence classes as those obtained when the sample size was 5,000. To compare the performance of all methods in detecting individual edges, we reported their marginal detection probability, defined as the relative frequency of detecting a specific edge, summing over all individual causal graphs across 100 replications. As the sample size increased, the estimation performance of the proposed model, LiNG-D, and PCMCI$^+$ improved. For example, when using X_{j1} as the instrument variable for Y_{j1}, the probability of detecting the edge between Y_{j1} and Y_{j4} was $23\% = 12\% + 10\% + 1\%$, $37\% = 21\% + 15\% + 1\%$, and $47\% = 45\% + 1\% + 1\%$ under the proposed model, LiNG-D, and PCMCI$^+$, respectively, with a sample size of 1,000; however, the corresponding probability was $98\% = 52\% + 46\%$, $96\% = 53\% + 43\%$, and $100\% = 83\% + 10\% + 7\%$ when sample size was 5,000.

The proposed model did not uniformly outperform alternative methods. For example, when using $Y_{j-1,1}$ as the instrumental variable for Y_{j1} with a sample size of 1,000, 2,500, and 5,000, the probability of detecting the edge between Y_{j1} and Y_{j4} under the proposed model was $27\% = 15\% + 11\% + 1\%$, $72\% = 41\% + 30\% + 1\%$, and $95\% = 48\% + 45\% + 2\%$, respectively, whereas PCMCI$^+$ achieved a true positive rate of $52\% = 36\% + 7\% + 5\% + 2\% + 2\%$, $91\% = 78\% + 13\%$, and $100\% = 81\% + 17\% + 2\%$, respectively. When compared to LiNG-D, the proposed model demonstrated a higher false positive rate when using $Y_{j-1,1}$ as the instrumental variable with a sample size of 1,000 (see Appendix Figures S14-S15).

Lastly, to further showcase the ability of the proposed Bayesian learning algorithm to robustly and accurately identify ICA equivalence classes, we conducted additional simulation studies using a different simulated true graph from the one discussed in this Section. Detailed simulation results, highlighting the superior performance of our method compared to alternative methods (i.e., LiNG-D, VAR-LiNGAM, and PCMCI$^+$), are provided in Appendix C.1. To empirically verify the causal identification theory under the proposed model with a varying number of visits J_i, we conducted additional simulation studies. Detailed descriptions of the setups and results, demonstrating that our causal identification theory applies effectively to general cases with varying J_i, are provided in Appendix C.2.

5.2 Simulation Scenario II

In this scenario, we evaluated the performance of the proposed model by generating synthetic datasets that mimic our motivating HIV dataset. Assume that there were 200 individuals with $Q = 3$ longitudinal health outcomes and $S = 3$ covariates including one binary covariate and two continuous covariates, i.e., $\boldsymbol{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})$, where X_{ij1}'s were generated from Bernoulli(0.6), and X_{ij2}, X_{ij3}'s were generated from independent standard normal distributions, $i = 1, \ldots, 200$. The number of longitudinal observations J_i for each individual i was randomly sampled from our motivating HIV dataset without replacement, resulting in the number ranging from 3 to 46, and a total sample size of $3,684$. We set the simulated true values for the intercepts to be $\mu_1 = 1$, $\mu_2 = -1$, $\mu_3 = 0$, and the variances of the Laplace errors to be 1, yielding $\sigma_q^2 = 1/8$ for $q = 1, \ldots, Q$. Then we generated the simulated true Y_{ijq} using the following data-generating process,

$$\begin{cases} Y_{ij1} = \mu_1 + 0.5\eta \times Y_{ij2} + 0.5\eta \times Y_{i,j-1,1} + 0.25\eta \times Y_{i,j-1,2} + 0.75\eta \times X_{ij1} + e_{ij1}, \\ Y_{ij2} = \mu_2 + 0.25\eta \times Y_{ij3} + 0.5\eta \times Y_{i,j-1,2} + 0.125\eta \times Y_{i,j-1,3} - 0.5\eta \times X_{ij2} + e_{ij2}, \\ Y_{ij3} = \mu_3 + 0.1\eta \times Y_{ij1} + 0.5\eta \times Y_{i,j-1,3} + 0.25\eta \times X_{ij3} + e_{ij3}, \end{cases} \tag{6}$$

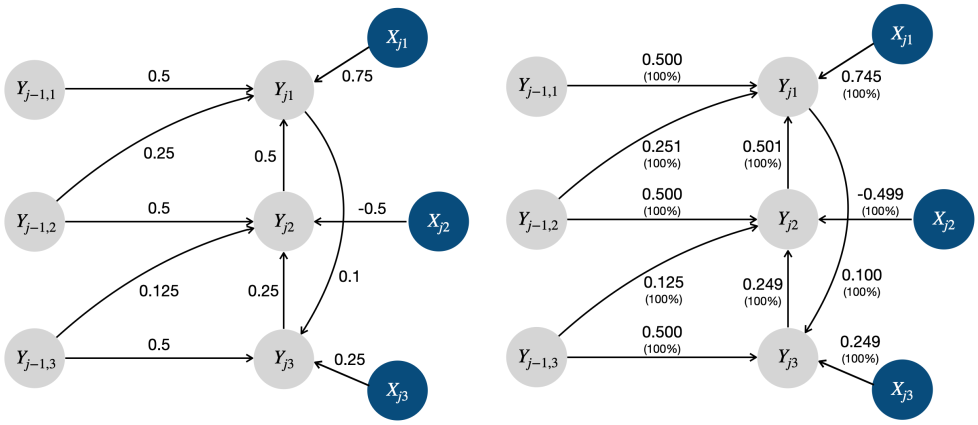
where $\eta \in \{0.5, 0.75, 1\}$ controls the causal effect size. The left panels of Figure 10(a,b,c) plot the simulated true causal graph $\mathcal{G}$ with $\eta = 1$, 0.75, and 0.5. The true number of time lags are $L_y^{\text{truth}} = 1$ and $L_x^{\text{truth}} = 0$ for longitudinal health outcomes and covariates, respectively. Note that X_{j1} and $Y_{j-1,1}$ are instrumental variables for Y_{j1}, and X_{j2} and X_{j3} are instrumental variables for Y_{j2} and Y_{j3}, respectively.

We applied the proposed Bayesian structural learning algorithm to the simulated datasets with the following hyper-parameter values: $\nu_0 = 5e\text{-}5 \times \eta$ for $\eta \in \{0.5, 0.75, 1\}$, $a_\nu = 5$, $b_\nu = 50$, $a_\rho = b_\rho = 0.5$, $\sigma_\mu^2 = 100$, and $a_\sigma = b_\sigma = 1$. Considering the simulated true graph $\mathcal{G}$ in this scenario with a substantial number of time-lagged and instantaneous causal relationships, we determined the number of time lags L_y and L_x in the proposed model from the data following the idea from Pamfil et al. (2020). They observed a significant decrease in the largest absolute value of the estimated coefficients $\boldsymbol{B}_\ell$ toward zero as ℓ increased from L_y^{truth} to $L_y^{\text{truth}} + 1$. This phenomenon provided a basis for determining the number of time lags L_y. The same rule was applied to select L_x based on $\boldsymbol{A}_\ell$. We performed 100 repeated experiments for each configuration of the simulated dataset with the same MCMC setup as in scenario I. To determine the presence of a directed edge in the estimated causal graph, we used the same median probability model criteria as in scenario I.

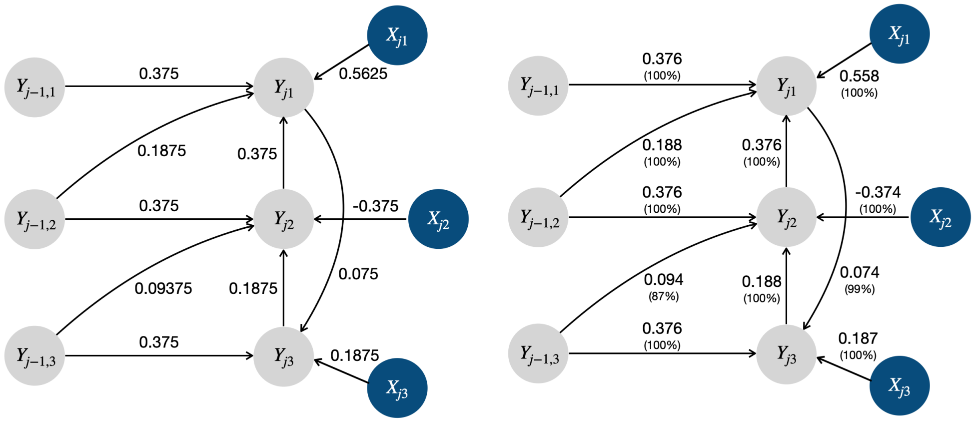
For comparison, we again applied PCMCI$^+$, LiNG-D, and VAR-LiNGAM using the same setup as in scenario I. In addition, to illustrate the advantages of the proposed model in capturing both time-lagged and instantaneous causal relationships, we considered two special cases of the proposed model and the learning algorithm for comparison. The first model is the (cross-sectional) structural causal model (SCM, Bollen 1989), which only estimates the instantaneous causal effects by setting the time-lagged causal effects to zero in the proposed model. The second model is the vector autoregressive model (VAR, Swanson and Granger 1997), which solely estimates the time-lagged causal effects by fixing the instantaneous causal effects at zero in the proposed model.

We now report the simulation results for scenario II. As shown in Appendix Figure S18, the proposed model successfully identified $L_y = 1$ and $L_x = 0$. Figure 10(a,b,c) summarize the estimated causal graphs under the proposed model when $\eta = 1$, 0.75, and 0.5, including the estimated causal effects averaged over 100 replications and the probabilities for detecting the corresponding edges. We found that the proposed model not only perfectly recovered the simulated true causal graph $\mathcal{G}$, but also estimated both the time-lagged and instantaneous causal effects with high accuracy.

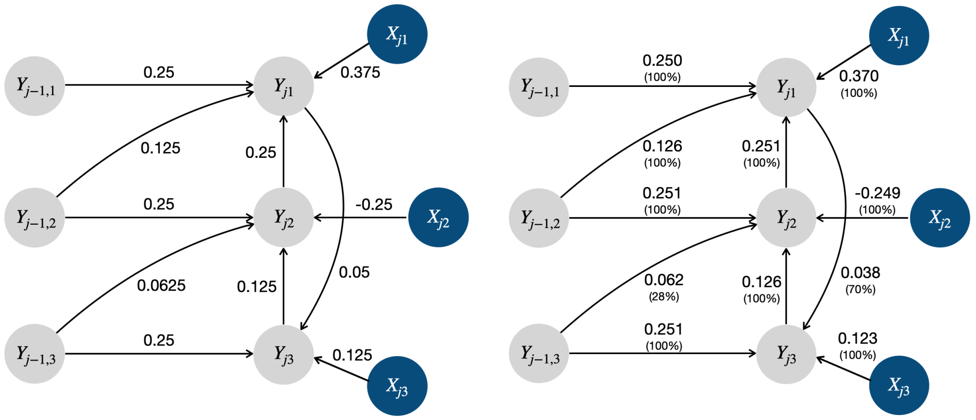
Moreover, Figure 11 displays the estimated causal graphs under the alternative methods PCMCI$^+$, LiNG-D, VAR-LiNGAM, SCM, and VAR when $\eta = 1$. Due to relying on the assumption of acyclicity, PCMCI$^+$ detected all directed edges in the instantaneous directed cycle $Y_{j1} \leftarrow Y_{j2} \leftarrow Y_{j3} \leftarrow Y_{j1}$ without orientation, and VAR-LiNGAM reversed the direction of the edge $Y_{j1} \rightarrow Y_{j3}$ to avoid the formation of cycles. While LiNG-D successfully identified the instantaneous cycle, it did not capture any time-lagged causal relationships. In addition,



(a) $\eta = 1$: $\mathcal{G}$ (left) and $\widehat{\mathcal{G}}_{\text{Proposed}}$ (right)

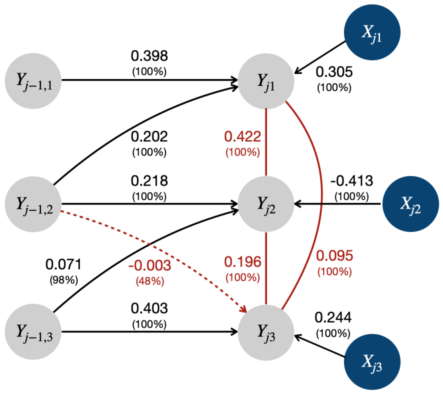


(b) $\eta = 0.75$: $\mathcal{G}$ (left) and $\widehat{\mathcal{G}}_{\text{Proposed}}$ (right)

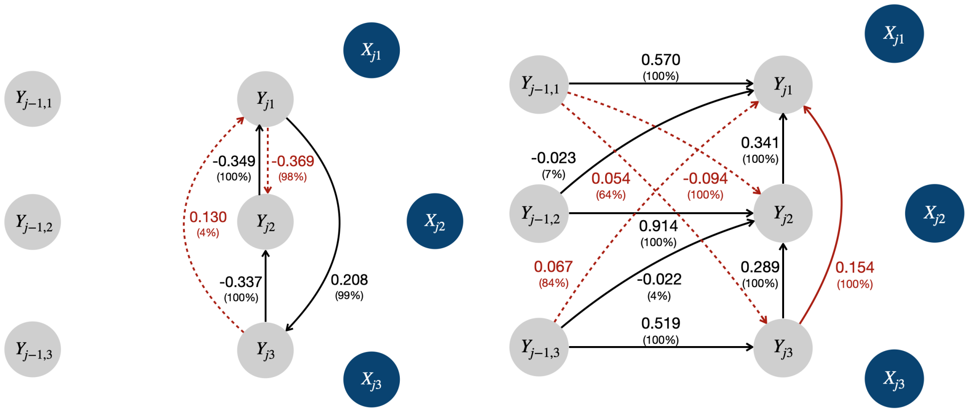


(c) $\eta = 0.5$: $\mathcal{G}$ (left) and $\widehat{\mathcal{G}}_{\text{Proposed}}$ (right)

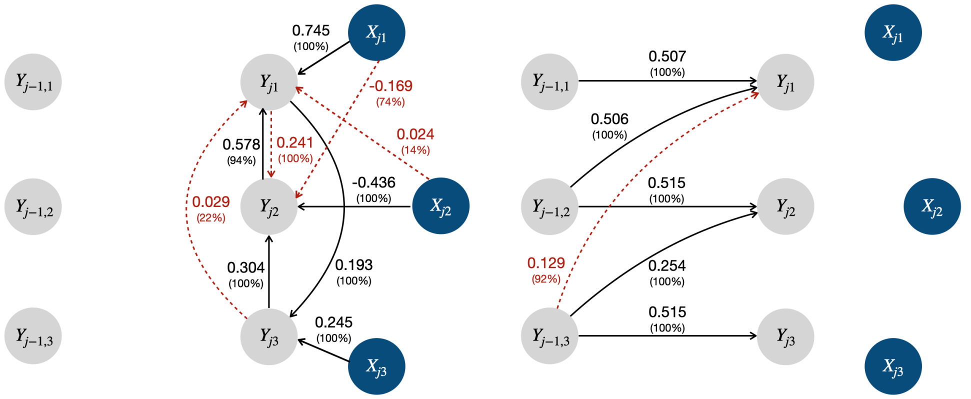
Figure 10: Simulated true causal graphs, and the estimated causal graphs under the proposed model in scenario II with $\eta = \{0.5, 0.75, 1\}$. The grey and blue circles represent the longitudinal health outcomes and covariates, respectively.



(a) $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$



(b) $\widehat{\mathcal{G}}_{\text{LiNG-D}}$ (c) $\widehat{\mathcal{G}}_{\text{VAR-LiNGAM}}$



(d) $\widehat{\mathcal{G}}_{\text{SCM}}$ (e) $\widehat{\mathcal{G}}_{\text{VAR}}$

Figure 11: The estimated causal graphs under alternative methods (i.e., PCMCI$^+$, LiNG-D, VAR-LiNGAM, SCM, and VAR) in scenario II with $\eta = 1$. The grey and blue circles represent the longitudinal health outcomes and covariates, respectively. The solid red lines indicate the unoriented causal links. The solid and dashed red arrows indicate reversed and spurious causal links, respectively.

all the alternative methods detected spurious causal links. Therefore, the proposed model outperformed all the alternatives in terms of recovering the underlying causal structure.

Lastly, we provide simulation results for alternative methods under scenario II with the causal effect size being $\eta = 0.75$ and 0.5 in Appendix Figure S19 and S20, respectively. When the effect size was small, all methods exhibited a lower true positive rate for detecting edges with small causal effects. For example, the probability of detecting the edge between Y_{j1} and Y_{j3} was 70%, 81%, 3%, 11%, and 65% under $\widehat{\mathcal{G}}_{\text{Proposed}}$, $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$, $\widehat{\mathcal{G}}_{\text{LiNG-D}}$, $\widehat{\mathcal{G}}_{\text{VAR-LiNGAM}}$ and $\widehat{\mathcal{G}}_{\text{SCM}}$ with $\eta = 0.5$, respectively; while the probability of detecting the edge between Y_{j1} and Y_{j3} was almost 100% for all the methods with $\eta = 1$. Note that when the effect size was small, PCMCI$^+$ had a slightly higher true positive rate than the proposed model. In addition, all the alternative methods exhibited a lower false positive rate with a small effect size. For instance, the probability of detecting $Y_{j-1,2} \to Y_{j3}$ was 48%, 1%, and 1% by $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$ when $\eta = 1$, 0.75 and 0.5, respectively; the probability of detecting $Y_{j1} \to Y_{j2}$ was 98%, 10%, and 0% by $\widehat{\mathcal{G}}_{\text{LiNG-D}}$ when $\eta = 1$, 0.75 and 0.5, respectively; the probability of detecting $Y_{j-1,3} \to Y_{j1}$ was 84%, 14%, and 0% by $\widehat{\mathcal{G}}_{\text{VAR-LiNGAM}}$ when $\eta = 1$, 0.75 and 0.5, respectively; the probability of detecting $Y_{j1} \to Y_{j2}$ was 100%, 70%, and 4% by $\widehat{\mathcal{G}}_{\text{SCM}}$ when $\eta = 1$, 0.75 and 0.5, respectively; the relative frequency of detecting $Y_{j-1,3} \to Y_{j1}$ was 92%, 12%, and 0% by $\widehat{\mathcal{G}}_{\text{VAR}}$ when $\eta = 1$, 0.75 and 0.5, respectively.

6. Application: WIHS Data Analysis

The Women's Interagency HIV Study (WIHS, Adimora et al. 2018) is a large prospective, observational, multicenter study designed to investigate the impact of HIV infection on multimorbidity in women with HIV or at risk for HIV in the United States. Semi-annually, participants follow up with their physicians for data collection, including assessments of sociodemographic, clinical, and behavioral characteristics. For the present analysis, we included all women from the Washington, D.C. site with at least two visits, yielding a total of 298 individuals. We were interested in investigating the causal relationships among the following $Q = 8$ longitudinal health outcomes that HIV physicians commonly account for when making treatment decisions in clinical practice: depression scores evaluated through the Center for Epidemiological Studies Depression Scale (Radloff, 1977) spanning somatic symptoms (e.g., sleep and appetite difficulties), negative affect (e.g., loneliness and sadness), lack of positive affect (e.g., hopelessness), and interpersonal symptoms (e.g., people are unfriendly); viral load of HIV RNA, CD4 count, estimated glomerular filtration rate (eGFR; a kidney function indicator), and body mass index (BMI). To verify the assumption of non-Gaussian noise, we conducted the Shapiro–Wilk test (Shaphiro and Wilk, 1965), where all the health outcomes rejected the null hypothesis. We also extracted risk factors including age, race, diabetes, smoking status, marital status, and education level as covariates.

We applied the proposed model to the WIHS dataset. We slightly modified the spike-and-slab prior on $\beta_{\ell qp}$ by assuming that (1) $\beta_{\ell qp} = 0, \ell > 0$ if $\beta_{0qp} = 0$; and (2) $\text{sign}(\beta_{\ell qp})$ is the same for all $\ell \geq 0$ for incorporating prior clinical knowledge and better interpretation. The same modification was applied to $\alpha_{\ell qs}$. The details of the prior specification are described in Appendix D. In addition, we selected the number of time lags using the same criteria as in the simulation study. As shown in Appendix Figure S21, $L_y = 1$ and $L_x = 0$.

We ran 25,000 MCMC iterations after an initial burn-in of 25,000 iterations, and a thinning factor of 50. We also applied PCMCI$^+$ to the same WIHS dataset for comparison.

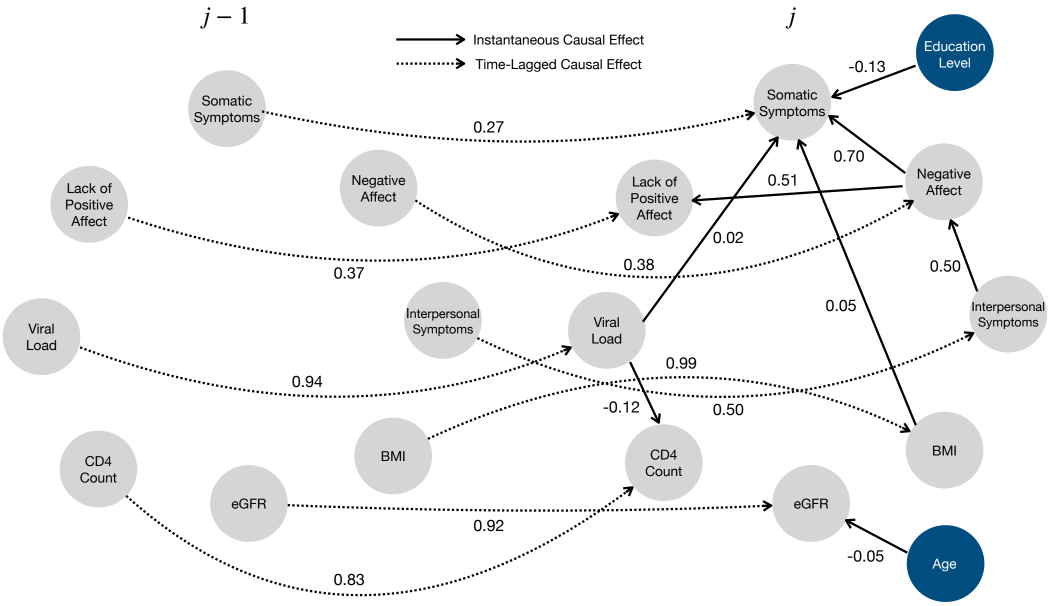
Figure 12(a,b) summarize the estimated causal graphs under the proposed model and the alternative method PCMCI$^+$ in the WIHS data analysis. The grey and blue circles represent the longitudinal health outcomes (i.e., Y_{jq}) and the covariates (i.e., X_{js}), respectively. The solid and dashed black lines indicate the instantaneous (i.e., $Y_{jq} \leftarrow Y_{jp}$ or $Y_{jq} \leftarrow X_{js}$) and time-lagged (i.e., $Y_{jq} \leftarrow Y_{j-1,p}$) causal effects, respectively. The solid red lines indicate the unoriented causal effects detected by PCMCI$^+$. To achieve the unique causal identifiability of the proposed model, we utilize the preceding measurement $Y_{j-1,q}$ of each longitudinal health outcome Y_{jq} as an instrumental variable. We will discuss the validity of this approach and highlight the advantages of using preceding measurements as instrumental variables compared to covariates such as age and educational level later.

We first report on the estimation of the instantaneous causal effects among longitudinal health outcomes. As shown in Figure 12(a), most of the instantaneous causal relationships revealed by the proposed model were associated with four depression items (i.e., somatic symptoms, negative affect, lack of positive affect, and interpersonal symptoms), which are important clinical measurements reflecting physical well-being and overall quality of life for people with HIV. For example, negative affect was identified as a direct cause for both somatic symptoms and lack of positive affect. Diener and Emmons (1984) examined both the instantaneous and the time-lagged causal effects between negative affect and lack of positive affect, and found that they were positively correlated in the short time period, but were independent in the long term. Charles and Almeida (2006) reported that people's prior-day negative affect influenced their current-day somatic symptoms in the positive direction. In addition, non-depression health outcomes such as viral load and BMI were also estimated to be direct causes for somatic symptoms. A high level of viral load and obesity/over-weight are both crucial risk factors for somatic symptoms in people with HIV (Roberts et al., 2000; Jain et al., 2021). Furthermore, instantaneous causal relationships were also detected among non-depression health outcomes. For instance, the proposed model discovered a negative causal effect from viral load to CD4 count, which is well-known in the HIV literature since CD4 cell is the primary target of HIV in human body (Vidya Vijayan et al., 2017).

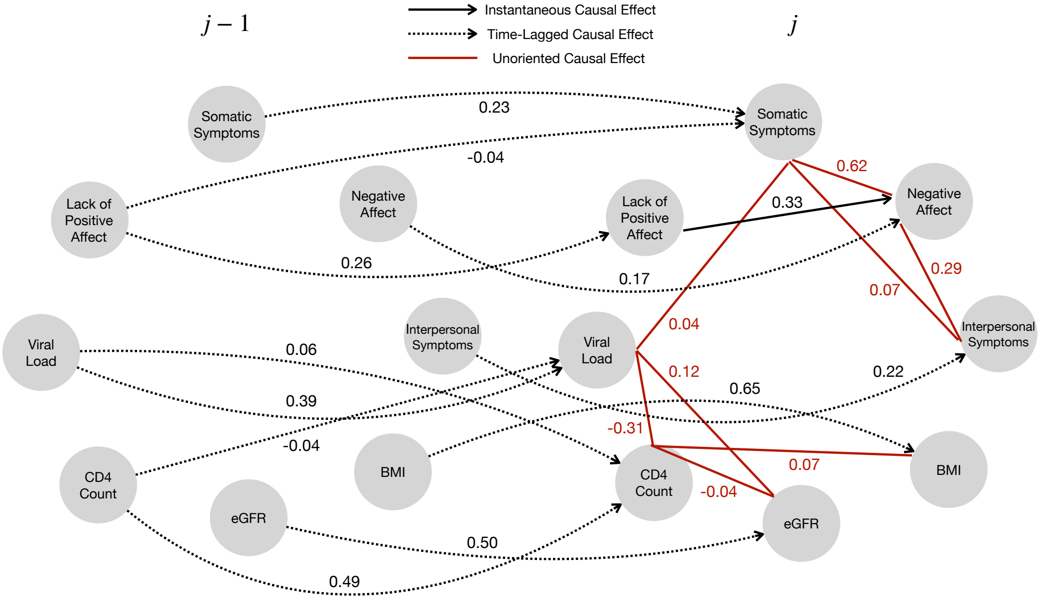
As shown in Figure 12(b), $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$ shares similar instantaneous causal graph skeleton among health outcomes with $\widehat{\mathcal{G}}_{\text{Proposed}}$. However, compared to $\widehat{\mathcal{G}}_{\text{Proposed}}$, most of the instantaneous causal relationships detected by PCMCI$^+$ were unoriented. For example, the causal relationship between viral load and CD4 count was unoriented in $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$, which should be from the former to the latter according to known clinical knowledge.

Next, we summarize the estimated instantaneous effects between covariates and health outcomes. As shown in Figure 12(a), we found that a lower educational level was estimated to increase the risk for somatic symptoms, which is consistent with existing knowledge (Sayar et al., 2003). Moreover, with aging, people exhibit progressive decreases in eGFR, which eventually leads to the loss of kidney function (Weinstein and Anderson, 2010).

We then present the estimated time-lagged causal effects among different longitudinal health outcomes. For each health outcome Y_{jq}, the only significant time-lagged causal effect was estimated to be from its own preceding measurement (i.e, $Y_{jq} \leftarrow Y_{j-1,q}$ for all q) by the proposed model (shown in Figure 12(a)). This implies that, for each q, $Y_{j-1,q}$ can be used as an instrumental variable for Y_{jq}, which guarantees the instantaneous causal identifiability



(a) $\widehat{\mathcal{G}}_{\text{Proposed}}$



(b) $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$

Figure 12: The estimated causal graphs under the proposed model (i.e., $\widehat{\mathcal{G}}_{\text{Proposed}}$) and the alternative method PCMCI$^+$ (i.e., $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$) in the WIHS data analysis. The grey and blue circles represent the longitudinal health outcomes and covariates, respectively. The solid and dashed black lines indicate the instantaneous and time-lagged causal effects, respectively. The solid red lines indicate the unoriented causal effects.

according to Corollary 18. However, determining the suitability of age and education level as instrumental variables for eGFR and somatic symptoms, respectively, poses challenges. Specifically, aging is recognized as a risk factor for kidney function loss but may also impact depressive symptoms in individuals with HIV (Jin et al., 2022). Similarly, education level may be a risk factor for non-somatic depressive symptoms (Lee, 2011). This underscores the robust causal identification capabilities of the proposed model compared to general SCMs.

Furthermore, all of these time-lagged causal relationships in $\widehat{\mathcal{G}}_{\text{Proposed}}$ were also captured by PCMCI$^+$ (shown in Figure 12(b)). However, there exist a few additional time-lagged causal relationships between different health outcomes in $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$. For example, viral load at the last visit (i.e., visit $j-1$) was estimated to positively influence CD4 count at the current visit (i.e., visit j), which was possibly a spurious causal link due to the following three reasons: (i) it is well-known in the HIV literature that a higher level of viral load decreases the CD4 count in both short-term and long-term (Vidya Vijayan et al., 2017); (ii) viral load was estimated to negatively influence CD4 count in the instantaneous causal graph of $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$; (iii) the estimated time-lagged causal effect between viral load and CD4 count was negligible (i.e., 0.06), compared to both the instantaneous causal effect between viral load and CD4 count (i.e., -0.31), as well as the time-lagged causal effect between viral load at different visits (i.e., 0.39) in $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$.

Lastly, we did not detect any cycles in the WIHS data analysis. There are two potential explanations: (i) the underlying true causal relationships among these health outcomes for people with HIV may be best represented by a directed acyclic graph, in which case the proposed method correctly identified the causal structure; or (ii) there may be cycles among these health outcomes, but due to data variability and limited sample size, our method may not be able to detect certain edges with small effect sizes. For example, a hypothetical cyclic causal relationship could exist between viral load and somatic symptoms. However, the effect of viral load on somatic symptoms is very small (i.e., 0.02 as shown in Figure 12(a)). The potential effect of somatic symptoms on viral load may be even smaller, making it undetectable by the proposed method given the limited sample size.

7. Conclusion

We developed a novel framework for simultaneously discovering the time-lagged and possibly cyclic instantaneous causalities from longitudinal/time-series observational data. To achieve the unique causal identifiability of the proposed model, we required an instrumental condition for directed graphs with joint cycles, which can be viewed as an extension of the causal identification result of Lacerda et al. (2008) from directed graphs with disjoint cycles to all possible directed graphs. We also proposed a Bayesian structural learning procedure that inferred robust and interpretable causal graphs by selecting a parsimony cause set while adjusting for covariate effects. Through both synthetic and real-world data experiments, we demonstrated the advantages of the proposed model in terms of accurately and robustly recovering the underlying causal mechanisms from longitudinal observational data by comparison with state-of-the-art alternative methods. Importantly, by applying the proposed model to a large-scale longitudinal HIV cohort study, we found interesting and clinically meaningful causal relationships among longitudinal health outcomes for people with HIV.

There are several future extensions. First, the unique causal identifiability of the proposed model relies on the assumption of causal sufficiency, since our causal identification theories are established upon the identifiability results of ICA, where the number of latent sources equals the number of observed variables. To account for unmeasured confounders, one potential approach is to leverage the identifiability results of the overcomplete ICA (Hoyer et al., 2008b; Geiger et al., 2015; Salehkaleybar et al., 2020; Adams et al., 2021), in which the number of latent sources exceeds the number of observed variables. Second, the proposed framework is built upon a linear additive model with non-Gaussian noises. Theoretical analyses for more flexible modeling choices such as the non-linear additive model and the functional model, including the development of new causal identification theories and structural learning methods could be a future direction. Lastly, the causal discovery results in the WIHS data analysis illustrate the potential clinical utility of the proposed framework. In particular, obesity/over-weight was estimated to be a direct cause for depressive symptoms, demonstrating the need of effective weight management for people with HIV (Parra-Rodriguez and O'Halloran, 2023). Therefore, the proposed framework can be used by physicians to develop better combination therapies for commorbidites of HIV, potentially improving the long-term health outcomes and quality of life for people with HIV.

Acknowledgments

Xu's research was partially supported by NSF 1940107, NSF 1918854, and NIH R01MH128085. Ni's research was partially supported by NSF DMS-2112943 and NIH 1R01GM148974-01. We are particularly grateful to the editor and three anonymous referees for their constructive and detailed comments that have significantly improved the paper.

Appendix A. Technical Proofs

A.1 Proof of Theorem 9

Proof Let $\mathcal{G}$ denote a directed graph with a total number of N directed cycles $\mathcal{O}_1, \ldots, \mathcal{O}_N$, where each pair of the two directed cycles $\mathcal{O}_n$ and $\mathcal{O}_{n'}$ can be joint with each other without complete overlap, i.e., it is possible to have $\mathcal{O}_n \cap \mathcal{O}_{n'} \neq \emptyset$, but not $\mathcal{O}_n \cap \mathcal{O}_{n'} = \mathcal{O}_n = \mathcal{O}_{n'}$, for $1 \leq n \neq n' \leq N$. Assume that there exist N variables $Y_1 \in \mathcal{O}_1, \ldots, Y_N \in \mathcal{O}_N$ in $\mathcal{G}$, each of which has its own instrumental variable I_{Y_n}, for $n = 1, \ldots, N$. Note that these N variables are not necessarily distinct, i.e., it is possible to have $Y_n = Y_{n'}$, for $1 \leq n \neq n' \leq N$.

Let $\mathcal{G}^+$ denote the directed graph formed by incorporating these instrumental variables into $\mathcal{G}$. By the definition of the instrumental variable (i.e., Definition 2), if the collection of instrumental variables $\mathcal{I} = \{I_{Y_1}, \ldots, I_{Y_N}\}$ introduces an additional directed cycle $\mathcal{O}^+$ in $\mathcal{G}^+$, then $\mathcal{O}^+$ only involves variables in $\mathcal{I}$, and will be disjoint with $\mathcal{O}_1, \ldots, \mathcal{O}_N$. By Lemma 17, the reversal of disjoint cycles is a necessary condition for two directed graphs to belong to the same ICA equivalence class. Therefore, there exists a one-to-many mapping from the ICA equivalence class of $\mathcal{G}$ to the ICA equivalence class of $\mathcal{G}^+$. In particular, if $\mathcal{I}$ introduces additional directed cycles, each graph in the ICA equivalence class of $\mathcal{G}$ will correspond to multiple directed graphs in the ICA equivalence class of $\mathcal{G}^+$, and the latter only differ from each other in the part that *only* involves the collection of instrumental variables $\mathcal{I}$. On the other hand, if $\mathcal{I}$ does not introduce additional directed cycles, then each graph in the ICA equivalence class of $\mathcal{G}$ will correspond to only one directed graph in the ICA equivalence class of $\mathcal{G}^+$. We will show later that the unique identification of $\mathcal{G}$ is independent of the part that *only* involves the collection of instrumental variables $\mathcal{I}$. Therefore, without loss of generality, we assume that $\mathcal{I}$ does not introduce additional directed cycles, and then drive the unique identification of $\mathcal{G}$ from the unique identification of $\mathcal{G}^+$.

Consider any directed cyclic graph $(\mathcal{G}^+)'$ in the ICA equivalence class of $\mathcal{G}^+$ obtained by performing the two steps described in Lemma 17. Note that by the definition of instrumental variable (i.e., Definition 2), $I_{Y_{n^\star}}$ is a special case of $\mathrm{pa}_{\mathcal{G}^+ \setminus \widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star})$, for $n^\star = 1, \ldots, N^\star$. In particular, $I_{Y_{n^\star}} \in \mathcal{G}^+ \setminus \widetilde{\mathcal{O}}_{n^\star}$ is a parent of $Y_{n^\star}$, and $Y_{n^\star} \in \widetilde{\mathcal{O}}_{n^\star}$ is the only child of $I_{Y_{n^\star}}$ in $\mathcal{G}$. By Lemma 17(ii), the edge $I_{Y_{n^\star}} \rightarrow Y_{n^\star}$ in $\mathcal{G}^+$ will be changed to $I_{Y_{n^\star}} \rightarrow \mathrm{pa}_{\widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star})$ in $(\mathcal{G}^+)'$. In other words, the only child of $I_{Y_{n^\star}}$ will be different in $\mathcal{G}$ and $\mathcal{G}'$ (i.e., $Y_{n^\star} \in \mathcal{G} \neq \mathrm{pa}_{\widetilde{\mathcal{O}}_{n^\star}}(Y_{n^\star}) \in \mathcal{G}'$ due to no self-loops), and thus $I_{Y_{n^\star}}$ will not be an instrumental variable for $Y_{n^\star}$ in $(\mathcal{G}^+)'$. This is the critical point that leads to the identification of $\mathcal{G}^+$, and it is independent of the part that *only* involves the collection of instrumental variables $\mathcal{I}$ in $\mathcal{G}^+$.

In summary, we can identify $\mathcal{G}^+$ from $(\mathcal{G}^+)'$ by utilizing a collection of instrumental variables $I_{Y_{n^\star}}$ for $Y_{n^\star} \in \widetilde{\mathcal{O}}_{n^\star}$, where $n^\star = 1, \ldots, N^\star$. We finish the proof by noting that the above argument can be applied to any $(\mathcal{G}^+)'$ within the ICA equivalence class of $\mathcal{G}^+$. ■

A.2 Proof of Lemma 14

Proof Let $\phi : R \rightarrow R$, $R = \{r_1, \ldots, r_K\}$, denote an irreducible row-permutation applied to the unmixing matrix $\boldsymbol{W}$ associated with a directed graph $\mathcal{G}$ is admissible. Let $\mathcal{G}'$ denote the directed graph associated with the resulting unmixing matrix $\boldsymbol{W}'$ by applying ϕ to $\boldsymbol{W}$.

(i) Note that if $\phi(r_k) = r_{k'}$, then we have $\boldsymbol{W}_{r_k, r_{k'}} \neq 0$, otherwise ϕ is not admissible. In other words, there exists a directed edge $(Y_{r_{k'}} \to Y_{r_k}) \in \mathcal{E}$ in the directed graph $\mathcal{G}$.

Without loss of generality, we assume that $\phi(r_1) = r_2$, which implies $(Y_{r_2} \to Y_{r_1}) \in \mathcal{E}$. Then we need to determine $\phi(r_2) = r_{k'}$ for some $k' \neq 2$. In addition, we have $k' \neq 1$ due to the irreducible assumption. Similarly, we assume that $\phi(r_2) = r_3$, which implies $(Y_{r_3} \to Y_{r_2}) \in \mathcal{E}$. Same argument can be applied to $k = 3, \dots, K-1$, i.e., $\phi(r_k) = r_{k+1}$, which indicates the existence of directed edges $(Y_{r_3} \to Y_{r_2}) \in \mathcal{E}, \dots, (Y_{r_K} \to Y_{r_{K-1}}) \in \mathcal{E}$. Now we need to determine $\phi(r_K) = r_{k'}$ for some $k' \neq K$. Due to the irreducible assumption, we have $k' \neq 2, \dots, K-1$. Therefore, $k' = 1$, which implies $(Y_{r_1} \to Y_{r_K}) \in \mathcal{E}$, and also the existence of the directed cycle $\mathcal{O} : Y_{r_1} \leftarrow Y_{r_2} \leftarrow \cdots \leftarrow Y_{r_K} \leftarrow Y_{r_1}$ in $\mathcal{G}$.

In addition, ϕ reverses the direction of $\mathcal{O}$, i.e., we have $\mathcal{O}' : Y_{r_1} \to Y_{r_2} \to \cdots \to Y_{r_K} \to Y_{r_1}$ in the directed graph $\mathcal{G}'$ associated with the resulting unmixing matrix $\boldsymbol{W}'$.

(ii) Without loss of generality, assume $\phi(r_k) = r_{k+1}$, for $k = 1, \dots, K-1$, and $\phi(r_K) = r_1$, which indicates the existence of the directed cycle $\mathcal{O} : Y_{r_1} \leftarrow Y_{r_2} \leftarrow \cdots \leftarrow Y_{r_K} \leftarrow Y_{r_1}$ in $\mathcal{G}$, by the above argument. Here we provide proof for the cases where k ranges from 1 to $K-1$, but note that the case $k = K$ can be proven in a similar manner.

Since $\phi(r_k) = r_{k+1}$, for $k = 1, \dots, K-1$, then the r_{k+1}-th row of $\boldsymbol{W}'$ is equivalent to the r_k-th row of $\boldsymbol{W}$ after appropriate row-scaling. This implies that $\mathrm{pa}_{\mathcal{G} \setminus \mathcal{O}}(Y_{r_k})$ will be the parent of $Y_{r_{k+1}}$ in $\mathcal{G}'$, for $k = 1, \dots, K-1$. The proof is finished by noting that $Y_{r_{k+1}}$ is the only parent of Y_{r_k} inside the directed cycle $\mathcal{O}$ in $\mathcal{G}$, i.e., $Y_{r_{k+1}} = \mathrm{pa}_{\mathcal{O}}(Y_{r_k})$.

■

A.3 Proof of Proposition 16

Proof The proposition is trivial if the admissible row-permutation ϕ is irreducible.

Now assume that ϕ is not irreducible. By the definition of irreducible row-permutation, there exists a row-permutation $\phi_1 : R_1 \to R_1$, where $R_1 \subset R$, such that $\phi_1(R_1) = \phi(R_1) = R_1$. Let $R_2 = R \setminus R_1$, then there exists another row-permutation $\phi_2 : R_2 \to R_2$ such that $\phi_2(R_2) = \phi(R_2) = R_2$. This is because if ϕ_2 is not a bijective mapping from R_2 to R_2, then ϕ can not be a bijective mapping from R to R. Furthermore, since $R_1 \cap R_2 = \emptyset$, both ϕ_1 and ϕ_2 are admissible, otherwise ϕ can not be an admissible row-permutation. Therefore, $\{\phi_1, \phi_2\}$ forms a collection of disjoint admissible row-permutations that is equivalent to ϕ.

If both ϕ_1 and ϕ_2 are irreducible, then the proposition is proved. Since $|R| < \infty$, if ϕ_1 and ϕ_2 are not irreducible, then we can always apply the aforementioned procedure finite times to find two collections of disjoint admissible row-permutations that are equivalent to ϕ_1 and ϕ_2, respectively, each of which only consists of row-permutations that are irreducible. Combining these two collections, we have a collection of disjoint admissible irreducible row-permutations that is equivalent to ϕ. ■

A.4 Proof of Lemma 17

Proof The first part of this Lemma is a direct corollary of Lemma 14 and Proposition 16.

Now consider applying steps (i) and (ii) to any choice of disjoint cycles in any directed graph $\mathcal{G}$. Without loss of generality, assume that the disjoint directed cycles are $\widetilde{\mathcal{O}}_{n^\star} : Y_{r_1} \leftarrow Y_{r_2} \leftarrow \cdots \leftarrow Y_{r_{K_{n^\star}}} \leftarrow Y_{r_1}$, for $n^\star = 1, \ldots, N^\star$. Then performing steps (i) and (ii) on $\mathcal{G}$ is equivalent to applying the row-permutation ϕ to the unmixing matrix $\boldsymbol{W}$ associated with $\mathcal{G}$, where $\phi(r_{k_{n^\star}}) = r_{k_{n^\star}+1}$ for $k_{n^\star} = 1, \ldots, K_{n^\star} - 1$ and $\phi(r_{K_{n^\star}}) = \phi(r_1)$, $n^\star = 1, \ldots, N^\star$. Note that due to the existence of edges $Y_{r_{k_{n^\star}}} \leftarrow Y_{r_{k_{n^\star}+1}}$ for $k_{n^\star} = 1, \ldots, K_{n^\star} - 1$ and $Y_{r_{K_{n^\star}}} \leftarrow Y_{r_1}$, the row-permutation ϕ is admissible. Consequently, the resulting directed graph $\mathcal{G}'$ remains in the same ICA equivalence class as $\mathcal{G}$. ■

A.5 Proof of Corollary 18

Proof Consider the following coefficient matrix $\boldsymbol{B}$ of the proposed model (1):

$$\boldsymbol{B} = \begin{array}{ccccccccccccccc|c}
\boldsymbol{Y}_J & \boldsymbol{Y}_{J-1} & \cdots & \boldsymbol{Y}_{J-L_y} & \boldsymbol{Y}_{J-L_y-1} & \cdots & \boldsymbol{Y}_1 & \boldsymbol{X}_J & \boldsymbol{X}_{J-1} & \cdots & \boldsymbol{X}_{J-L_x} & \boldsymbol{X}_{J-L_x-1} & \cdots & \boldsymbol{X}_1 & \\
\boldsymbol{B}_0 & \boldsymbol{B}_1 & \cdots & \boldsymbol{B}_{L_y} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots & \boldsymbol{A}_{L_x} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{Y}_J \\
\boldsymbol{0} & \boldsymbol{B}_0 & \cdots & \boldsymbol{B}_{L_y-1} & \boldsymbol{B}_{L_y} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_0 & \cdots & \boldsymbol{A}_{L_x-1} & \boldsymbol{A}_{L_x} & \cdots & \boldsymbol{0} & \boldsymbol{Y}_{J-1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B}_0 & \boldsymbol{B}_1 & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{A}_0 & \boldsymbol{A}_1 & \cdots & \boldsymbol{0} & \boldsymbol{Y}_{J-L_y} \\
\boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{B}_0 & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{A}_0 & \cdots & \boldsymbol{0} & \boldsymbol{Y}_{J-L_y-1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{B}_0 & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_0 & \boldsymbol{Y}_1
\end{array},$$

and the following simpler (cross-sectional) SCM defined by the instantaneous causal effects $\boldsymbol{B}_0$ at each time point j,

$$\boldsymbol{Y}_j = \boldsymbol{B}_0 \boldsymbol{Y}_j + \boldsymbol{E}_j,$$

where $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jQ})$ and $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jS})$, $j = 1, 2, \ldots, J$. We will first prove that the unique identification of the proposed model (1) (with the graph $\mathcal{G}$) can be achieved by leveraging the unique identification of the SCM (with the graph $\mathcal{G}_j$) described above.

By Lemma 17, any two directed graphs within the same ICA equivalence class can be obtained from each other by performing two steps, both of which are uniquely determined by the disjoint cycles to be reversed. Consequently, for any two directed graphs that share the same cycles, a one-to-one correspondence exists between their ICA equivalence classes. Note that the *only* potential cycles in $\boldsymbol{B}$ are within the instantaneous causal effects $\boldsymbol{B}_0$, which characterizes the same causal relationships for $\boldsymbol{Y}_j$'s at any time point j. In other words, if $\mathcal{G}'$ is obtained by reversing disjoint cycles in $\boldsymbol{B}_0$ of $\mathcal{G}$, this reversal of disjoint cycles will be applied *simultaneously* to all the rows associated with $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_J$ in $\boldsymbol{B}$. Therefore, $\mathcal{G}$ and $\mathcal{G}_j$ essentially share the same cycles, and there exists a one-to-one correspondence between their ICA equivalence classes. Then the unique identification of $\mathcal{G}$ can be achieved through the unique identification of $\mathcal{G}_j$.

The proof is concluded by demonstrating the unique identification of $\mathcal{G}_j$. By Theorem 9, for any cycle $\mathcal{O}_j$ in $\mathcal{G}_j$, if there exists a variable $Y_{jq} \in \mathcal{O}_j$ which has an instrumental variable $I_{Y_{jq}}$, then $\mathcal{G}_j$ can be uniquely identified from its ICA equivalence class. ■

Appendix B. MCMC Algorithm

1. Update $\beta_{\ell qp}$ and $\alpha_{\ell qs}$

   The full conditional distribution for $\beta_{\ell qp}$ and $\alpha_{\ell qs}$ is

$$p(\beta_{\ell qp}, \alpha_{\ell qs} \mid \cdot) \propto p(\beta_{\ell qp}) p(\alpha_{\ell qs}) \prod_{i} \prod_{j=\ell+1}^{J_i} \mathcal{N}(Y_{ijq} \mid Y^*_{ijq}, \frac{\sigma_q^2}{\tau_{ijq}}) |\boldsymbol{I} - \boldsymbol{B}_0|,$$

   where

$$Y^*_{ijq} = \mu_q + \sum_{\ell'=0}^{L'_y} \sum_{p=1}^{Q} Y_{i,j-\ell',p} \beta_{\ell' qp} + \sum_{\ell'=0}^{L'_x} \sum_{s=1}^{S} X_{i,j-\ell',s} \alpha_{\ell' qs},$$
$$L'_y = \min(j-1, L_y), \text{ and } L'_x = \min(j-1, L_x).$$

   1.1. Update β_{0qp}

   Since the full conditional distribution for β_{0qp} involves the additional term $|\boldsymbol{I} - \boldsymbol{B}_0|$, there is no closed form solution. Therefore, we will update it using the Metropolis-Hasting algorithm. At each Metropolis-Hasting step, we will accept the move only if the maximum modulus of $\boldsymbol{B}_0$'s eigenvalues is strictly less than 1, to ensure that the proposed model is well-defined.

   1.2. Update $\beta_{\ell qp}, \ell > 0$ and $\alpha_{\ell qs}, \ell \geq 0$

   The full conditional distribution for $\beta_{\ell qp}, \ell > 0$ is

$$p(\beta_{\ell qp} \mid \cdot) \propto \mathcal{N}(0, \gamma^{\beta}_{\ell qp} \nu^{\beta}_{\ell qp}) \prod_{i} \prod_{j=\ell+1}^{J_i} \mathcal{N}(Y_{ijq} \mid Y^*_{ijq}, \frac{\sigma_q^2}{\tau_{ijq}}) \propto \mathcal{N}(\mu_\beta, \sigma_\beta^2),$$

   where

$$\sigma_\beta^2 = \left(\sum_{i} \sum_{j=\ell+1}^{J_i} \frac{(Y_{i,j-\ell,p})^2}{\sigma_q^2 / \tau_{ijq}} + \frac{1}{\gamma^{\beta}_{\ell qp} \nu^{\beta}_{\ell qp}} \right)^{-1},$$
$$\mu_\beta = \sigma_\beta^2 \left(\sum_{i} \sum_{j=\ell+1}^{J_i} \frac{Y_{i,j-\ell,p}}{\sigma_q^2 / \tau_{ijq}} \widetilde{Y}_{ijq} \right),$$
$$\widetilde{Y}_{ijq} = Y_{ijq} - \mu_q - \sum_{\ell', p' \neq \ell, p} Y_{i,j-\ell',p'} \beta_{\ell' qp'} - \sum_{\ell'=0}^{L'_x} \sum_{s=1}^{S} X_{i,j-\ell',s} \alpha_{\ell' qs}.$$

   We can update $\alpha_{\ell qs}, \ell \geq 0$ analogously.

2. Update $\gamma^{\beta}_{\ell qp}$ and $\gamma^{\alpha}_{\ell qs}$

The full conditional distribution for $\gamma^{\beta}_{\ell qp}$ is

$$p(\gamma^{\beta}_{\ell qp} \mid \cdot) \propto \left\{\rho^{\beta}\delta_1(\gamma^{\beta}_{\ell qp}) + (1-\rho^{\beta})\delta_{\nu_0}(\gamma^{\beta}_{\ell qp})\right\}\mathcal{N}(\beta_{\ell qp} \mid 0, \gamma^{\beta}_{\ell qp}\nu^{\beta}_{\ell qp}),$$

i.e.,

$$\frac{P(\gamma^{\beta}_{\ell qp} = 1 \mid \beta_{\ell qp}, \nu^{\beta}_{\ell qp}, \rho^{\beta})}{P(\gamma^{\beta}_{\ell qp} = \nu_0 \mid \beta_{\ell qp}, \nu^{\beta}_{\ell qp}, \rho^{\beta})} = \frac{\sqrt{\nu_0}\rho^{\beta}}{1-\rho^{\beta}} \exp\left\{\frac{(1-\nu_0)(\beta_{\ell qp})^2}{2\nu_0\nu^{\beta}_{\ell qp}}\right\}.$$

We can update $\gamma^{\alpha}_{\ell qs}$ analogously.

3. Update $\nu^{\beta}_{\ell qp}$ and $\nu^{\alpha}_{\ell qs}$

   The full conditional distribution for $\nu^{\beta}_{\ell qp}$ is

$$\begin{aligned} p(\nu^{\beta}_{\ell qp} \mid \cdot) &\propto \text{Inverse-Gamma}(a_\nu, b_\nu)\mathcal{N}(\beta_{\ell qp} \mid 0, \gamma^{\beta}_{\ell qp}\nu^{\beta}_{\ell qp}) \\ &\propto \text{Inverse-Gamma}\left(a_\nu + \frac{1}{2}, b_\nu + \frac{(\beta_{\ell qp})^2}{2\gamma^{\beta}_{\ell qp}}\right). \end{aligned}$$

   We can update $\nu^{\alpha}_{\ell qs}$ analogously.

4. Update ρ^{β} and ρ^{α}

   The full conditional distribution for ρ^{β} is

$$\begin{aligned} p(\rho^{\beta} \mid \cdot) &\propto \text{Beta}(a_\rho, b_\rho)\prod_{\ell=0}^{L_y}\prod_{q=1}^{Q}\prod_{p=1}^{Q}\left\{\rho^{\beta}\delta_1(\gamma^{\beta}_{\ell qp}) + (1-\rho^{\beta})\delta_{\nu_0}(\gamma^{\beta}_{\ell qp})\right\} \\ &\propto \text{Beta}\left(a_\rho + \sum_{\ell=0}^{L_y}\sum_{q=1}^{Q}\sum_{p=1}^{Q}\delta_1(\gamma^{\beta}_{\ell qp}), b_\rho + \sum_{\ell=0}^{L_y}\sum_{q=1}^{Q}\sum_{p=1}^{Q}\delta_{\nu_0}(\gamma^{\beta}_{\ell qp})\right). \end{aligned}$$

   We can update ρ^{α} analogously.

5. Update τ_{ijq}

   The full conditional distribution for τ_{ijq} is

$$\begin{aligned} p(\tau_{ijq} \mid \cdot) &\propto \text{Inverse-Gamma}(1, \frac{1}{8})\mathcal{N}(Y_{ijq} \mid Y^*_{ijq}, \frac{\sigma_q^2}{\tau_{ijq}}) \\ &\propto \text{Inverse-Gaussian}\left(\frac{\sigma_q}{2|Y_{ijq} - Y^*_{ijq}|}, \frac{1}{4}\right). \end{aligned}$$

6. Update μ_q

   The full conditional distribution for μ_q is

$$p(\mu_q \mid \cdot) \propto \mathcal{N}(0, \sigma_\mu^2)\prod_{i}\prod_{j=1}^{J_i}\mathcal{N}(Y_{ijq} \mid Y^*_{ijq}, \frac{\sigma_q^2}{\tau_{ijq}}) \propto \mathcal{N}(\mu_n, \sigma_n^2),$$

where

$$\sigma_n^2 = \left(\sum_i \sum_{j=1}^{J_i} \frac{\tau_{ijq}}{\sigma_q^2} + \frac{1}{\sigma_\mu^2} \right)^{-1},$$

$$\mu_n = \sigma_n^2 \left(\sum_i \sum_{j=1}^{J_i} \widetilde{Y}_{ijq} \frac{\tau_{ijq}}{\sigma_q^2} \right),$$

$$\widetilde{Y}_{ijq} = Y_{ijq} - \sum_{\ell'=0}^{L_y'} \sum_{p=1}^{Q} Y_{i,j-\ell',p} \beta_{\ell' qp} - \sum_{\ell'=0}^{L_x'} \sum_{s=1}^{S} X_{i,j-\ell',s} \alpha_{\ell' qs}.$$

7. Update σ_q^2

   The full conditional distribution for σ_q^2 is

$$\begin{aligned} p(\sigma_q^2 \mid \cdot) &\propto \text{Inverse-Gamma}(a_\sigma, b_\sigma) \prod_i \prod_{j=1}^{J_i} \mathcal{N}(Y_{ijq} \mid Y_{ijq}^*, \frac{\sigma_q^2}{\tau_{ijq}}) \\ &\propto \text{Inverse-Gamma} \left(a_\sigma + \frac{1}{2} \sum_i \sum_{j=1}^{J_i} 1, b_\sigma + \frac{1}{2} \sum_i \sum_{j=1}^{J_i} (Y_{ijq} - Y_{ijq}^*)^2 \tau_{ijq} \right). \end{aligned}$$

Appendix C. Additional Simulation Studies

C.1 Identifying ICA Equivalence Classes

To further illustrate the superior performance of the proposed Bayesian learning algorithm in identifying ICA equivalence classes, we conducted additional simulation studies using a different simulated true causal graph from the one discussed in Section 5.1. In particular, the simulated true causal graph was set to be $\mathcal{G}$ in the left panel of Figure C1 with an instrumental variable $I_{Y_{j2}}$ for Y_{j2}. The right panel of Figure C1 plots the only graph $\mathcal{G}'$ in the ICA equivalence class of $\mathcal{G}$ for which the corresponding SCM is stable. We considered two types of instrumental variables: $I_{Y_{j2}} = X_{j1} \sim N(0,1)$, and $I_{Y_{j2}} = Y_{j-1,2}$. We generated the simulated true Y_{ijq} using the following data-generating process,

$$\begin{cases} Y_{ij1} = \mu_1 - 0.4 \times Y_{ij2} + 0.4 \times Y_{ij3} + e_{ij1}, \\ Y_{ij2} = \mu_2 + 2 \times Y_{ij1} - 0.5 \times I_{Y_{ij2}} + e_{ij2}, \\ Y_{ij3} = \mu_3 + 0.5 \times Y_{ij2} + e_{ij3}. \end{cases}$$

The rest of the simulation setup was the same as in Section 5.1.

Figures C2 summarizes all the individual casual graphs identified by the proposed model when X_{j1} and $Y_{j-1,2}$ were used as the instrumental variable $I_{Y_{j2}}$. The proposed Bayesian learning algorithm successfully identified $\mathcal{G}$ and $\mathcal{G}'$ in the ICA equivalence class.

Figures C3-C6 present the estimated causal graphs under alternative methods. When $Y_{j-1,2}$ was used as the instrumental variable $I_{Y_{j2}}$, PCMCI$^+$ and VAR-LiNGAM identified 14 and 32 individual graphs across 100 replications, respectively. Therefore, for clarity and effective presentation, we reported only the top 10 individual causal graphs identified by PCMCI$^+$ and VAR-LiNGAM based on their relative frequency across 100 replications.

As shown in Figures C3-C6, PCMCI$^+$ and VAR-LiNGAM failed to identify the instantaneous directed cycles. While LiNG-D successfully recovered the ICA equivalence classes when X_{j1} was used as the instrumental variable $I_{Y_{j2}}$, it produced spurious causal links when the time-lagged variable $Y_{j-1,2}$ was used as the instrumental variable $I_{Y_{j2}}$. Similarly, both PCMCI$^+$ and VAR-LiNGAM also detected spurious causal links under these scenarios.

In summary, the proposed Bayesian learning algorithm demonstrated the best performance in identifying ICA equivalence classes compared to all alternative methods.

C.2 Varying Number of Visits J_i

To empirically verify the causal identification theory under the general case where different individuals have varying numbers of visits J_i, we conducted additional simulation studies. Specifically, we assumed that there were 1,000 individuals and the number of visits J_i was randomly generated between four and eight, instead of fixing at five. The other simulation setups remained the same as in simulation scenario I.

Figures C7 and C8 present all the individual causal graphs identified by the proposed model across 100 repeated experiments, using X_{j1} and $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$, respectively. We observed similar results to those in simulation scenario I, indicating that our causal identification theory applies effectively to general cases with varying J_i.

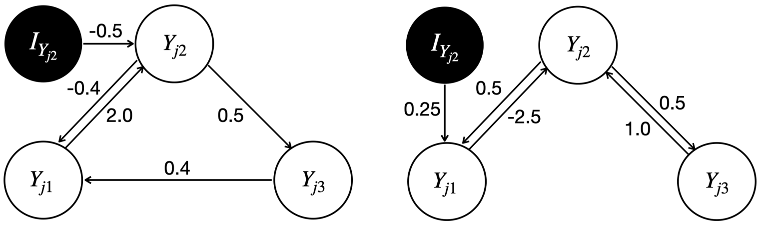
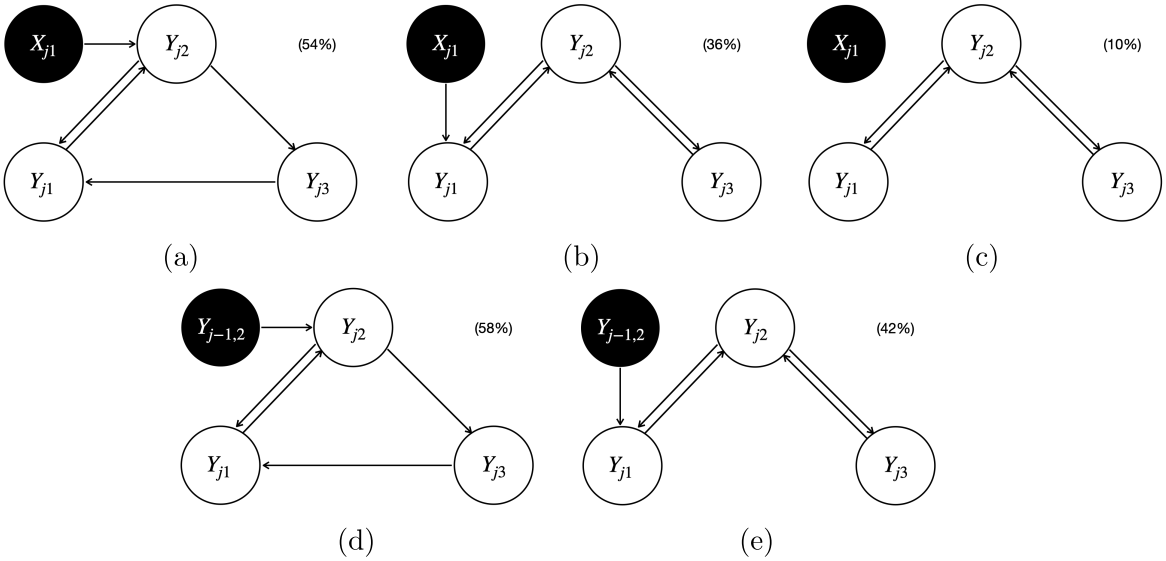


Figure C1: The simulated true causal graph $\mathcal{G}$ (left) and the only graph $\mathcal{G}'$ (right) associated with a stable SCM in its ICA equivalence class in the additional simulation studies. The instrumental variable $I_{Y_{j2}}$ for Y_{j2} is highlighted by the black circle.



(a) (b) (c)

(d) (e)

Figure C2: The estimated causal graphs under the proposed model with a sample size of 5,000, using X_{j1} (Panel(a)-(c)) and $Y_{j-1,2}$ (Panel(d)-(e)) as the instrumental variable $I_{Y_{j2}}$ in the additional simulation studies.

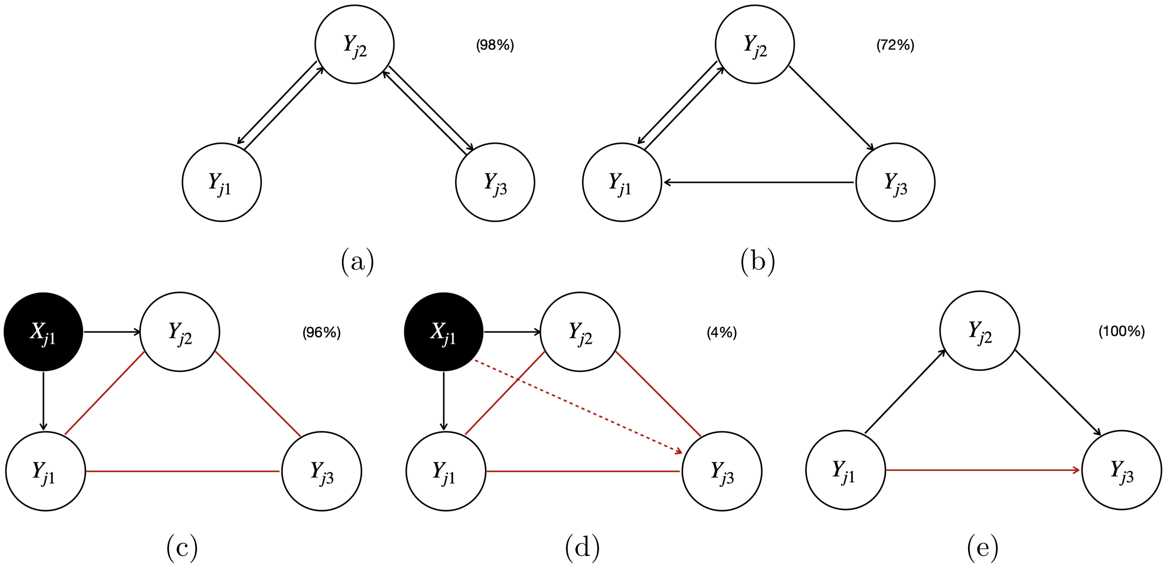


Figure C3: The estimated causal graphs under LiNG-D (Panel(a)-(b)), PCMCI$^+$ (Panel(c)-(d)), and VAR-LiNGAM (Panel(e)) with a sample size of 5,000, using X_{j1} as the instrumental variable $I_{Y_{j2}}$ in the additional simulation studies.

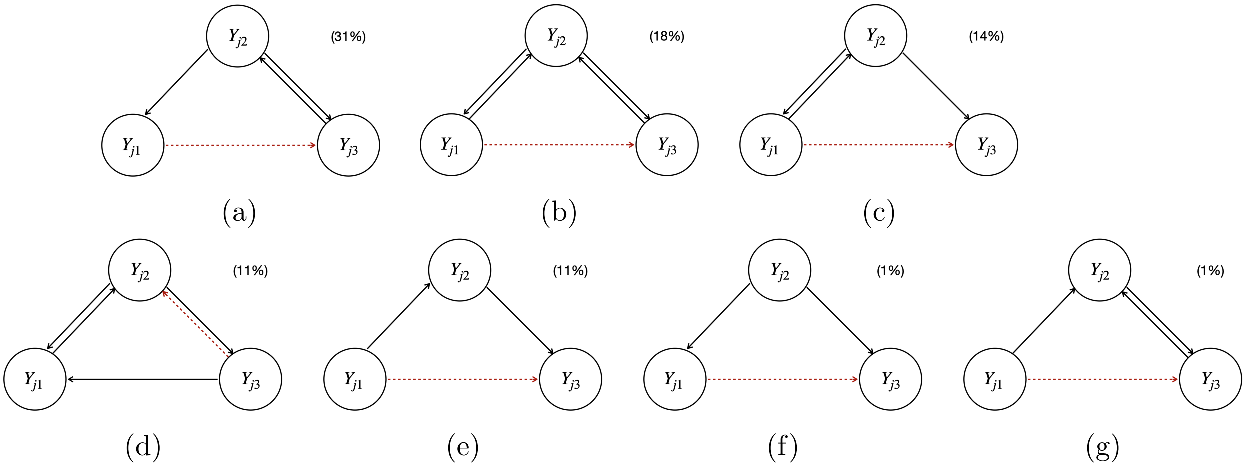


Figure C4: The estimated causal graphs under LiNG-D with a sample size of 5,000, using $Y_{j-1,2}$ as the instrumental variable $I_{Y_{j2}}$ in the additional simulation studies.

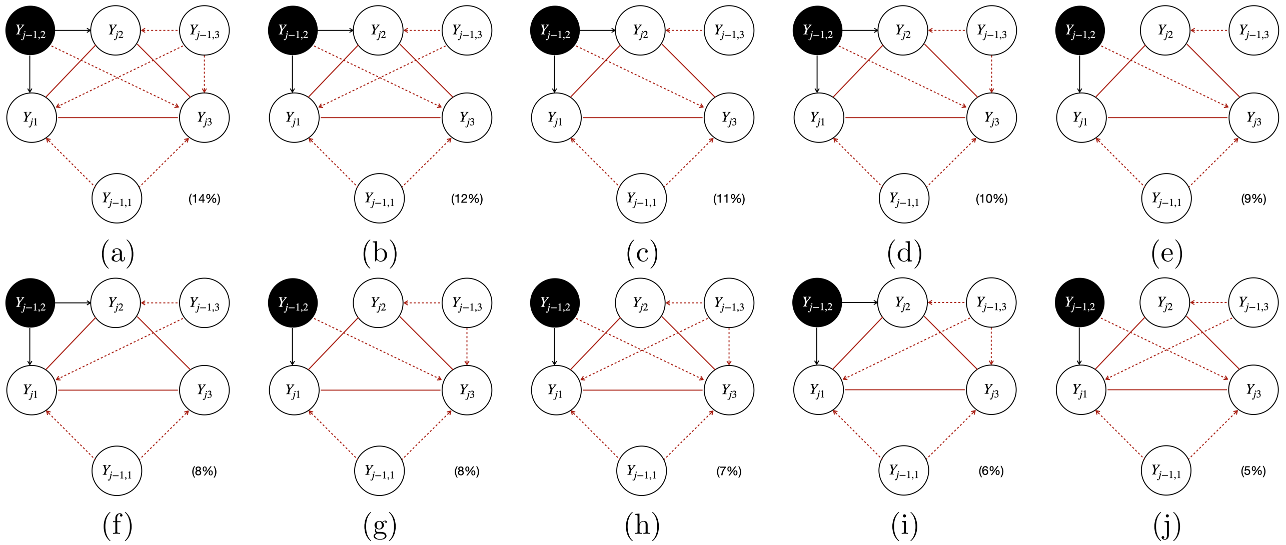


Figure C5: The estimated causal graphs under PCMCI$^+$ with a sample size of 5,000, using $Y_{j-1,2}$ as the instrumental variable $I_{Y_{j2}}$ in the additional simulation studies.

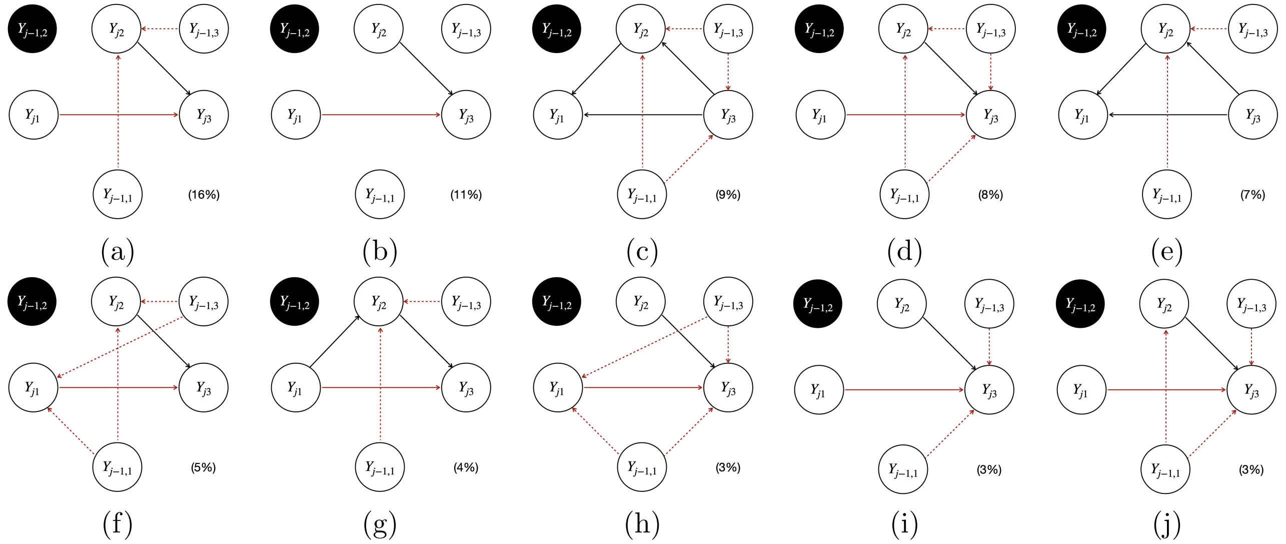


Figure C6: The estimated causal graphs under VAR-LiNGAM with a sample size of 5,000, using $Y_{j-1,2}$ as the instrumental variable $I_{Y_{j2}}$ in the additional simulation studies.

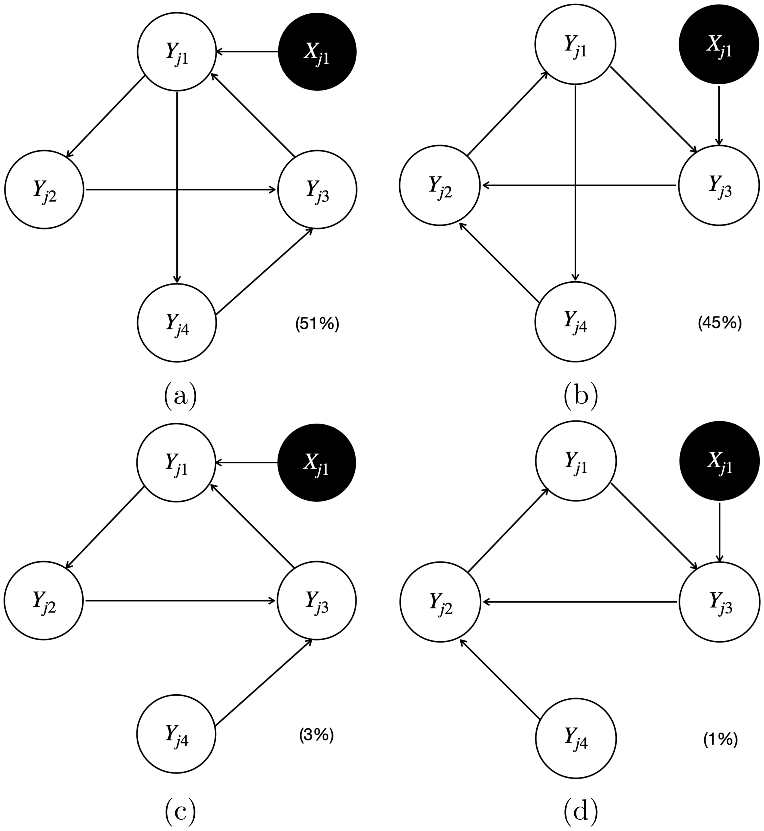


Figure C7: The estimated causal graphs under the proposed model with a varying number of visits J_i, using X_{j1} as the instrumental variable $I_{Y_{j1}}$.

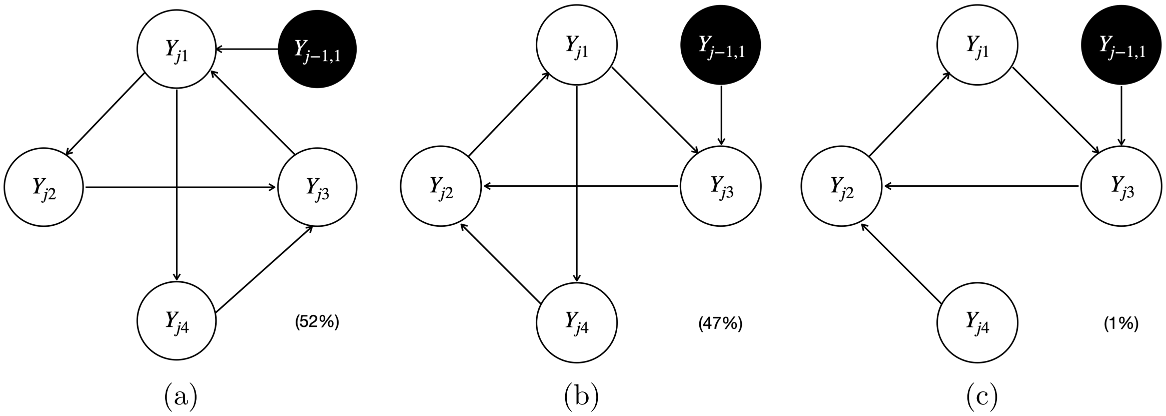


Figure C8: The estimated causal graphs under the proposed model with a varying number of visits J_i, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$.

Appendix D. Prior Specification for WIHS Data Analysis

For better biomedical interpretations in the WIHS data analysis, we slightly modified the spike-and-slab priors on both $\beta_{\ell qp}$ and $\alpha_{\ell qs}$ introduced in Section 4 of the manuscript. We describe the prior for $\beta_{\ell qp}$, and the prior for $\alpha_{\ell qs}$ is analogously defined.

Specifically, we impose a conditional prior on the instantaneous (i.e., β_{0qp}) and time-lagged effects $\beta_{\ell qp}, \ell > 0$ by assuming that (i) $\beta_{\ell qp} = 0, \ell > 0$ if $\beta_{0qp} = 0$; and (ii) $\mathrm{sign}(\beta_{\ell qp})$ is the same for all $\ell \geq 0$ such that $\beta_{\ell qp} \neq 0$. We expand $\beta_{\ell qp} = \eta_{\ell qp}\xi_{qp}$ to be a product of two scalars $\eta_{\ell qp}$ and ξ_{qp}. The scalar ξ_{qp} is a sign indicator for $\beta_{\ell qp}$, which represents whether the causal effect is positive or negative when $\beta_{\ell qp} \neq 0$. We assume that if there exist causal relationships between two variables, then the instantaneous (i.e., $Y_{jq} \leftarrow Y_{jp},\ p \neq q$) and time-lagged (i.e., $Y_{jq} \leftarrow Y_{j-\ell,p}, \ell > 0$) effects are on the same direction, therefore, ξ_{qp} does not depend on ℓ. We assign equal prior probabilities for either direction, i.e., $P(\xi_{qp} = 1) = P(\xi_{qp} = -1) = 0.5$. The (non-negative) scalar $\eta_{\ell qp}$ is assigned a spike-and-slab prior $\eta_{\ell qp} \sim \mathcal{N}^+(0, \gamma_{\ell qp}\nu_{\ell qp})$ with $\nu_{\ell qp} \sim \text{Inverse-Gamma}(a_\nu, b_\nu)$ and $\gamma_{\ell qp} \sim \rho\delta_1(\gamma_{\ell qp}) + (1-\rho)\delta_{\nu_0}(\gamma_{\ell qp})$, where $\mathcal{N}^+$ denotes the normal distribution truncated by $[0, \infty)$, and $\delta_x(\cdot)$ denotes the Dirac measure at x. By the conditional prior assumption, $\gamma_{\ell qp} = \nu_0, \ell > 0$ if $\gamma_{0qp} = \nu_0$.

The rest of the prior specifications are the same as before. For the MCMC algorithm, we only need to replace Step 1 in Appendix B with the following Step 1*.

1*. Update $\beta_{\ell qp} = \eta^{\beta}_{\ell qp}\xi^{\beta}_{qp}$ and $\alpha_{\ell qs} = \eta^{\alpha}_{\ell qs}\xi^{\alpha}_{qs}$

The full conditional distribution for $\beta_{\ell qp}$ and $\alpha_{\ell qs}$ is

$$p(\beta_{\ell qp}, \alpha_{\ell qs} \mid \cdot) \propto p(\beta_{\ell qp})p(\alpha_{\ell qs}) \prod_i \prod_{j=\ell+1}^{J_i} \mathcal{N}(Y_{ijq} \mid Y^*_{ijq}, \frac{\sigma_q^2}{\tau_{ijq}})|\boldsymbol{I} - \boldsymbol{B}_0|,$$

where

$$Y^*_{ijq} = \mu_q + \sum_{\ell'=0}^{L'_y}\sum_{p=1}^{Q} Y_{i,j-\ell',p}\beta_{\ell' qp} + \sum_{\ell'=0}^{L'_x}\sum_{s=1}^{S} X_{i,j-\ell',s}\alpha_{\ell' qs},$$
$$L'_y = \min(j-1, L_y), \text{ and } L'_x = \min(j-1, L_x).$$

1.1*. Update ξ^{β}_{qp} and ξ^{α}_{qs}

The full conditional distribution for ξ^{β}_{qp} is

$$\frac{P(\xi^{\beta}_{qp} = 1 \mid \cdot)}{P(\xi^{\beta}_{qp} = -1 \mid \cdot)} \propto \exp\left(\sum_i \sum_{j=1}^{J_i} \left\{ 2\frac{\widetilde{Y}_{ijq}\sum_{\ell'=0}^{L'_y} Y_{i,j-\ell',p}\eta^{\beta}_{\ell' qp}}{\sigma_q^2/\tau_{ijq}} + \log\left(\frac{|\boldsymbol{I} - \boldsymbol{B}_0^+|}{|\boldsymbol{I} - \boldsymbol{B}_0^-|}\right) \right\}\right),$$

where $\boldsymbol{B}_0^+$ and $\boldsymbol{B}_0^-$ denote $\boldsymbol{B}_0$ with $\xi^{\beta}_{qp} = 1$ and $\xi^{\beta}_{qp} = -1$, respectively, and

$$\widetilde{Y}_{ijq} = Y_{ijq} - \mu_q - \sum_{\ell'=0}^{L'_y}\sum_{p' \neq p} Y_{i,j-\ell',p'}\beta_{\ell' qp'} - \sum_{\ell'=0}^{L'_x}\sum_{s=1}^{S} X_{i,j-\ell',s}\alpha_{\ell' qs}.$$

The full conditional distribution for ξ^{α}_{qs} is similar, except it does not involve the additional term $|\boldsymbol{I} - \boldsymbol{B}_0^+|/|\boldsymbol{I} - \boldsymbol{B}_0^-|$. We can update ξ^{α}_{qs} analogously.

1.2*. Update η^{β}_{0qp}

Since the full conditional distribution for η^{β}_{0qp} involves the additional term $|\boldsymbol{I} - \boldsymbol{B}_0|$, there is no closed form solution. Therefore, we will update it using the Metropolis-Hasting algorithm. At each Metropolis-Hasting step, we will accept the move only if the maximum modulus of $\boldsymbol{B}_0$'s eigenvalues is strictly less than 1, to ensure that the proposed model is well-defined.

1.3*. Update $\eta^{\beta}_{\ell qp}, \ell > 0$ and $\eta^{\alpha}_{\ell qs}, \ell \geq 0$

The full conditional distribution for $\eta^{\beta}_{\ell qp}, \ell > 0$ is

$$p(\eta^{\beta}_{\ell qp} \mid \cdot) \propto \mathcal{N}^{+}(0, \gamma^{\beta}_{\ell qp}\nu^{\beta}_{\ell qp}) \prod_{i} \prod_{j=\ell+1}^{J_i} \mathcal{N}(Y_{ijq} \mid Y^{*}_{ijq}, \frac{\sigma^2_q}{\tau_{ijq}}) \propto \mathcal{N}^{+}(\mu_\eta, \sigma^2_\eta),$$

where

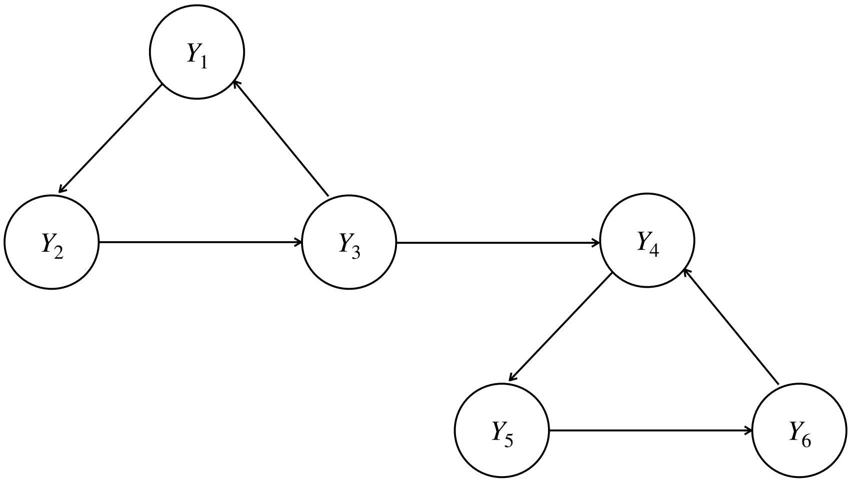
$$\sigma^2_\eta = \left(\sum_i \sum_{j=\ell+1}^{J_i} \frac{(Y_{i,j-\ell,p}\xi^{\beta}_{qp})^2}{\sigma^2_q/\tau_{ijq}} + \frac{1}{\gamma^{\beta}_{\ell qp}\nu^{\beta}_{\ell qp}} \right)^{-1},$$

$$\mu_\eta = \sigma^2_\eta \left(\sum_i \sum_{j=\ell+1}^{J_i} \frac{Y_{i,j-\ell,p}\xi^{\beta}_{qp}}{\sigma^2_q/\tau_{ijq}} \widetilde{Y}_{ijq} \right),$$

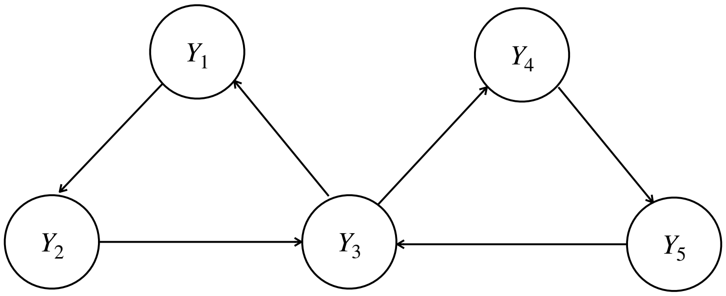
$$\widetilde{Y}_{ijq} = Y_{ijq} - \mu_q - \sum_{\ell',p' \neq \ell,p} Y_{i,j-\ell',p'}\beta_{\ell' qp'} - \sum_{\ell'=0}^{L'_x} \sum_{s=1}^{S} X_{i,j-\ell',s}\alpha_{\ell' qs}.$$

We can update $\eta^{\alpha}_{\ell qs}, \ell \geq 0$ analogously.

Appendix E. Supplementary Figures

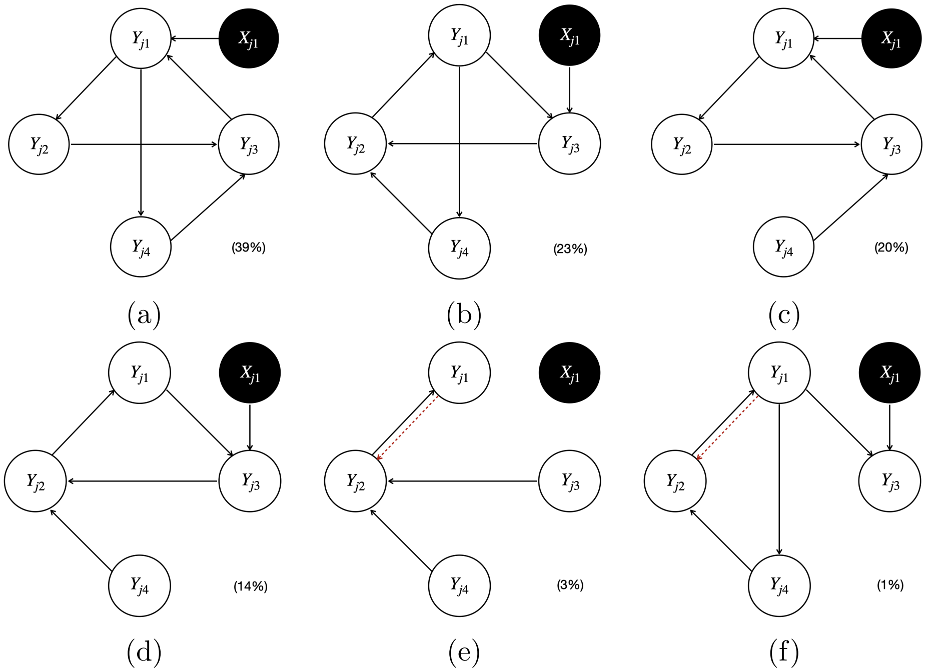


(a) $\mathcal{G}_1$ with disjoint cycles



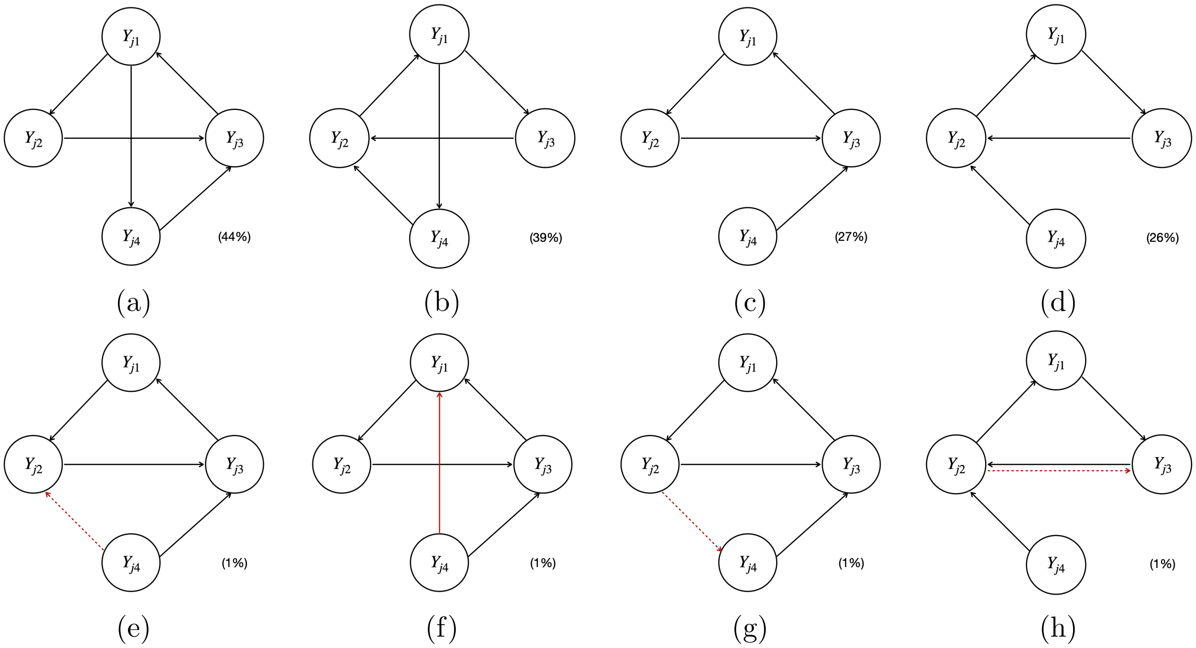
(b) $\mathcal{G}_2$ with joint cycles

Figure S1: Directed graphs that contain disjoint cycles (top) and joint cycles (bottom). In $\mathcal{G}_1$, $\mathcal{O}_1 : Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_1$ and $\mathcal{O}_2 : Y_4 \rightarrow Y_5 \rightarrow Y_6 \rightarrow Y_4$, where $\mathcal{O}_1 \cap \mathcal{O}_2 = \emptyset$. In $\mathcal{G}_2$, $\mathcal{O}_3 : Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_1$ and $\mathcal{O}_4 : Y_3 \rightarrow Y_4 \rightarrow Y_5 \rightarrow Y_3$, where $\mathcal{O}_3 \cap \mathcal{O}_4 = \{Y_3\}$.



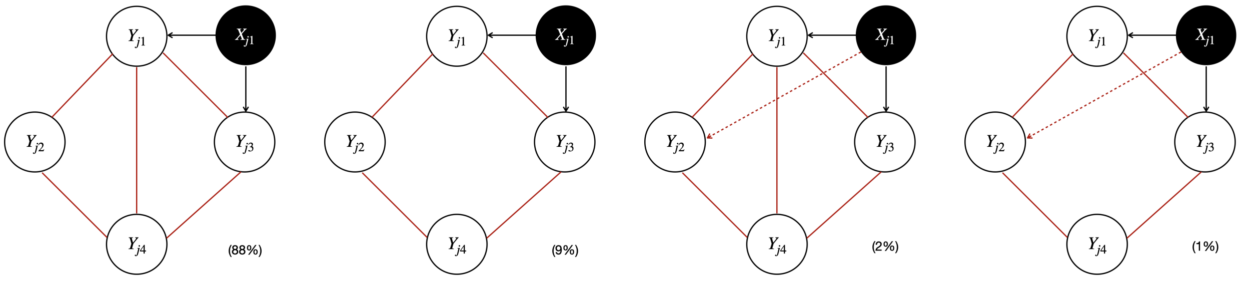
(a) (b) (c) (d) (e) (f)

Figure S2: The estimated causal graphs under the proposed model with a sample size of 2,500, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



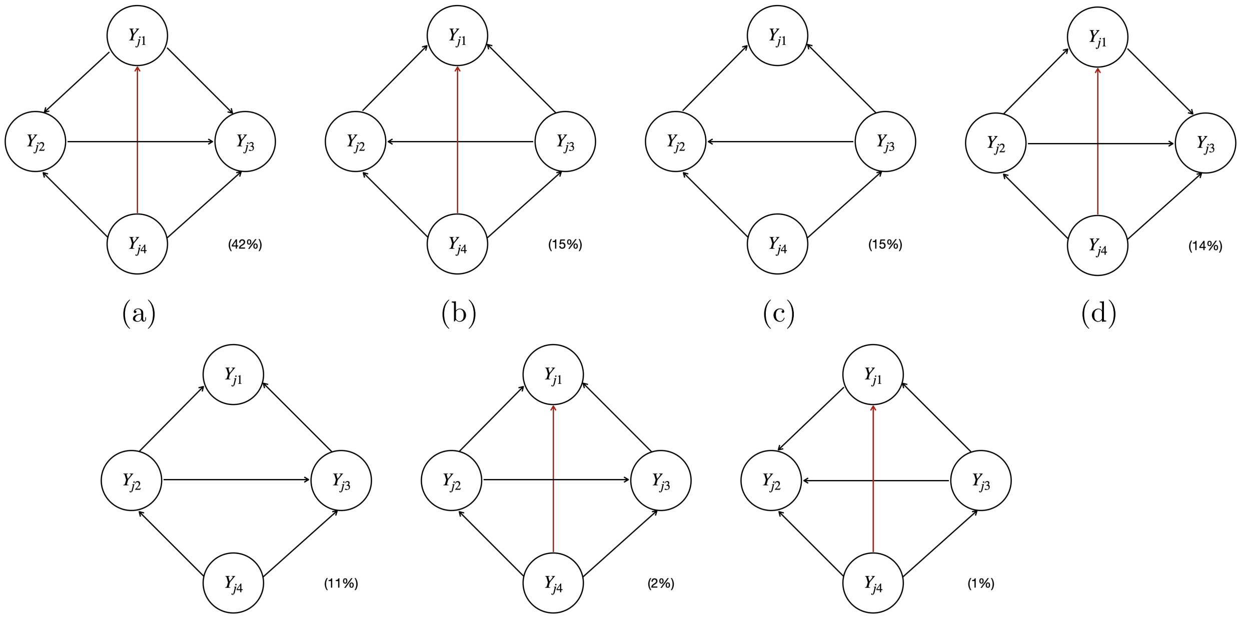
(a) (b) (c) (d) (e) (f) (g) (h)

Figure S3: The estimated causal graphs under LiNG-D with a sample size of 2,500, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c) (d)

Figure S4: The estimated causal graphs under PCMCI$^+$ with a sample size of 2,500, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c) (d)

(e) (f) (g)

Figure S5: The estimated causal graphs under VAR-LiNGAM with a sample size of 2,500, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

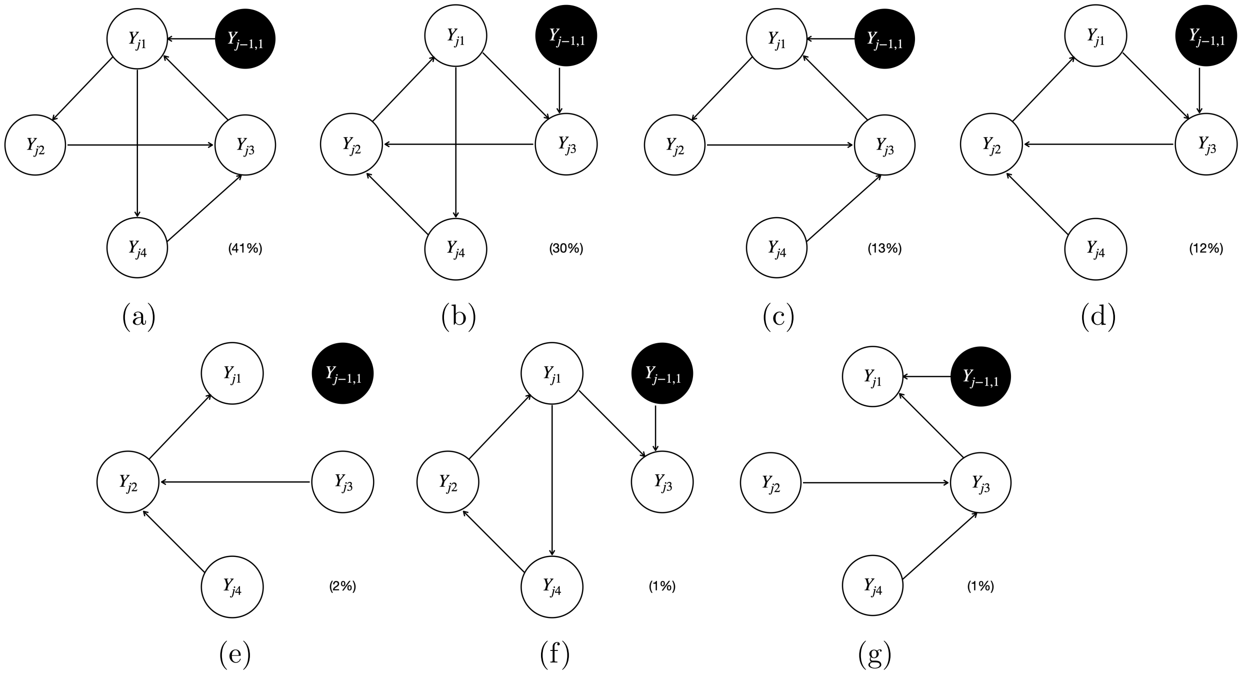


Figure S6: The estimated causal graphs under the proposed model with a sample size of 2,500, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

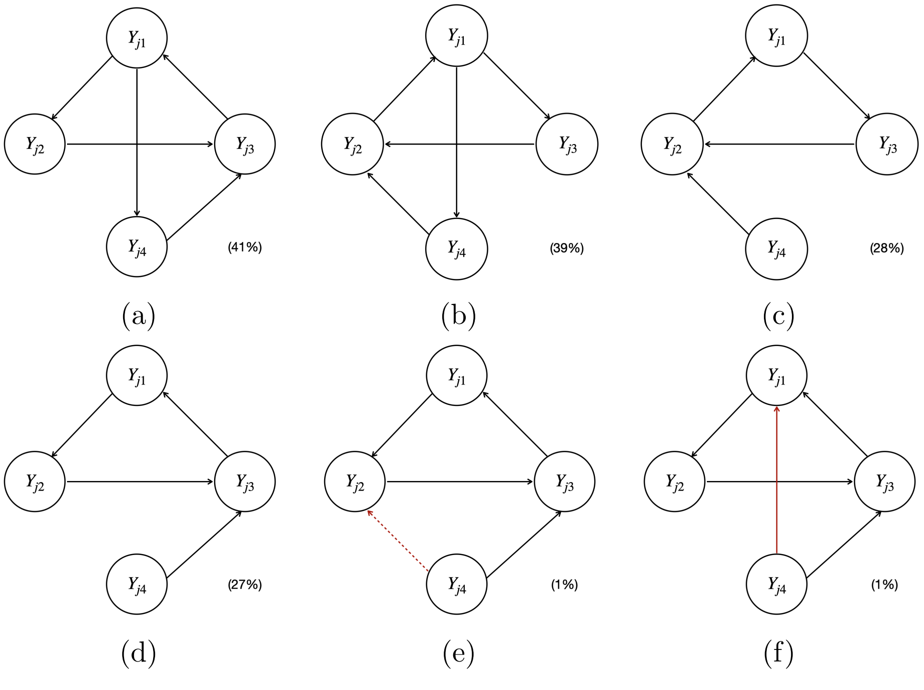
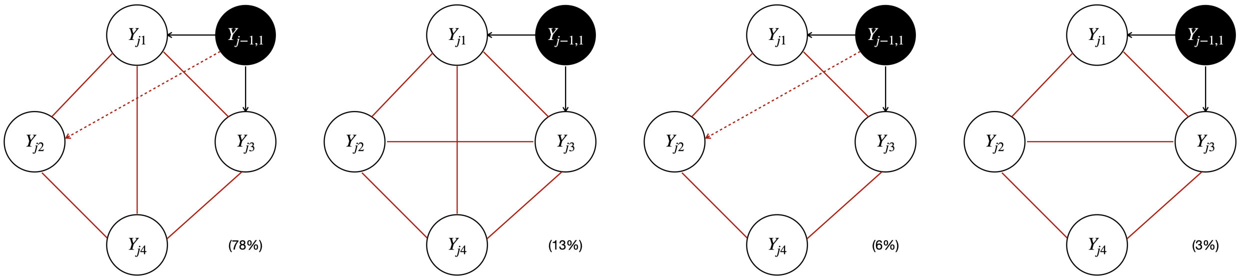
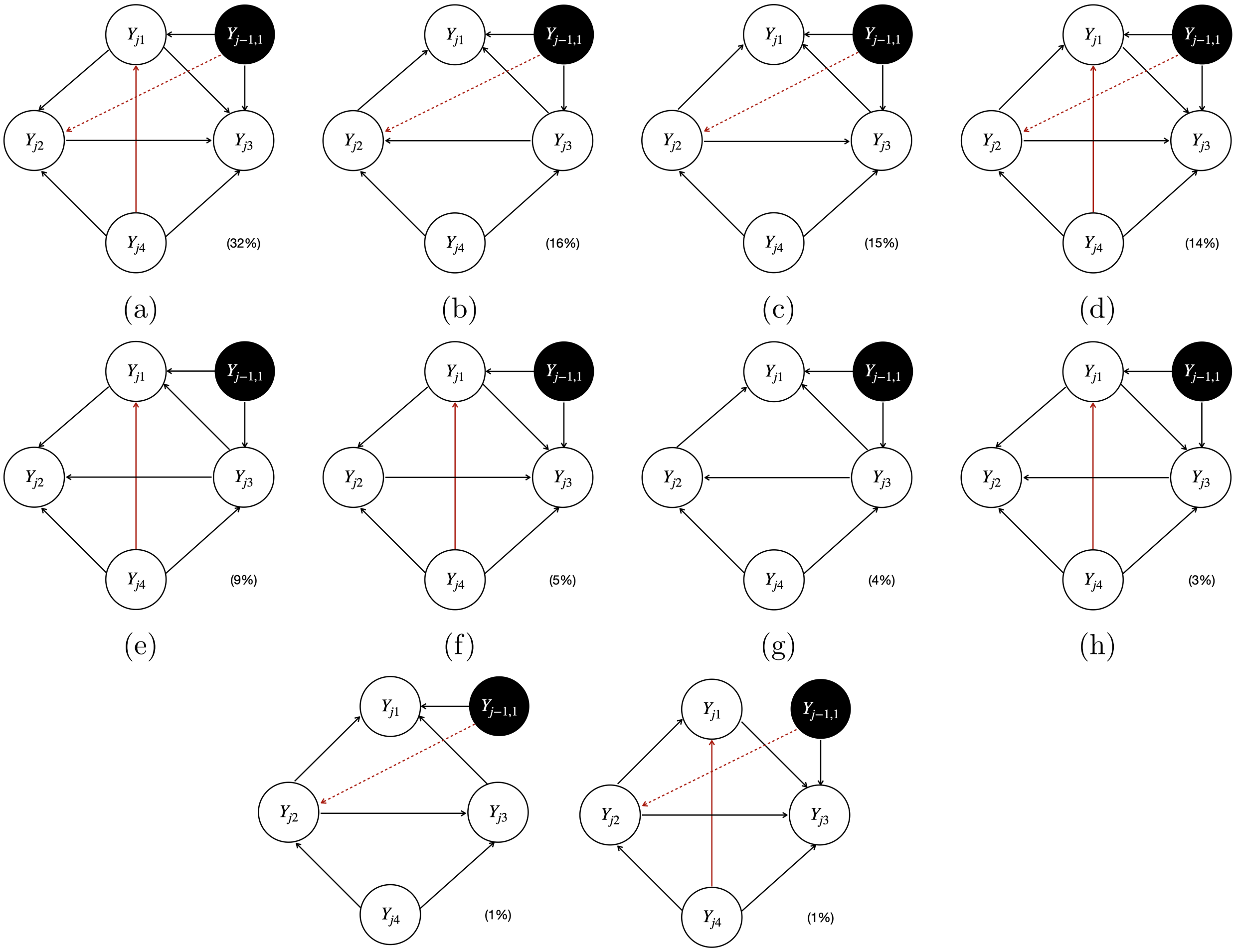


Figure S7: The estimated causal graphs under LiNG-D with a sample size of 2,500, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c) (d)

Figure S8: The estimated causal graphs under PCMCI^{+} with a sample size of 2,500, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

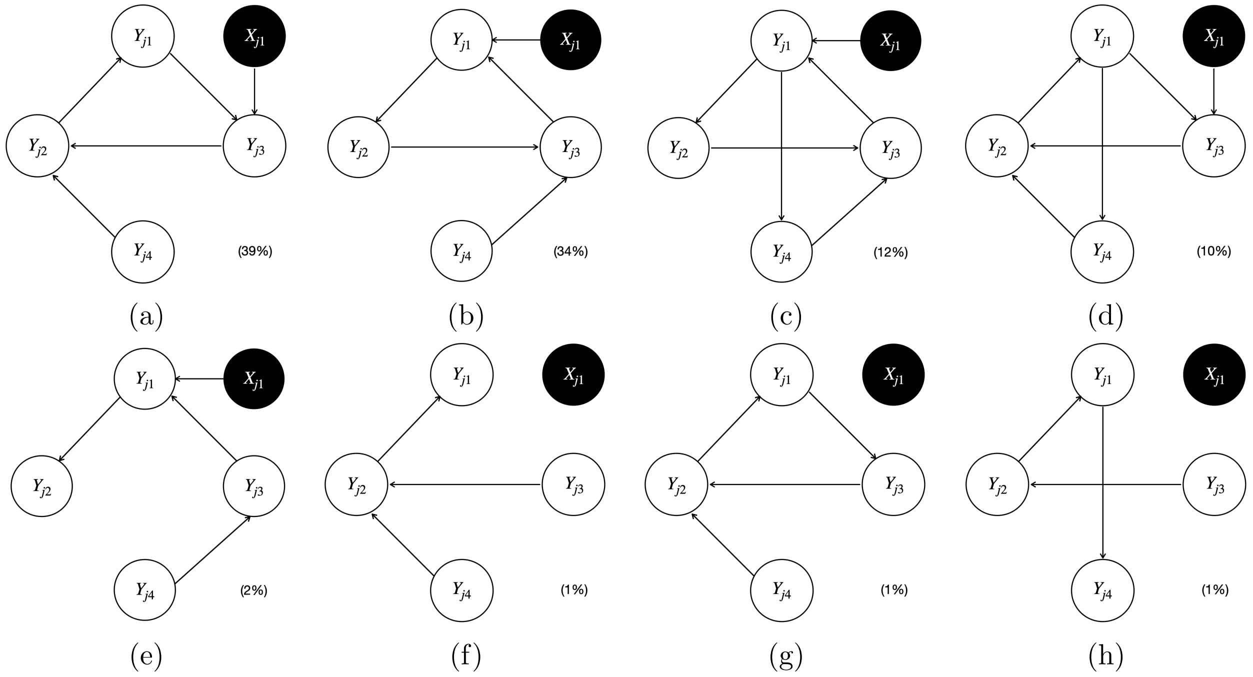


(a) (b) (c) (d)

(e) (f) (g) (h)

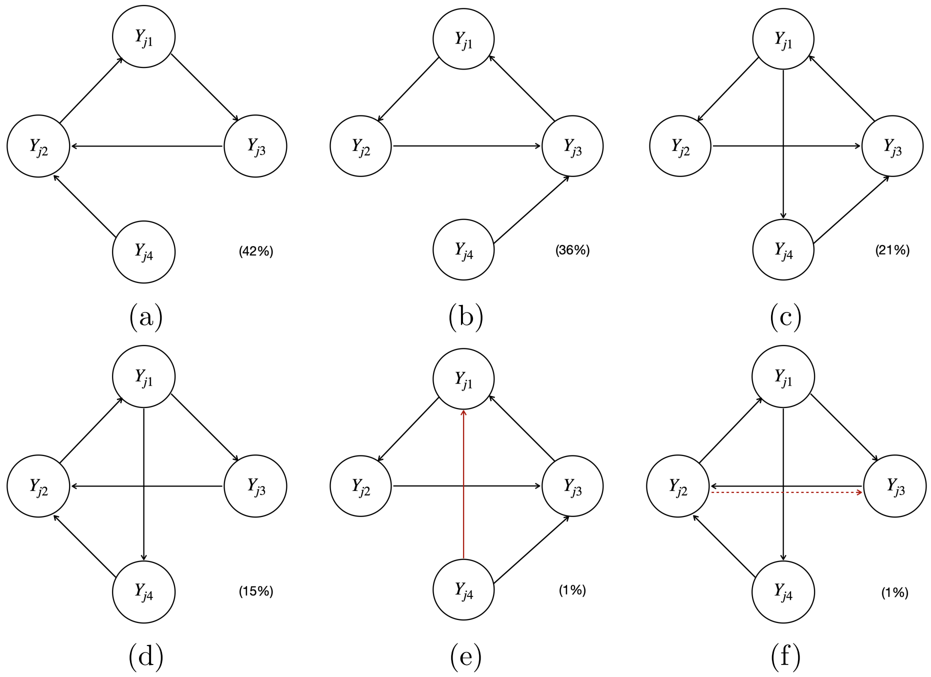
(i) (j)

Figure S9: The estimated causal graphs under VAR-LiNGAM with a sample size of 2,500, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



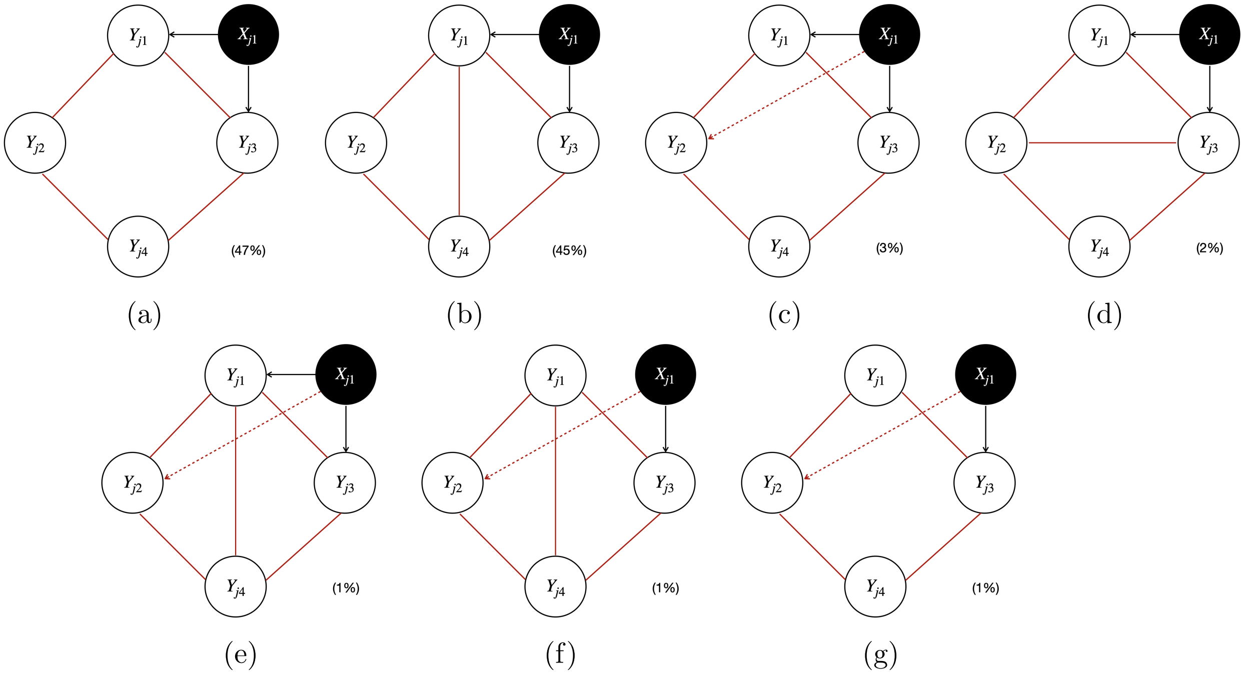
(a) (b) (c) (d)
(e) (f) (g) (h)

Figure S10: The estimated causal graphs under the proposed model with a sample size of 1,000, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c)
(d) (e) (f)

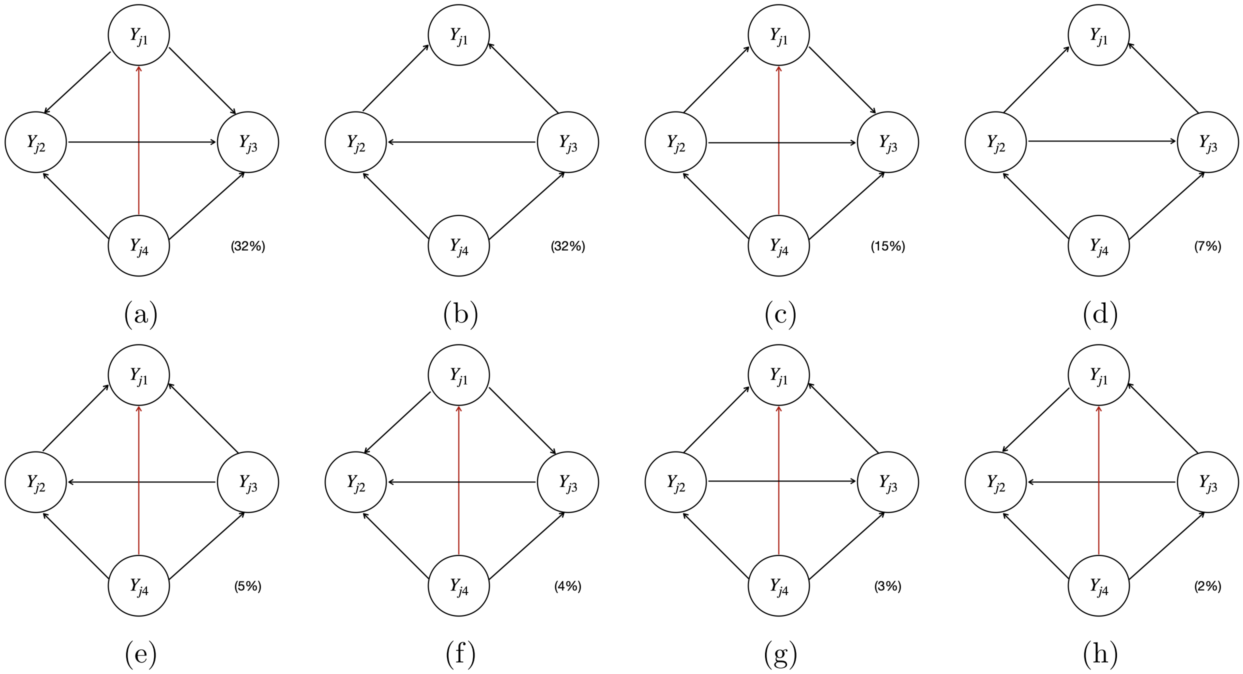
Figure S11: The estimated causal graphs under LiNG-D with a sample size of 1,000, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c) (d)

(e) (f) (g)

Figure S12: The estimated causal graphs under PCMCI$^+$ with a sample size of 1,000, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure S13: The estimated causal graphs under VAR-LiNGAM with a sample size of 1,000, using X_{j1} as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

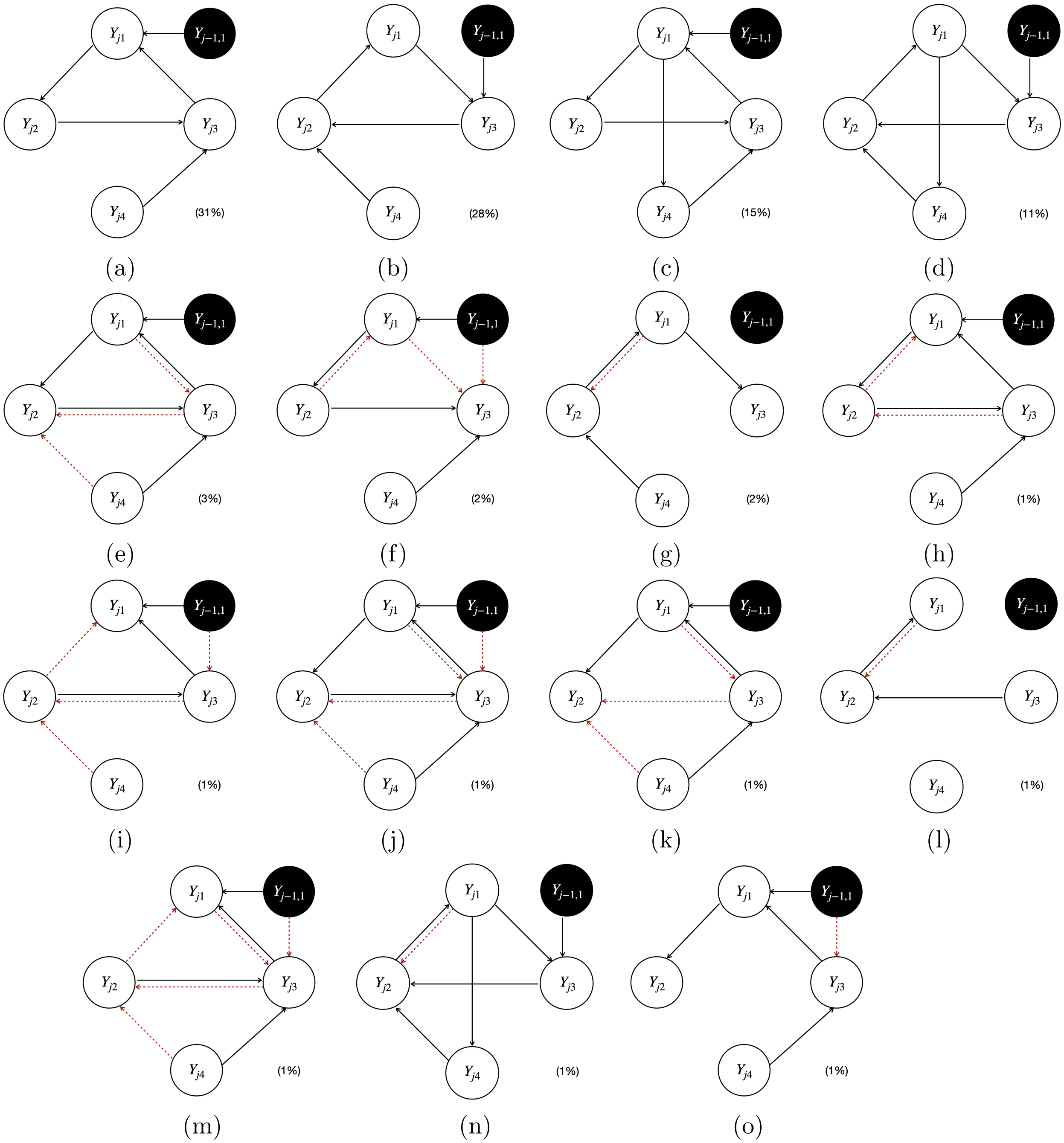


Figure S14: The estimated causal graphs under the proposed model with a sample size of 1,000, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

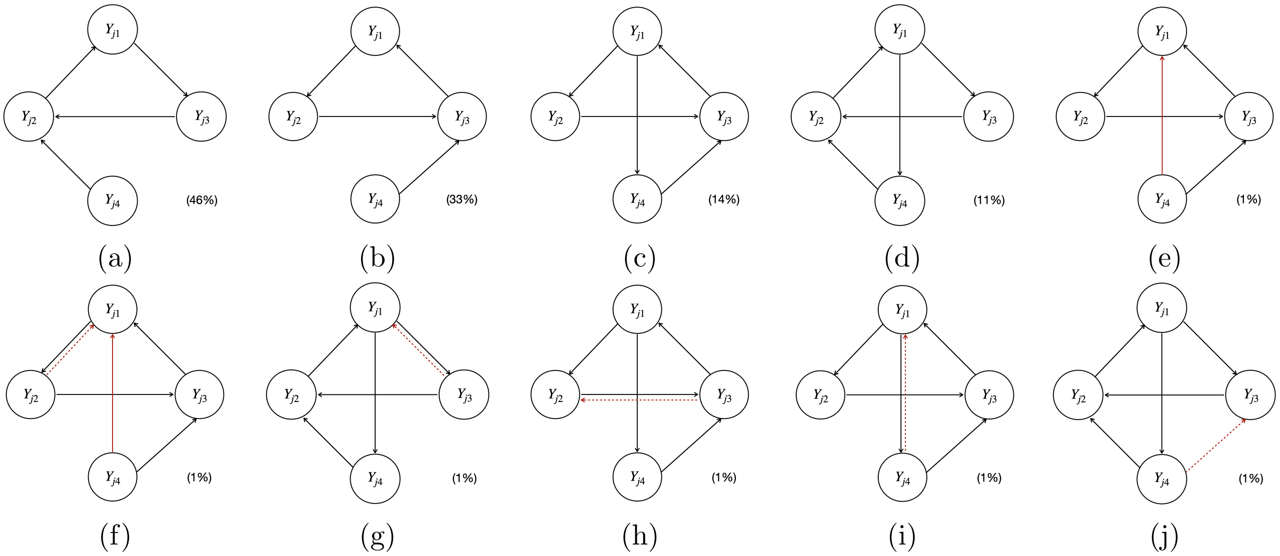


Figure S15: The estimated causal graphs under LiNG-D with a sample size of 1,000, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

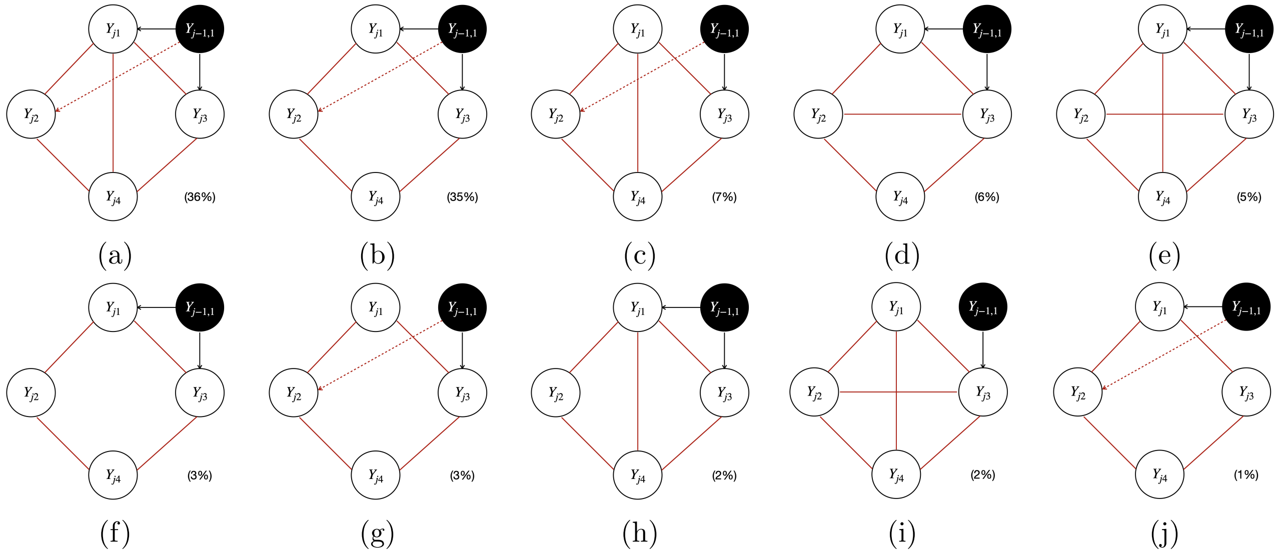
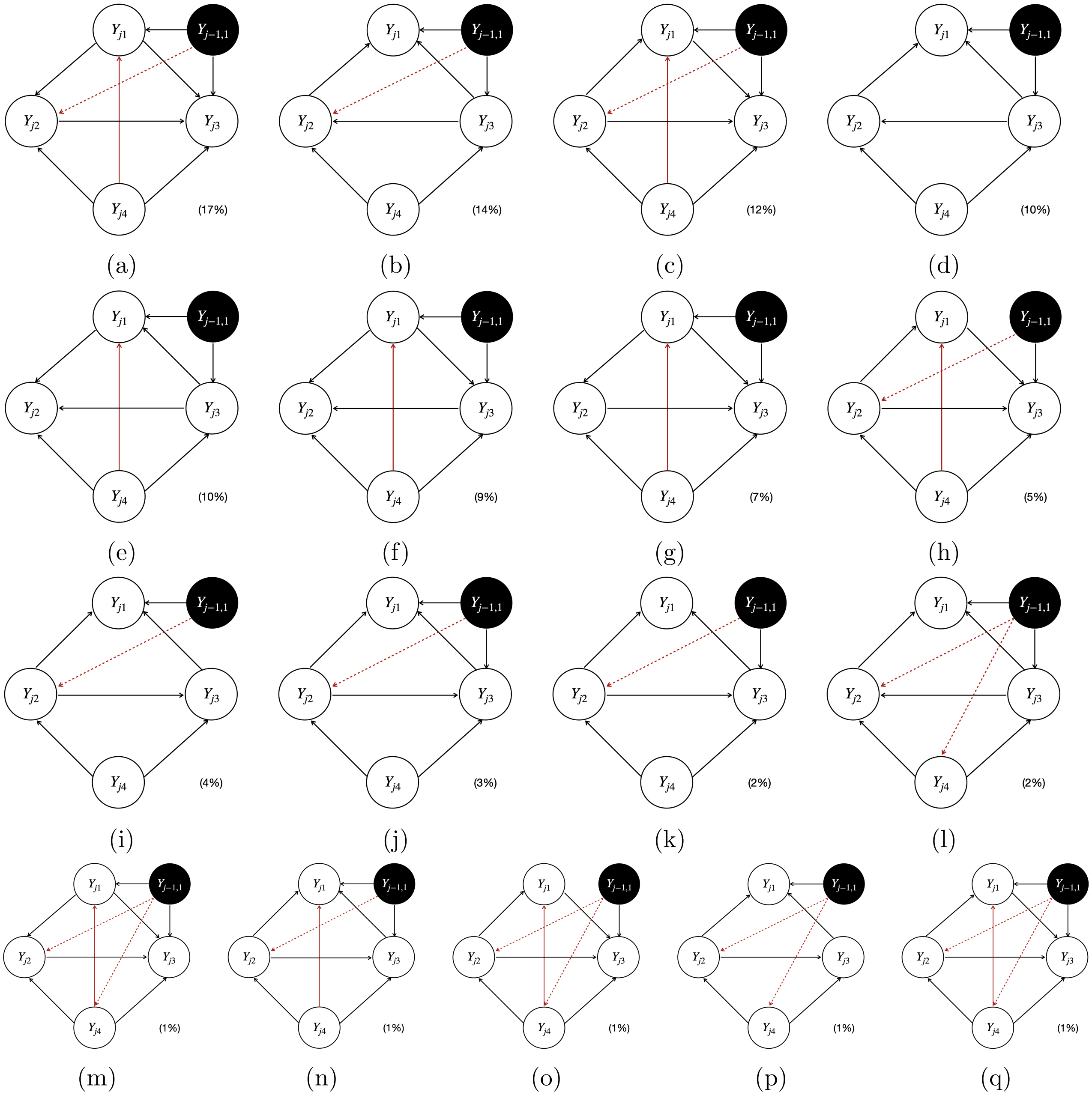
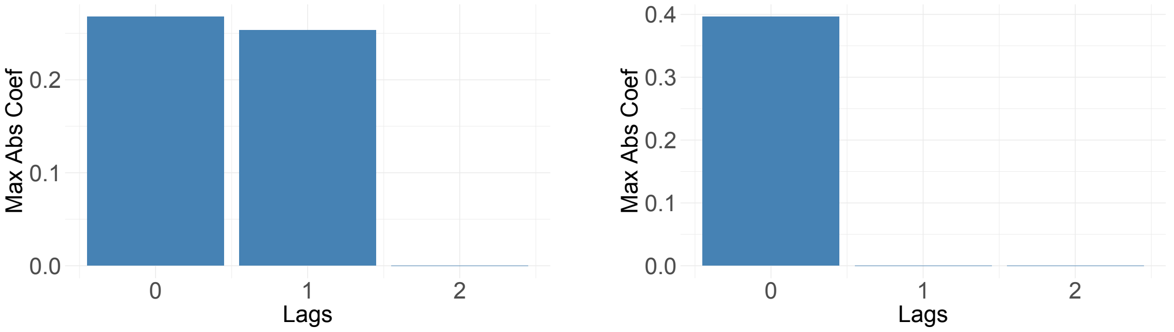


Figure S16: The estimated causal graphs under PCMCI$^+$ with a sample size of 1,000, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.

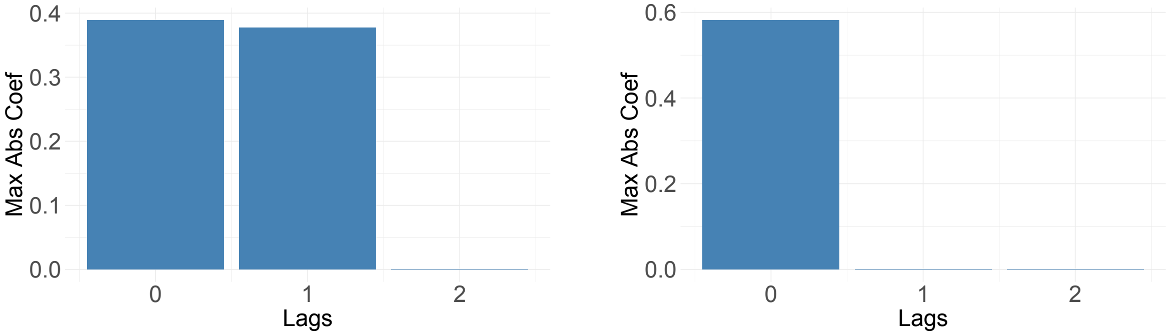


(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o) (p) (q)

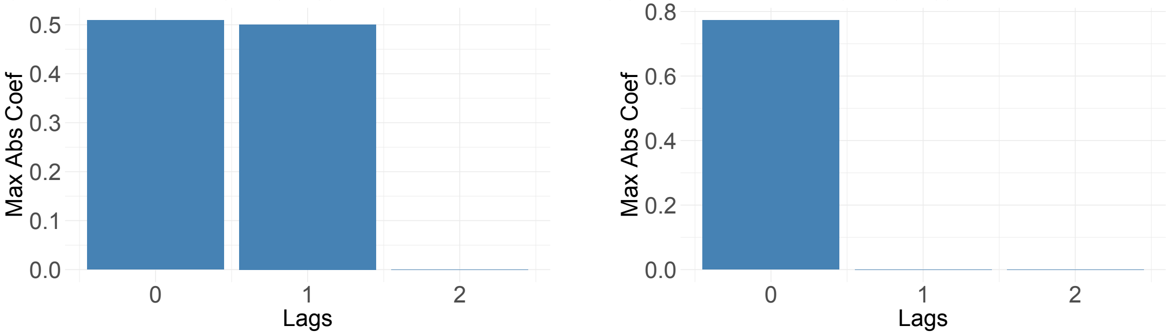
Figure S17: The estimated causal graphs under VAR-LiNGAM with a sample size of 1,000, using $Y_{j-1,1}$ as the instrumental variable $I_{Y_{j1}}$ in simulation scenario I.



(a) $\eta = 0.5 : \max(\text{abs}(\boldsymbol{B}_\ell))$ as a function of ℓ (b) $\eta = 0.5 : \max(\text{abs}(\boldsymbol{A}_\ell))$ as a function of ℓ

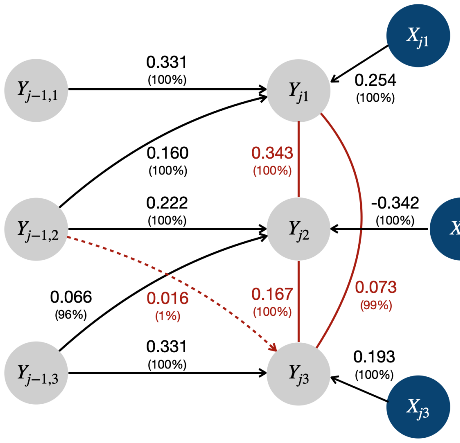


(c) $\eta = 0.75 : \max(\text{abs}(\boldsymbol{B}_\ell))$ as a function of ℓ (d) $\eta = 0.75 : \max(\text{abs}(\boldsymbol{A}_\ell))$ as a function of ℓ

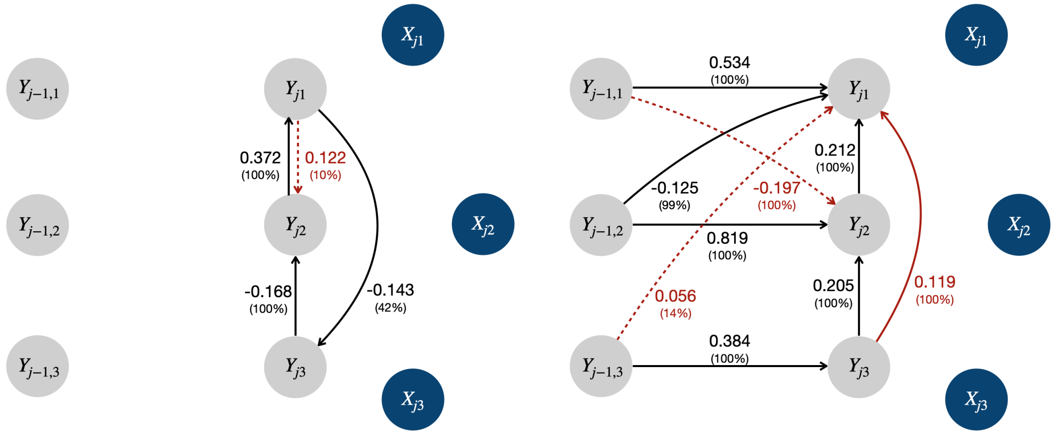


(e) $\eta = 1 : \max(\text{abs}(\boldsymbol{B}_\ell))$ as a function of ℓ (f) $\eta = 1 : \max(\text{abs}(\boldsymbol{A}_\ell))$ as a function of ℓ

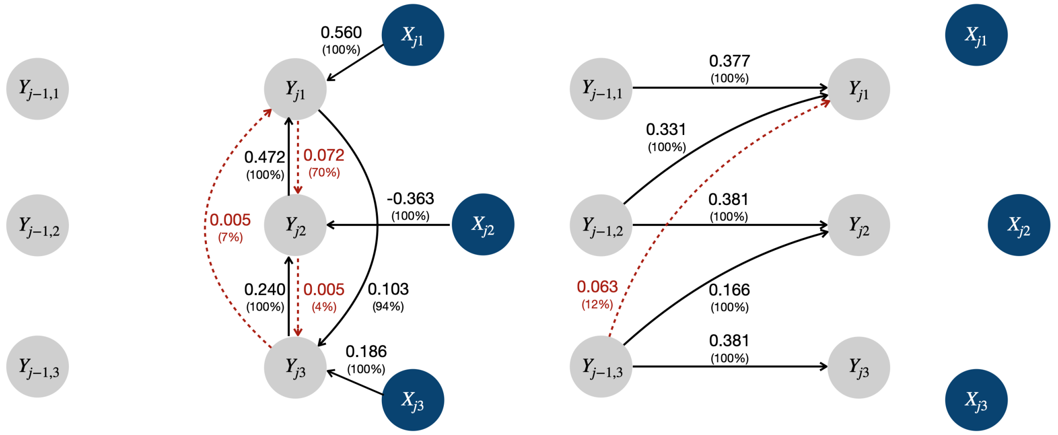
Figure S18: Largest absolute values of the estimated coefficients $\boldsymbol{B}_\ell$ and $\boldsymbol{A}_\ell$ for one randomly selected simulated dataset from 100 experiments in the simulation scenario II. The simulated true number of time lags are $L_y^{\text{truth}} = 1$ and $L_x^{\text{truth}} = 0$. To select the number of lags, note that $\max(\text{abs}(\boldsymbol{B}_\ell))$ and $\max(\text{abs}(\boldsymbol{A}_\ell))$ decrease significantly towards zero from $\ell = L_y^{\text{truth}}$ to $\ell = L_y^{\text{truth}} + 1$, and from $\ell = L_x^{\text{truth}}$ to $\ell = L_x^{\text{truth}} + 1$, respectively.



(a) $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$

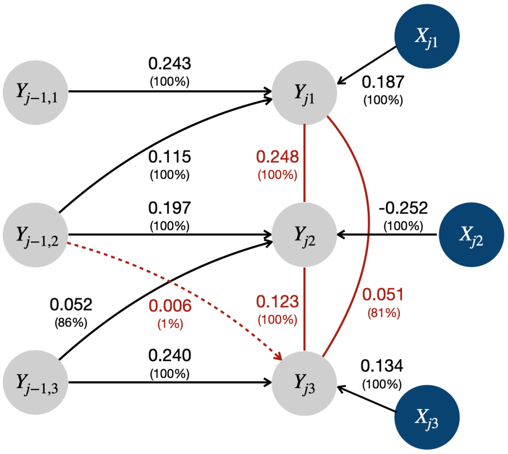


(b) $\widehat{\mathcal{G}}_{\text{LiNG-D}}$ (c) $\widehat{\mathcal{G}}_{\text{VAR-LiNGAM}}$

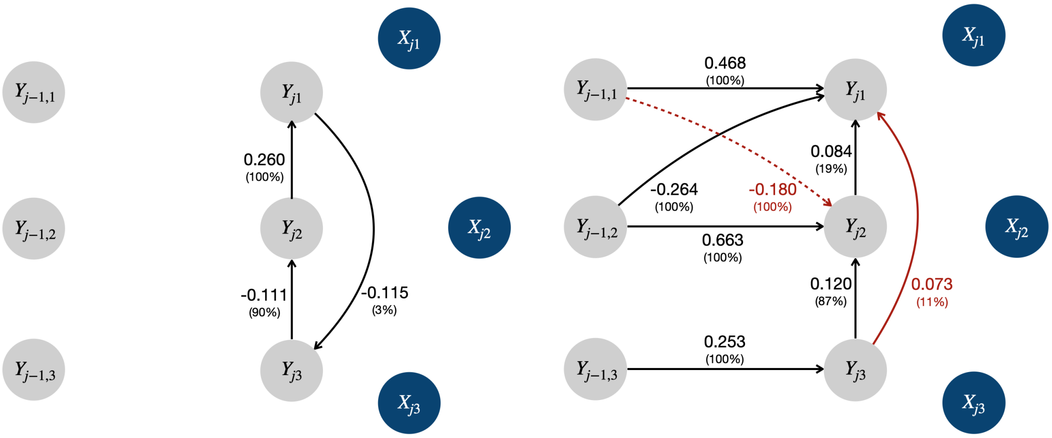


(d) $\widehat{\mathcal{G}}_{\text{SCM}}$ (e) $\widehat{\mathcal{G}}_{\text{VAR}}$

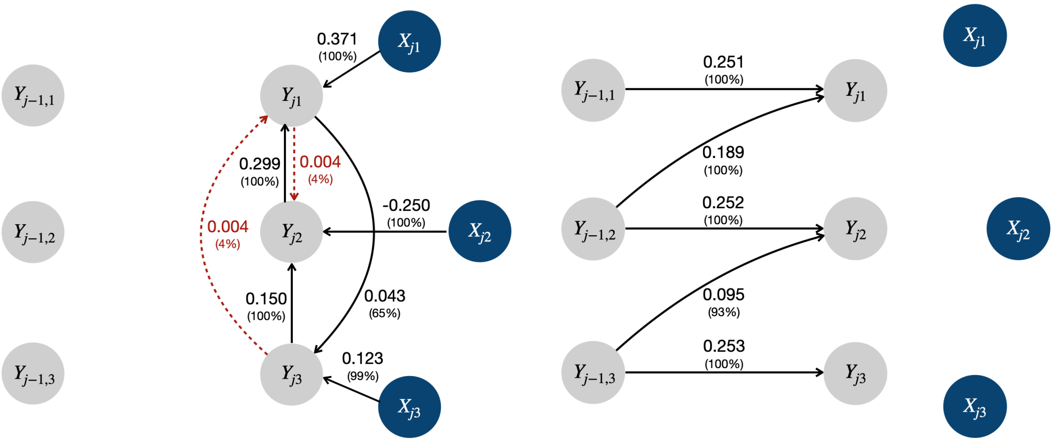
Figure S19: The estimated causal graphs under alternative methods (PCMCI$^+$, LiNG-D, VAR-LiNGAM, SCM, and VAR) in scenario II with $\eta = 0.75$.



(a) $\widehat{\mathcal{G}}_{\text{PCMCI}^+}$

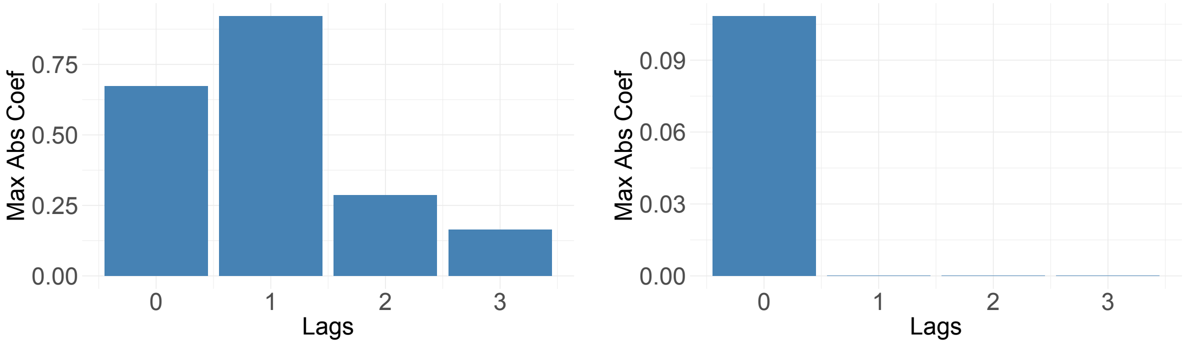


(b) $\widehat{\mathcal{G}}_{\text{LiNG-D}}$ (c) $\widehat{\mathcal{G}}_{\text{VAR-LiNGAM}}$



(d) $\widehat{\mathcal{G}}_{\text{SCM}}$ (e) $\widehat{\mathcal{G}}_{\text{VAR}}$

Figure S20: The estimated causal graphs under alternative methods (PCMCI$^+$, LiNG-D, VAR-LiNGAM, SCM, and VAR) in scenario II with $\eta = 0.5$.



(a) $\max(\text{abs}(\boldsymbol{B}_\ell))$ as a function of ℓ (b) $\max(\text{abs}(\boldsymbol{A}_\ell))$ as a function of ℓ

Figure S21: Largest absolute values of the estimated coefficients $\boldsymbol{B}_\ell$ and $\boldsymbol{A}_\ell$ in the WIHS data analysis. The number of time lags were selected to be $L_y = 1$ and $L_x = 0$, since $\max(\text{abs}(\boldsymbol{B}_\ell))$ and $\max(\text{abs}(\boldsymbol{A}_\ell))$ decrease significantly towards zero from $\ell = L_y$ to $\ell = L_y + 1$, and from $\ell = L_x$ to $\ell = L_x + 1$, respectively.

References

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-Gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34:22822–22833, 2021.

Adaora A Adimora, Catalina Ramirez, Lorie Benning, Ruth M Greenblatt, Mirjam-Colette Kempf, Phyllis C Tien, Seble G Kassaye, Kathryn Anastos, Mardge Cohen, Howard Minkoff, et al. Cohort profile: the women's interagency HIV study (WIHS). *International Journal of Epidemiology*, 47(2):393–394i, 2018.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

Cristina Brickman, Kathleen J Propert, Chelsea Voytek, David Metzger, and Robert Gross. Association between depression and condom use differs by sexual behavior group in patients with HIV. *AIDS and Behavior*, 21(6):1676–1683, 2017.

Susan T Charles and David M Almeida. Daily reports of symptoms and negative affect: Not all symptoms are the same. *Psychology and Health*, 21(1):1–17, 2006.

Li Chen, Chunlin Li, Xiaotong Shen, and Wei Pan. Discovery and inference of a causal network with hidden confounding. *Journal of the American Statistical Association*, pages 1–13, 2023.

Hee Min Choi and James P Hobert. Analysis of MCMC algorithms for Bayesian linear regression with Laplace errors. *Journal of Multivariate Analysis*, 117:32–40, 2013.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3): 287–314, 1994.

Ed Diener and Robert A Emmons. The independence of positive and negative affect. *Journal of Personality and Social Psychology*, 47(5):1105, 1984.

Todd W Dunn, Jeffrey R Vittengl, Lee Anna Clark, Thomas Carmody, Michael E Thase, and Robin B Jarrett. Change in psychosocial functioning and depressive symptoms during acute-phase cognitive therapy for depression. *Psychological Medicine*, 42(2):317–326, 2012.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using FCI. *Probabilistic Graphical Models*, pages 121–128, 2010.

Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, pages 1917–1925. PMLR, 2015.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pages 424–438, 1969.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Uncertainty in Artificial Intelligence*, pages 766–776. PMLR, 2023.

Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, 2016.

Roger A Horn and Charles R Johnson. *Matrix Analysis.* Cambridge University Press, 2012.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems*, 21, 2008a.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008b.

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

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Mamta K Jain, Xilong Li, Beverley Adams-Huet, Yordanos M Tiruneh, Amneris E Luque, Piper Duarte, Joseph M Trombello, and Ank E Nijhawan. The risk of depression among racially diverse people living with HIV: the impact of HIV viral suppression. *AIDS Care*, 33(5):645–653, 2021.

Wei Jin, Yang Ni, Leah H Rubin, Amanda B Spence, and Yanxun Xu. A Bayesian nonparametric approach for inferring drug combination effects on mental health in people with HIV. *Biometrics*, 78(3):988–1000, 2022.

Kento Kadowaki, Shohei Shimizu, and Takashi Washio. Estimation of causal structures in longitudinal data using non-Gaussianity. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2013.

Jan TA Koster. Markov properties of nonrecursive causal models. *The Annals of Statistics*, 24(5):2148–2177, 1996.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 366–374, 2008.

Markku Lanne, Mika Meitz, and Pentti Saikkonen. Identification and estimation of non-Gaussian structural vector autoregressions. *Journal of Econometrics*, 196(2):288–304, 2017.

Jinkook Lee. Pathways from education to depression. *Journal of Cross-Cultural Gerontology*, 26(2):121–135, 2011.

Wei Li, Rui Duan, and Sai Li. Discovery and inference of possibly bi-directional causal relationships with invalid instrumental variables. *arXiv preprint arXiv:2407.11646*, 2024.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pages 23–47. PMLR, 2018.

Joris M Mooij and Tom Claassen. Constraint-based causal discovery using partial ancestral graphs in the presence of cycles. In *Conference on Uncertainty in Artificial Intelligence*, pages 1159–1168. PMLR, 2020.

Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. *Advances in Neural Information Processing Systems*, 24, 2011.

Yuichi Murata, Mahito Kimura, and Robert G Robinson. Does cognitive impairment cause poststroke depression? *The American Journal of Geriatric Psychiatry*, 8(4):310–317, 2000.

Chris J Oates, Jim Q Smith, and Sach Mukherjee. Estimating causal structure using conditional DAG models. *Journal of Machine Learning Research*, 17(54):1–23, 2016.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

Luis Parra-Rodriguez and Jane A O'Halloran. HIV and obesity: updates in management strategies. *Current Opinion in HIV and AIDS*, 18(2):68–74, 2023.

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

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26, 2013.

Lenore Sawyer Radloff. The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401, 1977.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, pages 454–461, 1996.

Robert E Roberts, George A Kaplan, Sarah J Shema, and William J Strawbridge. Are the obese at greater risk for depression? *American Journal of Epidemiology*, 152(2):163–170, 2000.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR, 2020.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553, 2019a.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019b.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-Gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21(1):1436–1459, 2020.

Kemal Sayar, Laurence J Kirmayer, and Suzanne S Taillefer. Predictors of somatic symptoms in depressive disorder. *General Hospital Psychiatry*, 25(2):108–114, 2003.

James G Scott and James O Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010.

S Shaphiro and MBJB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3): 591–611, 1965.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. *Scientific Reports*, 10(1):2975, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 491–498, 1995.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000.

Norman R Swanson and Clive WJ Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92(437):357–367, 1997.

Nikolaj Thams, Rikke Søndergaard, Sebastian Weichwald, and Jonas Peters. Identifying causal effects using instrumental time series: Nuisance IV and correcting for the past. *Journal of Machine Learning Research*, 25(302):1–51, 2024.

US Department of Health & Human Services. Guidelines for the use of antiretroviral agents in adults and adolescents with HIV, 2020.

KK Vidya Vijayan, Krithika Priyadarshini Karthigeyan, Srikanth P Tripathi, and Luke Elizabeth Hanna. Pathophysiology of CD4+ T-cell depletion in HIV-1 and HIV-2 infections. *Frontiers in Immunology*, 8:580, 2017.

Jessica R Weinstein and Sharon Anderson. The aging kidney: physiological changes. *Advances in Chronic Kidney Disease*, 17(4):302–307, 2010.

Kerstyn C Zalesin, Barry A Franklin, Wendy M Miller, Eric D Peterson, and Peter A McCullough. Impact of obesity on cardiovascular disease. *Endocrinology and Metabolism Clinics of North America*, 37(3):663–684, 2008.