Discovering Cyclic Causal Models in Psychological Research

Kyuri Park[1,2*], Lourens J. Waldorp[3], and Oisín Ryan[4]

[1]Department of Methodology and Statistics, Utrecht University
[2]Computational Science Lab, Informatics Institute, University of Amsterdam
[3]Department of Psychology, University of Amsterdam
[4]Department of Data Science and Biostatistics, Julius Centre, University Medical Centre Utrecht

Abstract

Statistical network models have become popular tools for analyzing multivariate psychological data. In empirical practice, network parameters are often interpreted as reflecting causal relationships – an approach that can be characterized as a form of causal discovery. Recent research has shown that undirected network models are likely to perform poorly as causal discovery tools in the context of discovering acyclic causal structures, a task for which many alternative methods are available. However, acyclic causal models are likely unsuitable for many psychological phenomena, such as psychopathologies, which are often characterized by cycles or feedback loop relationships between symptoms. A number of cyclic causal discovery methods have been developed, largely in the computer science literature, but they are not as well studied or widely applied in empirical practice. In this paper, we provide an accessible introduction to the basics of cyclic causal discovery for empirical researchers. We examine three different cyclic causal discovery methods and investigate their performance in typical psychological research contexts by means of a simulation study. We also demonstrate the practical applicability of these methods using an empirical example and conclude the paper with a discussion of how the insights we gain from cyclic causal discovery relate to statistical network analysis.

Keywords: Constraint-based, cyclic causal discovery, directed cyclic graph (DCG), partial ancestral graph (PAG), statistical network model

*Correspondence concerning this paper should be addressed to: Kyuri Park, Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. e-mail: kyurheep@gmail.com

1 Introduction

A fundamental task in various disciplines of science is to understand the mechanisms or causal relations underlying the phenomena of interest. In psychology, one of the core questions is how psychopathology comes about, with the network theory positing that mental disorder is produced by a system of direct causal interactions between symptoms (Borsboom & Cramer, 2013). Given this theoretical framework, statistical network models have become popular tools for analyzing multivariate observational data (Robinaugh et al., 2020; Epskamp et al., 2018a). In practice, empirical researchers often interpret the conditional statistical relationships estimated in a network model as reflecting the causal relationships between symptoms of mental disorder and/or other psychological variables — an approach that can be characterized as a form of causal discovery (Spirtes et al., 2000; Peters et al., 2017; Ryan et al., 2022). However, it has been shown that network models are likely to perform poorly as causal discovery tools; relations in the network may not reflect the direct causal effects that researchers aim to discover, as they may be produced by, amongst other inferential issues, unwittingly conditioning on common effects (Dablander & Hinne, 2019; Ryan et al., 2022).

In the field of causal discovery, one class of methods that utilizes patterns of statistical (in)dependence estimated from observational data to infer causal structures is known as *constraint-based* causal discovery (Spirtes & Glymour, 1991). Ryan et al. (2022) suggest that network models could be replaced by purpose-built constraint-based causal discovery methods. However, the most popular and well-studied constraint-based methods assume that causal relationships are *acyclic*; that is if X causes Y, then Y does not cause X (Glymour et al., 2019). Although so-called Directed Acyclic Graphs (DAGs) (Pearl, 2009) are popular tools for causal modeling (Tennant et al., 2021), the acyclicity assumption is problematic when studying a phenomena such as mental disorders, since *cyclic* relationships or *feedback loops* are critical to the theoretical understanding of psychopathology (Borsboom, 2017; Haslbeck et al., 2021). For example, Wittenborn et al. (2016) suggest that several different causal feedback loops, such as *perceived stress* $\longrightarrow$ *negative affect* $\longrightarrow$ *rumination* $\longrightarrow$ *perceived stress* play a key role in sustaining depression. Such theoretical expectations necessitate the use of *cyclic causal discovery* methods.

Although some constraint-based cyclic causal discovery algorithms have been developed (Mooij & Claassen, 2020; Richardson, 1996a; Strobl, 2019), they have not been as well studied as their acyclic counterparts. In part, this is due to the conceptual and practical difficulties in fitting and interpreting cyclic causal models. Conceptually, a number of researchers in the causal modeling literature have shown that, under certain conditions, cyclic causal models fit to cross-sectional data may be interpreted as reflecting causal relations between *equilibriums* or resting states of a dynamic system (Iwasaki & Simon, 1994; Dash, 2005; Strotz & Wold, 1960; Spirtes, 1995; Mooij et al., 2013; Weinberger, 2021; Bongers et al., 2022). From this perspective, cyclic causal relations should be interpreted as a kind of coarse-grained or time-averaged representation of (reciprocal) causal re-

lations between processes that evolve over time; for a detailed treatment of cyclic equilibrium causal models in the context of psychological modeling, we refer readers to Ryan and Dablander (2022). On the practical side, in the context of structural equation modeling, it is well known that all acylic path-models (i.e. containing independent error terms and no latent variables) are statistically identified, whereas this is not generally the case for path models which contain cycles (Bongers et al., 2021; Bollen, 1989). Furthermore, interpreting the output of cyclic causal discovery algorithms is significantly more challenging than for their acyclic counterparts: Even in ideal scenarios, the same pattern of statistical dependence that can be used to deduce a direct causal link in a DAG may not reflect a direct causal link in a cyclic graph (Bongers et al., 2018; Hyttinen et al., 2013). Recent work has explored the application of *invariance-based* algorithms — another type of causal discovery methods capable of estimating cycles — to psychological data (Kossakowski et al., 2021). However, these methods require multiple datasets that measure the same variables but in different settings, such as a mix of observational and experimental data, which is a disadvantage compared to constraint-based methods that can be applied using only a single observational dataset (Peters et al., 2016; Glymour et al., 2019).

To our knowledge, little research has been done on the applicability of constraint-based cyclic causal discovery methods in psychology, and much remains unknown about their performance. Therefore, in this paper, we aim to address the following question: How well do constraint-based cyclic causal discovery methods perform in typical psychological research contexts? First, we will provide an accessible overview of the different cyclic causal discovery methods, including the assumptions under which they are expected to work and how the output of these methods should be interpreted. Second, we investigate, by means of a simulation study, how well these different methods perform under various circumstances. In the simulation study, we study more and less ideal situations, by varying the sample size, the size and density of the underlying network, and the presence or absence of unobserved confounding variables. Third, we demonstrate the practical applicability of these methods by applying them to empirical data (McNally et al., 2017) and discussing how the insights we gain relate to the statistical network analysis of the same data.

2 Background

In this section, we will establish the basic concepts of graphical models and examine how constraint-based causal discovery methods operate. We will first introduce different types of graphical models, while demonstrating the differences between statistical and causal graphical models using example graphs. Then, we will explain the assumptions that underlie constraint-based causal discovery methods, as well as the technical difficulties that arise in the presence of cycles. Lastly, we will illustrate each step of the constraint-based causal discovery procedure using a simple example of a directed graph.

2.1 Graphical Models

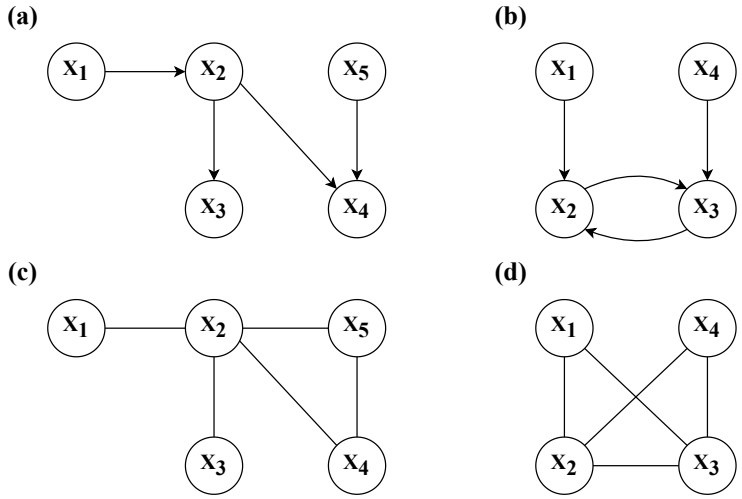
A graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ is a set of edges, describing the connections between the vertices. A probabilistic graphical model uses a graph to express the conditional (in)dependencies between random variables, where the vertices represent random variables, and the edges encode the conditional dependencies that hold among the set of variables (Lauritzen, 1996).

There exist various graphical models that differ in the types of edges they allow (e.g., directed vs. undirected) and how edges are interpreted in terms of statistical relationships. One commonly known graphical model in psychology is the *Pairwise Markov Random Field* (PMRF) — an undirected graph where edges indicate a statistical association between variables after controlling for all other variables — which forms the basis of statistical network models such as the Gaussian Graphical Model (GGM) (Epskamp et al., 2018b; Epskamp & Fried, 2018). In PMRFs, edges are strictly undirected, and the presence of an edge $A - B$ indicates that A and B are statistically dependent, conditional on the set of all other variables in the network (Borsboom et al., 2021). Models like the GGM and Ising network model can be seen as parameterizations of the PMRF. These models assume particular distributions for the variables involved, and, in the case of the GGM, assume that conditional (in)dependence relations can be captured by linear dependence parameters such as the partial correlation.

In a *causal* graphical model, on the other hand, the edges describe *causal* relationships between variables; the edges are typically directed, with $A \longrightarrow B$ denoting that intervening on A results in a change in the probability distribution of B (Geiger & Pearl, 1990). The simplest and most commonly used causal graph is a *directed acyclic graph* (DAG), which is commonly utilized to represent a *Bayesian network*. A DAG consists of directed edges without any cycles (Pearl, 1988). When a causal graph contains cycles, it is referred to as a *directed cyclic graph* (DCG). Two examples of causal graphical models are shown in Figure 1. Figure 1(a) does not contain any cycles, whereas Figure 1(b) does; hence they are called a DAG and DCG, respectively. In Figure 1(b), the cycle $X_2 \rightleftarrows X_3$ denotes that X_2 is a direct cause of X_3 and X_3 is a direct cause of X_2.

Causal graphical models also describe patterns of statistical independencies, which can be read off from structure of the graph using Pearl's *d-separation criterion* (Geiger et al., 1990). The idea of this criterion is to associate *dependence* with *connectedness* (i.e., the path between A and B is activated by C; *d-connected* by C) and *independence* with *separation* (i.e., the path between A and B is blocked by C, and so these variables are *d-separated* by C; for a detailed formal treatment of d-separation, see Tian et al., 1998). Formally, two variables A and B are said to be d-separated given C if and only if all paths between A and B are *blocked* when conditioning on C (Pearl, 2009). Different types of directed paths in a graph are either blocked or unblocked by conditioning on variables along them. For instance, in Figure 1(a), we see a *chain* structure $X_1 \longrightarrow X_2 \longrightarrow X_3$, which implies that X_1 and X_3 are marginally dependent ($X_1 \not\perp\!\!\!\perp X_3$) but independent conditional on X_2

Figure 1. Example causal graphical models and corresponding PMRF models.



Note. (a) is the example directed acyclic graph (DAG). (b) is the example directed cyclic graph (DCG). (c) is the PMRF (Pairwise Markov Random Field) corresponding to the DAG in (a). (d) is the PMRF corresponding to the DCG in (b).

($X_1 \perp\!\!\!\perp X_3 \mid X_2$). More formally, we would say X_1 and X_3 are *d-connected* given the empty set but *d-separated* by X_2. A *fork* structure $X_3 \leftarrow X_2 \rightarrow X_4$ implies the same pattern of independencies; X_3 and X_4 are marginally dependent ($X_3 \not\perp\!\!\!\perp X_4$) but independent conditional on X_2 ($X_3 \perp\!\!\!\perp X_4 \mid X_2$). However, a *collider* structure $X_2 \rightarrow X_4 \leftarrow X_5$ implies a contrasting pattern; here X_2 and X_5 are marginally independent (i.e. d-separated when conditioning on the empty set, $X_2 \perp\!\!\!\perp X_5$) but dependent conditional on X_4 ($X_2 \not\perp\!\!\!\perp X_5 \mid X_4$). This distinguishing characteristic of colliders is crucial when identifying the directions of causal relationships, as will be shown in Section 2.3. In principle, the d-separation criterion can be applied to both acyclic and cyclic causal graphs, as long as certain conditions, discussed in Section 2.2, are met.

Having established the basics, we can now examine how PMRF-based statistical network models relate to different types of directed causal models. In Figure 1(c), we show the PMRF model that corresponds to the DAG in Figure 1(a), where an additional edge is introduced between $X_2 - X_5$ due to conditioning on the common effect (i.e., collider), X_4. In Figure 1(d), the PMRF model corresponding to the DCG in Figure 1(b) is shown. Two spurious edges are induced in the PMRF network (e.g., $X_1 - X_3$ and $X_2 - X_4$) because of conditioning on the colliders X_2 and X_3. These examples illustrate the limitations of statistical network models in inferring patterns of directed causal relationships. PMRFs can contain spurious edges resulting from conditioning on common effects, and the possibility of producing collider structures is likely to be higher with the presence of cycles, exacerbating this issue. Notably, while the mapping from a causal graph to the statistical network we show here is unique, the two statistical networks presented may have been generated by various

distinct causal graphs, including those with or without cycles. For more details on the relationship between PMRF-based networks and causal graphs, we refer readers to Ryan et al. (2022).

Despite this limitation, in practice, PMRFs have often been interpreted as a *causal skeleton* — the undirected version of a causal graph (Haslbeck & Waldorp, 2018). Further elaborating on this, Ryan et al. (2022) demonstrated how PMRF-based network models can, in fact, be used to identify a so-called *equivalence class* of causal graphs under certain assumptions. However, these models are prone to suboptimal performance, as the equivalence class they identify is likely much larger than that of custom-built causal discovery methods. Consequently, the authors suggest that causal discovery methods specifically designed for this task are likely to outperform statistical network models in learning the underlying causal structure, indicating that network models may not be desirable tools for discovering causal relationships. In the following sections, we will focus on how constraint-based causal discovery methods recover the causal structure while looking into the assumptions they require. Additionally, we will explore the practical and conceptual difficulties involved in performing causal discovery in the presence of causal cycles.

2.2 Acyclic vs. Cyclic Causal Graphs

The d-separation criterion described above applies to all acyclic graphs, but for graphs with cycles, it applies only under certain conditions. To understand these conditions, we first need to introduce some graph terminology. In the field of graphical models, we use kinship terminology to describe a graph structure, as follows:

$$\text{if } \begin{Bmatrix} A \to B \\ A \leftarrow B \\ A \to \cdots \to B \text{ or } A = B \\ A \leftarrow \cdots \leftarrow B \text{ or } A = B \end{Bmatrix} \text{ in } \mathcal{G} \text{ then } A \text{ is a } \begin{Bmatrix} \text{parent} \\ \text{child} \\ \text{ancestor} \\ \text{descendant} \end{Bmatrix} \text{ of } B \text{ and } \begin{Bmatrix} A \in pa_{\mathcal{G}}(B) \\ A \in ch_{\mathcal{G}}(B) \\ A \in an_{\mathcal{G}}(B) \\ A \in de_{\mathcal{G}}(B) \end{Bmatrix}.$$

Also, when there exists an edge between two vertices $A - B$, A and B are said to be *adjacent*. For example, in Figure 1(b), $X_1 \in pa_{\mathcal{G}}(X_2)$, $X_2 \in ch_{\mathcal{G}}(X_1)$, $\{X_1, X_2, X_3, X_4\} \in an_{\mathcal{G}}(X_3)$, $\{X_1, X_2, X_3\} \in de_{\mathcal{G}}(X_1)$, and X_2 is adjacent to X_1 and X_3. With this terminology in place, we can define the conditions that relate patterns of causal dependency in a causal graph to patterns of statistical dependency between random variables. First, we introduce the *global Markov* condition, which states that d-separation relations represented in causal graphs can be used to read off statistical independence relations such that:

$$\text{if } X_A \perp\!\!\!\perp_{\mathcal{G}} X_B \mid X_C \implies X_A \perp\!\!\!\perp X_B \mid X_C \text{ for all disjoint subsets of } X_A, X_B, X_C,$$

where $\perp\!\!\!\perp_{\mathcal{G}}$ refers to d-separation in graphs, and $\perp\!\!\!\perp$ refers to statistical independence between random variables. If causal graphs are *acyclic*, such as DAGs, then the *global Markov* condition holds

regardless of the functional forms of causal relations and the distributions of variables involved (Lauritzen, 1996). In addition, in DAGs, the *global Markov* condition entails the *local Markov* condition, which states that a variable is independent of its non-descendants given its parents (Lauritzen, 2001). The fact that one Markov property implies the other comes in handy when reading off conditional independencies from a graph.

In contrast to the acyclic case, the situation is not as straightforward in *cyclic* graphs. In DCGs, the global Markov property does not always hold. Spirtes (1994) showed that this property does hold when causal relations are *linear* and the error terms are *independent*. However, even in this case, the global Markov property does not imply the local Markov property. For example, in Figure 1(b), the global Markov property is preserved ($X_1 \perp\!\!\!\perp_{\mathcal{G}} X_4 \mid \{X_2, X_3\} \implies X_1 \perp\!\!\!\perp X_4 \mid \{X_2, X_3\}$), but the local Markov property is violated as $X_2 \not\perp\!\!\!\perp_{\mathcal{G}} X_4 \mid X_3$ (i.e., X_2 is *not* independent of its non-descendant X_4 given its parent X_3). This is because X_3 serves as both a parent of X_2 and a collider on the path $X_2 \rightarrow X_3 \leftarrow X_4$ at the same time. In the current paper, we limit the scope of our study to cyclic causal graphs that represent *linear* causal relations with jointly *independent* error terms, so for which the global Markov condition is satisfied. This type of assumption is common in many statistical modeling traditions in psychology and social sciences. For example, structural equation models and popular statistical network models, such as the GGM, also rely on similar assumptions about the linearity of statistical relations (Epskamp et al., 2018b; Bollen & Long, 1993).

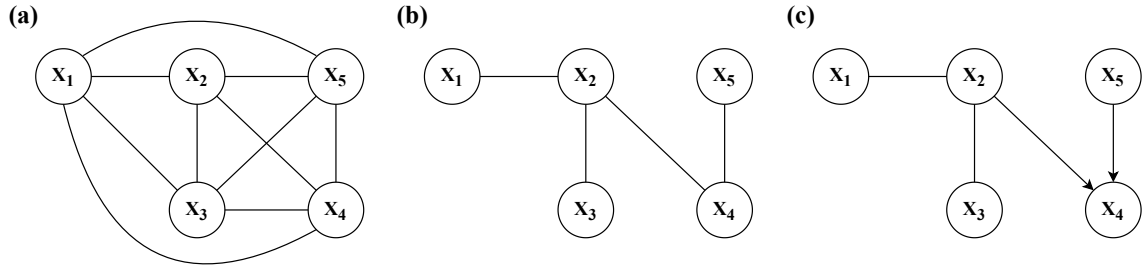
In addition to the above, constraint-based causal discovery methods typically make use of two additional assumptions (Pearl, 2009; Spirtes et al., 2000). The first is known as as *faithfulness*, which is essentially the reverse of the global Markov condition, stating that statistical independencies map onto the structure of causal graphs such that:

$$X_A \perp\!\!\!\perp X_B \mid X_C \implies X_A \perp\!\!\!\perp_{\mathcal{G}} X_B \mid X_C.$$

In other words, it postulates that if two variables are (conditionally) statistically independent of each other, then this implies that they are causally independent of each other, ruling out, for instance, that different paths in a causal system exactly cancel one another out.[1] The second assumption, known as *causal sufficiency*, relates to the absence of *unobserved* (i.e. *latent*) confounding variables — that is, any unobserved common causes of the variables within the causal graph. This assumption ensures that the statistical dependence between two variables can be explained by the patterns of causal dependence among the observed variables. Without this assumption, unobserved confounders can induce an edge between variables when there is no direct causal relation between them (Lauritzen, 1996; Spirtes et al., 2000). Crucially, not all causal discovery algorithms require the assumption of causal sufficiency; we will revisit a discussion of these methods in Section 3, but in the example below we assume sufficiency holds.

[1] For a discussion of this assumption in the context of psychological network analysis, see Ryan et al. (2022).

Figure 2. Steps of a constraint-based method.



Note. (a) shows the fully-connected graph for the example DAG from Figure 1(a), which is the initial starting point of the algorithm. (b) shows the estimated *skeleton* — an undirected graph of the underlying causal structure — after the first step. (c) shows the resulting graph after the second step, which represents the *Markov equivalence class* of DAGs (i.e., a set of DAGs that entail the same set of conditional independencies).

2.3 A Primer on Constraint-Based Causal Discovery

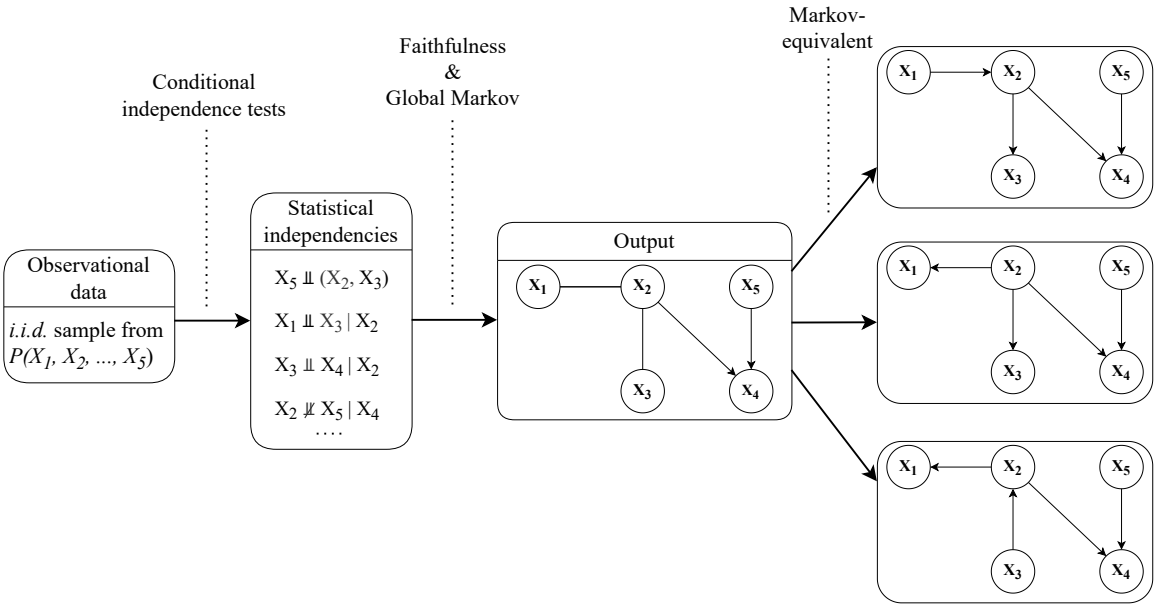
Under the aforementioned assumptions, constraint-based methods seek to recover the underlying causal structure by testing for conditional independence relations between variables from observational data (Scheines et al., 1998; Peters et al., 2017; Pearl, 1988). Assuming linear relations with additive Gaussian errors, conditional independence can be tested using partial correlations (Lawrance, 1976), although notably a number of non-parametric conditional independence tests can also be used under less stringent assumptions (Li & Fan, 2020; Huang et al., 2016). Constraint-based methods typically employ a two-step procedure; first, establishing the *skeleton* — an undirected version of the underlying graph — and second, attempting to assign directions to the edges. In general, constraint-based techniques, much like any methods relying on observational datasets, are unable to uniquely identify the underlying causal graph, but instead return a set of causal graphs that imply the same statistical independence relations (Spirtes et al., 2000).

To develop an intuition for how constraint-based methods work, we will examine how they operate on data generated by a system for which the causal graph is represented by the relatively simple DAG shown in Figure 1(a). The method begins with a fully-connected graph, as shown in Figure 2(a). In the first step, the *skeleton* is estimated by testing for conditional independence; if two variables are independent when conditioning on *any* subset of the remaining variables (e.g., $X_1 \perp\!\!\!\perp X_3 \mid X_2$, $X_1 \perp\!\!\!\perp X_4 \mid X_2$, $X_1 \perp\!\!\!\perp X_3 \mid \{X_2, X_4\}$, *etc.*), then the edge between those two variables is removed (see Figure 2(b)). This is based on the principle that, in acyclic graphs, two variables are always statistically dependent (regardless of any conditioning set), if and only if a direct causal relation exists between them. In other words, we can identify the presence of direct causal paths between variables by testing whether they are statistically dependent given any subset of the remaining variables. However, while this principle allows us to detect the presence of an edge, it does

not tell us the direction of that edge (e.g., $X_2 \rightarrow X_4$ or $X_2 \leftarrow X_4$). In the second step, we attempt to orient the edges in the skeleton by searching for *colliders* that induce distinctive patterns of independencies (e.g., X_4 is identified as a collider since $X_2 \perp\!\!\!\perp X_5$ and $X_2 \not\perp\!\!\!\perp X_5 \mid X_4$, thus $X_2 \rightarrow X_4 \leftarrow X_5$ is oriented; see Figure 2(c)). This procedure we have described is essentially the PC algorithm (Spirtes et al., 2000), and the output of the PC algorithm (Figure 2(c)) is called a *complete partially directed acyclic graph* (CPDAG). Note that the resulting CPDAG is not identical to the original true graph $\mathcal{G}$, as the two edges between $X_1 - X_2$ and $X_2 - X_3$ remain undirected. There are, in fact, three DAGs implied by the CPDAG, which are obtained by assigning directions to the undirected edges, excluding the combinations that introduce new colliders. These DAGs are called *Markov equivalent*, meaning that they encode the same conditional (in)dependencies (i.e., the same d-separation relations hold), and we call such a set of equivalent graphs a *Markov equivalence class*, denoted by $Equiv(\mathcal{G})$ (Spirtes et al., 2000). These Markov equivalent DAGs are shown in the right-hand side of Figure 3, which summarizes the constraint-based causal discovery procedure that we just described. This highlights a general difficulty in constraint-based causal discovery that relies solely on observational data, namely that there are typically multiple graphs that are consistent with an observed set of statistical independencies. Notice that the Markov equivalence class contains both the true graph shown in Figure 1(a) as well as two other distinct causal graphs; the algorithm is correct, in that the true graph is captured in this equivalence class, however, we cannot distinguish it from the other, equally plausible, members of the equivalence class.

Constraint-based methods for cyclic causal discovery operate under similar principles as those described earlier. However, cyclic causal discovery is in general more challenging, and the problem of having multiple graphs that are Markov equivalent is often exacerbated when cycles are allowed (Richardson & Spirtes, 1999). Consider how, in the DAG example above, we could identify the presence of direct causal relations between variables when they are statistically dependent given any subset of the remaining variables. Now suppose we apply the same rule to the DCG shown in Figure 1(b). In this cyclic graph, there is no direct causal path between X_1 and X_3, but instead there is the path $X_1 \rightarrow X_2 \leftrightarrows X_3$. When testing for conditional independencies, we find that X_1 and X_3 are not marginally independent because of the causal chain $X_1 \rightarrow X_2 \rightarrow X_3$. However, unlike in the acyclic case, we also find that X_1 and X_3 are not conditionally independent given X_2, since X_2 also acts as a collider on the path $X_1 \rightarrow X_2 \leftarrow X_3$. In cyclic graphs, two variables can be statistically dependent conditional on every subset of the remaining variables, even when there is no direct causal relation between them. Therefore, when a cycle exists, a constraint-based method often cannot directly identify parental relations but recovers only up to *ancestral* relations, typically leading to a larger Markov equivalence class. This also means that the estimated skeleton of a cyclic graph represents ancestral relations, and is thus called the *ancestral skeleton*. Although the same principles of causal discovery that we described for DAGs can be adapted and used for DCGs, some additional constraints and orientation rules are required to address the complexities arising from the presence of cycles. In the next section, we will further elaborate on these constraints and rules while introducing

Figure 3. Summary of the constraint-based causal discovery procedure.



Note. A constraint-based algorithm starts with performing a series of conditional independence tests on observational (*i.i.d.*: independent and identically distributed) data. Under the faithfulness and global Markov assumption, the algorithm estimates a graph structure based on the observed statistical independence patterns. The output is a *partially directed* graph (as some edges remain undirected). It can represent multiple graphs that are *Markov equivalent*, meaning that they imply the same statistical independence relations. This set of equivalent graphs is referred to as a *Markov equivalence class*, and in this example, it consists of three different DAGs.

several constraint-based cyclic causal discovery algorithms.

3 Causal Discovery Algorithms

In the previous section, we introduced the key concepts of graphical models and the fundamental principles of constraint-based causal discovery methods. In the following section, using the key concepts of constraint-based causal discovery methods introduced above, we provide a detailed description of three different constraint-based algorithms for cyclic graphs: *cyclic causal discovery* (CCD) (Richardson, 1996b), *fast causal inference* (FCI) (Spirtes et al., 1995), and *cyclic causal inference* (CCI) (Strobl, 2019). We will discuss their assumptions, steps involved, and output graphs along with their interpretation.

3.1 Assumptions of Algorithms

All three algorithms build upon the same principles and hence can be seen as extensions of the PC algorithm described in Section 2.3, but they entail slightly different assumptions. The CCD algorithm assumes *causal sufficiency*, described in Section 2.2. The other two algorithms, FCI and CCI, relax this sufficiency assumption and account for the possibility of latent confounders. In practice, this means that the output of these algorithms will often be more conservative than that of the CCD or PC algorithm, as statistical dependence between two variables, conditional on all other possible subsets of observed variables, may be induced by the presence of an unobserved confounder. However, similar to how the PC algorithm can use collider structures to determine the direction of causal relations, these algorithms can sometimes use particular patterns of multivariate dependencies to identify that some statistical dependence relations must be induced by direct or ancestral causal relations; for more detail on the general principles of causal discovery without sufficiency, we refer readers to Spirtes et al. (2000). That these algorithms do not rely on causal sufficiency makes them potentially more promising for psychological research, where the assumption of unobserved confounding is rarely warranted (Rohrer, 2018).

Another closely related concept to sufficiency is *selection bias*. Selection bias occurs when one conditions on an unobserved collider, for example, by selectively excluding a particular subgroup of samples, which leads to a similar problem of inducing spurious causal relations (Versteeg et al., 2022; Haslbeck et al., 2022). While the CCD algorithm assumes no selection bias, the FCI and CCI algorithms account for the possibility of selection bias. Although the FCI algorithm was initially designed for acyclic causal structures (Spirtes et al., 1995), it has been shown to perform well in cyclic settings under a more generalized Markov condition, while ruling out the presence of selection bias (Mooij & Claassen, 2020). Thus, we consider FCI as one of the cyclic causal discovery algorithms, but note that the suggested conditions by Mooij and Claassen (2020) hold under limited situations excluding linear and discrete cases. Table 1 summarizes the set of assumptions made by each of the algorithms, including the fundamental assumptions of global Markov condition, faithfulness, and acyclicity, as well as those related to the functional forms of causal relations and error terms.

In Section 2.3, we learned that constraint-based causal discovery works by testing for patterns of conditional (in)dependence, resulting in an equivalence class of causal graphs that convey the same statistical (in)dependencies. The PC algorithm was used to illustrate this process, which involves two steps and yields a CPDAG as the output. For cyclic causal graphs, constraint-based algorithms follow a similar approach, but with some caveats. First, the specific steps taken are more complex than those of the PC algorithm. Second, the output of these algorithms is not a CPDAG, but a different representation of equivalent cyclic graphs known as a *partial ancestral graph* (PAG).

Table 1. Assumptions of cyclic causal discovery algorithms.

	CCD	FCI	CCI
Global Markov condition	✓	✓	✓
Faithfulness	✓	✓	✓
Acyclicity	✗	–[a]	✗
Causal sufficiency	✓	✗	✗
Absence of selection bias	✓	–[a]	✗
Linearity	✓	–[a]	✓

Note. [a] FCI was originally designed to infer causal structure in the presence of selection bias assuming acyclicity, but a recent study has proposed that it performs comparably well in the cyclic settings under certain conditions (Mooij & Claassen, 2020). Specifically, these conditions require that selection bias is *absent* and variables share *non-linear* relations.

3.2 CCD Algorithm

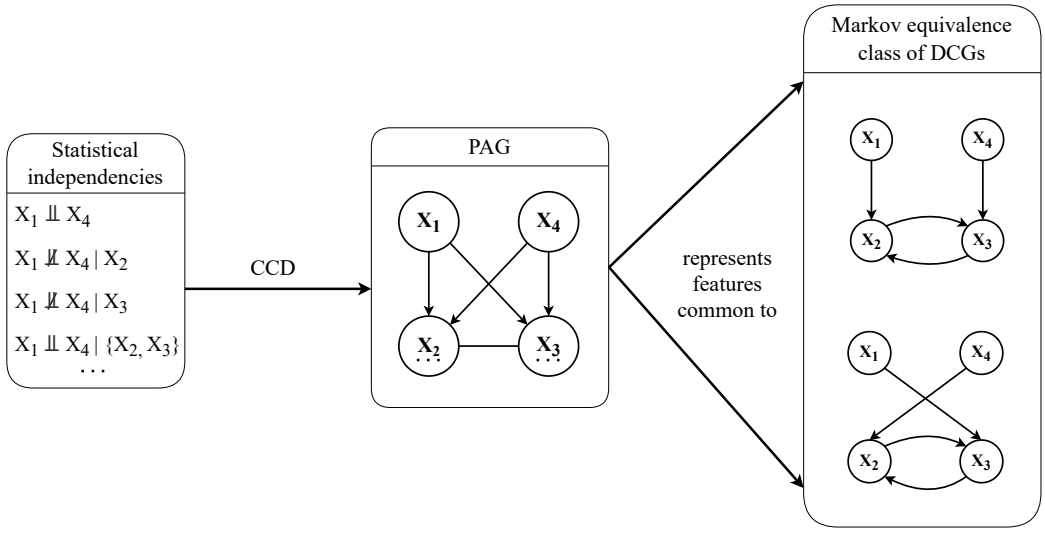
The CCD algorithm is considered relatively simple among the three algorithms, as it assumes that there is no unobserved latent confounding (i.e., *causal sufficiency*). The basic operation of CCD is summarized in Figure 4. The fundamental principles on which the CCD algorithm operates are similar to those of the PC algorithm, as illustrated in Figure 3. However, the output graph for CCD is a PAG, which represents the common features of equivalent directed cyclic graphs (DCGs). In what follows, we will explain how to interpret a PAG using the example PAG shown in Figure 4.

3.2.1 CCD Output Representation: Partial Ancestral Graph (PAG)

As was the case for DAGs, there typically exist multiple DCGs that imply the same statistical independencies and so are statistically indistinguishable from one another. To represent a set of equivalent DCGs, the CCD algorithm uses a PAG that characterizes the common features shared by all equivalent DCGs, $Equiv(\mathcal{G})$. As discussed in Section 2.3, the causal semantics of edges in PAGs become more complicated due to the presence of cyclic relations. In a CPDAG, an edge represents a direct causal relation between the corresponding vertices, while no edge implies its absence. In a PAG, the absence of an edge still indicates the absence of a direct causal relation, but the presence of an edge indicates causal *ancestry*, with $A \longrightarrow B$ meaning that A is an *ancestor* of B. In the causal graphs we have looked at so far, the types of edges are limited to directed ($\longrightarrow$) and undirected edges (—). In PAGs, however, three different types of edge-endpoints $\{\circ, >, -\}$ are utilized to represent the ancestral relations in $Equiv(\mathcal{G})$. The interpretation of each edge-endpoint in a PAG is as follows:[2]

[2]In the description of the semantics for PAGs (Richardson, 1996b), $*$ is used as a *meta-symbol*, indicating one of the three possible edge-endpoints. For instance, $A \longrightarrow\!\!* B$ indicates any of the following edges: A — B, $A \longrightarrow B$, or A—$\circ B$.

Figure 4. Summary of CCD algorithm operation.



Note. Given the observed statistical independencies, CCD constructs a partial ancestral graph (PAG), which represents the *ancestral* features that are common to every directed cyclic graph (DCG) in a Markov equivalence class. In this particular example, the Markov equivalence class consists of two different DCGs.

1. $A*\!\!\rightarrow B$ is interpreted as B is *not an ancestor* of A in every graph in $Equiv(\mathcal{G})$.
2. $A*\!\!-B$ is interpreted as B *is an ancestor* of A in every graph in $Equiv(\mathcal{G})$.
3. $A*\!\!-\!\!\circ B$ is interpreted as the ancestral relation of B with regard to A is undetermined or varies across graphs in $Equiv(\mathcal{G})$.

The PAG output of the CCD algorithm can also include a solid or dotted underlining to provide additional information about the causal relations in triplets. If there is a solid underlining $A*\!\!-\!\!*\underline{B}*\!\!-\!\!*C$, it indicates that B is an ancestor of (at least one of) A or C in every graph in $Equiv(\mathcal{G})$. If there is a dotted underlining added to a collider structure such as $A\rightarrow \underline{B} \leftarrow C$, it indicates that B is *not* a descendant of a common child of A and C in every graph in $Equiv(\mathcal{G})$. For example, from the PAG shown in Figure 4, we can read off the following:

1. X_2 and X_3 are not ancestors of X_1 and X_4 in every graph in $Equiv(\mathcal{G})$.
2. X_1 and X_4 are both ancestors of X_2 and X_3 in every graph in $Equiv(\mathcal{G})$.
3. X_2 is an ancestor of X_3 and X_3 is an ancestor of X_2 in every graph in $Equiv(\mathcal{G})$, indicating the presence of a cyclic relationship between them.
4. X_2 and X_3 are not descendants of a common child of X_1 and X_4 in every graph in $Equiv(\mathcal{G})$. This means that it is not possible for both $X_1 \rightarrow X_2$ and $X_4 \rightarrow X_2$, or both $X_1 \rightarrow X_3$ and $X_4 \rightarrow X_3$ to coexist in any graph in $Equiv(\mathcal{G})$. For instance, if X_1 were to be a parent of X_2, and considering that X_2 and X_3 are ancestors/descendants of each other, X_4 cannot also be a parent of X_2; otherwise this condition would be violated.

Given the causal ancestral relations represented by the example PAG described above, we can corre-

spondingly derive the Markov-equivalent DCGs, which are shown in the right-hand side of Figure 4.

3.2.2 Steps of CCD Algorithm

In this section, we will provide a description of the CCD algorithm, which is the first theoretically well-founded constraint-based method that can be applied in a cyclic setting. The FCI and CCI algorithms share essentially the same structure, but differ in specific orientation rules in the latter part of the algorithm. As such, here we provide a high-level overview of the CCD algorithm in so far as it shares features with the other two more complex algorithms. For this reason, we omit some of the more technical details for simplicity, and refer readers to Appendix A for a more in-depth and comprehensive description of the CCD algorithm.

The CCD algorithm consists of six steps. We illustrate each step using the example DCG from Figure 5(a), which is the same example DCG that we previously introduced in Figure 1(b). The algorithm starts with a fully-connected PAG with circle endpoints, as shown in Figure 5(b), which implies that the direction has not been determined yet. As it proceeds, (some) circles will be replaced by either an arrow head or a tail.

Step 1. This step is identical to the first step of the PC algorithm described above in Section 2.3; the algorithm tests whether two vertices, A and B, are statistically independent given any subset of the remaining variables. When such a subset is found, the algorithm removes $A \circ\!\!-\!\!\circ B$. Since X_1 and X_4 are marginally independent in our example DCG, $X_1 \circ\!\!-\!\!\circ X_4$ is removed, resulting in Figure 5(c).[3]

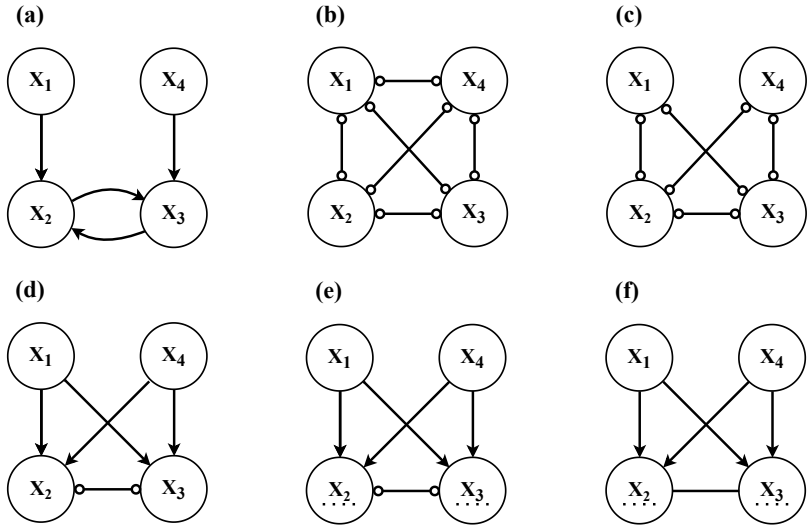
Step 2. Again, the algorithm proceeds in a similar manner to the PC algorithm by searching for collider structures in triplets $A \ast\!\!-\!\!\ast B \ast\!\!-\!\!\ast C$. Once the algorithm identifies B as a collider, the triplet is oriented as $A \rightarrow B \leftarrow C$. Given that X_2 and and X_3 are colliders in our example, $X_1 \circ\!\!-\!\!\circ X_2 \circ\!\!-\!\!\circ X_4$ and $X_1 \circ\!\!-\!\!\circ X_3 \circ\!\!-\!\!\circ X_4$ are oriented respectively as $X_1 \rightarrow X_2 \leftarrow X_4$ and $X_1 \rightarrow X_3 \leftarrow X_4$, resulting in Figure 5(d).

Step 3. The algorithm then checks for a different pattern of d-separating relations in triplets to perform additional orientations. It seeks adjacent variables $A \ast\!\!-\!\!\ast B$ for which it can find a third variable, C, which a) is not directly connected to either A or B, and b) is not d-separated from B given A. In our example, no such structures can be found, since X_1 and X_4 are the only variables not adjacent at this point. Hence, no further orientations are performed in step 3.

Step 4. In this step, the algorithm tries to refine the causal graph by introducing underlinings to collider structures. To do this, it searches for a **Supset** (super separation set), a set of variables that d-separate two endpoint vertices in a collider structure when conditioning on the collider. For each collider structure $A \rightarrow B \leftarrow C$, the algorithm examines the presence of any **Supsets**, and if one is found, a dotted-underlining $A \rightarrow \underline{B} \leftarrow C$ is added. Since X_2 and X_3 are identified as a **Supset** in our

[3]The resulting graph is referred to as an *ancestral* skeleton — an undirected graph of ancestral relations implied by the underlying structure.

Figure 5. Trace of CCD algorithm.



Note. (a) shows the true directed cyclic graph. (b) shows the fully-connected PAG, which is the starting point of the algorithm. (c) shows the *ancestral* skeleton (i.e., an undirected version of the PAG) estimated in step 1. (d) shows the state of the PAG after step 2, where some of the edges are oriented given the identified colliders. (e) shows the state of the PAG after step 4, where the *Supsets* are identified and the corresponding colliders are dotted-underlined. (f) shows the final state of the PAG after step 5, where an additional edge between X_2 and X_3 is oriented.

example, they are dotted-underlined as $X_1 \rightarrow \underset{\cdot\cdot}{X_2} \leftarrow X_4$ and $X_1 \rightarrow \underset{\cdot\cdot}{X_3} \leftarrow X_4$, resulting in Figure 5(e).[4]

Step 5. The last two steps of CCD concern additional orientation of the remaining undirected edges by examining **Supsets** in the context of quadruplets $\langle A, B, C, D \rangle$. In this step, the algorithm identifies quadruplets where B and D serve as colliders for A and C, while each being part of the **Supsets** in triplets involving A and C. When such structure exists, and B and D are connected, the algorithm orients that edge $B \ast\!\!\rightarrow\!\!\ast D$ as $B \ast\!\!-\!\!\ast D$. In our example, a quadruplet matching this criteria is found, resulting in the orientation of $X_2 \circ\!\!-\!\!\circ X_3$ as $X_2 - X_3$, as depicted in Figure 5(f).[5]

Step 6. In the final step, the algorithm searches for a different pattern in quadruplets where B remains a collider and is part of a **Supset**, but D is not adjacent to both A and C. If, in this case, A and D are d-connected given the **Supset** of $\langle A, B, C \rangle$, the algorithm orients the edge $B \ast\!\!\rightarrow\!\!\ast D$ as $B \rightarrow D$. In our example, no such quadruplets exist, so no additional orientation occurs. This ultimately leads to Figure 5(f) as the final PAG. With the final PAG in hand, we can determine the Markov equivalence class of DCGs by reading off all the ancestral relationships represented by the PAG, as discussed in Section 3.2.1.

[4]Recall that underlinings in PAGs can convey additional information on causal relations in triplets, as mentioned in Section 3.2.1. In this example, $X_1 \rightarrow \underset{\cdot\cdot}{X_2} \leftarrow X_4$ and $X_1 \rightarrow \underset{\cdot\cdot}{X_3} \leftarrow X_4$ together indicate that X_2 and X_3 are not descendants of a common child of X_1 and X_4.

[5]This indicates that X_2 and X_3 are ancestors of each other, implying a cyclic causal relationship between them.

3.3 FCI Algorithm

The FCI algorithm, originally proposed by Spirtes et al. (1995), is a constraint-based causal discovery method for directed acyclic graphs (DAGs), which takes into account the presence of latent confounding and possible selection bias. Recently, Mooij and Claassen (2020) demonstrated that the FCI algorithm can also be applied to cyclic causal discovery in the presence of latent confounding under more general faithfulness and Markov conditions, provided that the causal relationships are *non-linear*. For details on these conditions, we refer readers to Forré and Mooij (2017).

3.3.1 FCI Output Representation: Partial Ancestral Graph (PAG)

The FCI algorithm, like the CCD algorithm, aims to identify the underlying causal graph up to its Markov equivalence class and also employs a PAG to represent the common ancestral features among the equivalent graphs. However, allowing latent confounders adds a complication; DCGs are not closed under marginalization over latent confounders, meaning that there exist infinitely many DCGs of observed variables (O) and latent confounders (L) that entail the same set of independencies (Richardson & Spirtes, 2002). This problem arises from the fact that we do not know how many latent confounders are involved, and the algorithm has to account for the possibilities of arbitrarily many latent confounders (Zhang & Spirtes, 2005).

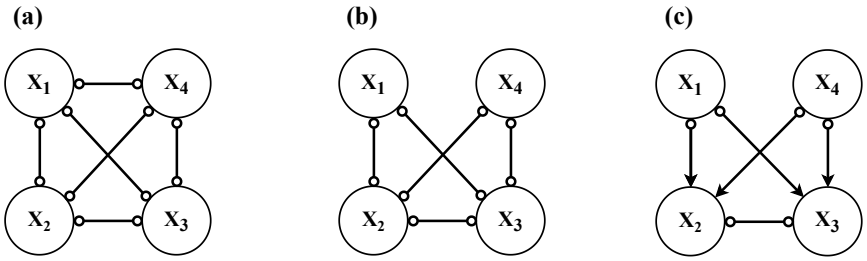
To represent the presence of latent confounders in the infinite space of causal graphs, we need to introduce a new type of edge into the PAG representation. Specifically, we take bidirected edges ($\leftrightarrow$) to reflect the presence of latent confounders, with $A \leftrightarrow B$ denoting a confounding variable between A and B.[6] The interpretation of edges in the PAGs estimated by the FCI algorithm is otherwise the same as in the CCD algorithm, except for the fact that in FCI PAGs, fully-connected vertices with circle endpoints ($\circ\!\!-\!\!\circ$) may indicate a possible cyclic structure (Mooij & Claassen, 2020).

3.3.2 Steps of FCI Algorithm

We will walk through the steps of the FCI algorithm given the same example DCG used for the CCD algorithm (Figure 1(b)). The algorithm consists of three main steps: skeleton discovery, collider structure orientation, and application of further orientation rules, where the first two steps are analogous to the CCD procedure. As with the CCD algorithm, the FCI algorithm begins with a fully-connected PAG with $\circ\!\!-\!\!\circ$ edges between every pair of variables (Figure 6(a)). Then, it estimates the ancestral skeleton (Figure 6(b)) by testing for statistical independence (see step 1 of CCD). Subsequently, the FCI algorithm searches for colliders in the same way as the CCD algorithm (see step 2 of CCD); when a collider (B) is identified, $A \ast\!\!-\!\!\ast B \ast\!\!-\!\!\ast C$ is oriented as $A \ast\!\!\rightarrow B \leftarrow\!\!\ast C$, resulting in Figure 6(c). Lastly, the FCI algorithm executes a set of orientation rules to further orient the edges.

[6] The class of graphs that make use of bidirected edges ($\leftrightarrow$) to represent latent confounding is called *directed mixed graphs* (DMGs), which can be seen as extensions of DCGs (Richardson, 2003).

Figure 6. Trace of FCI algorithm.



Note. (a) shows the fully-connected PAG, which is the starting point of the algorithm. (b) shows the *ancestral* skeleton estimated in the same manner as the CCD algorithm. (c) shows the state of the PAG after the orientation step using the collider structures identified in step 2.

For a complete list of orientation rules (Zhang, 2008), see Appendix B. In this case, no additional endpoints are oriented in further steps, leaving Figure 6(c) as the final resulting PAG. Given this PAG, we can read off that:

1. X_2 and X_3 are not ancestors of X_1 and X_4 in every graph in $Equiv(\mathcal{G})$.
2. X_2 and X_3 might be part of a cycle in $\mathcal{G}$ as they are fully-connected with circle endpoints.

Notice that the PAG produced by FCI has more circle endpoints ($\circ$) than the PAG produced by CCD, which indicates a greater degree of uncertainty about causal ancestral relationships. This is because the FCI algorithm accounts for the possibility of latent confounding, leading to a larger space of possible graphs. Consequently, there are many more equivalent graphs that conform to the relational structure implied by the FCI PAG, resulting in a larger Markov equivalence class, as illustrated in the right-hand side of Figure 7.

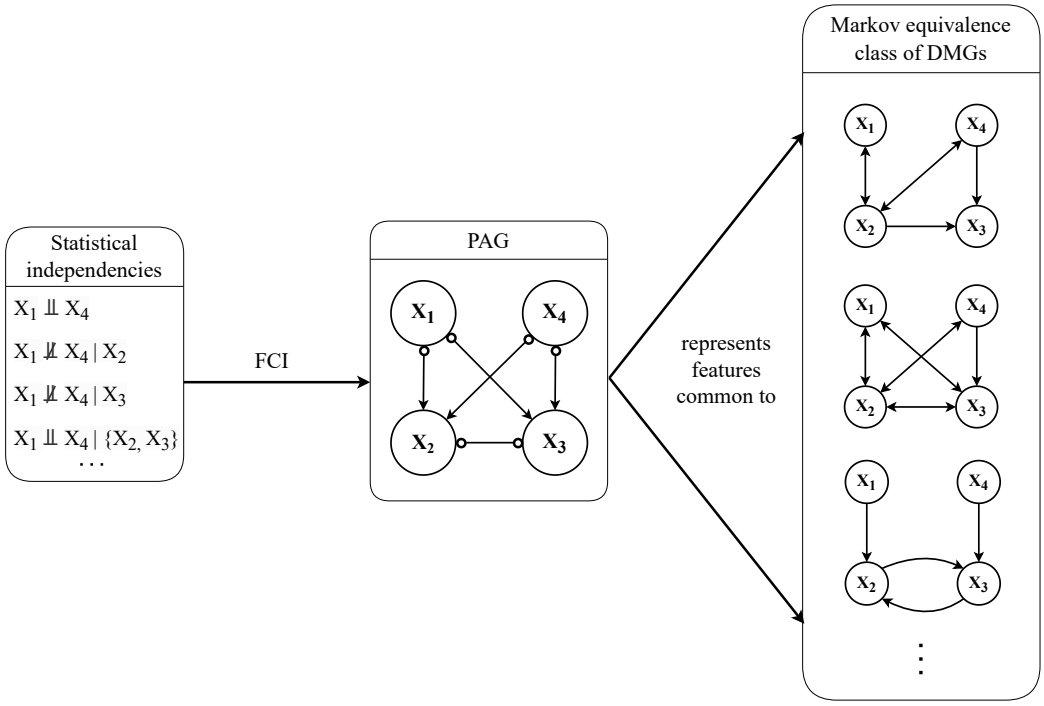
3.4 CCI Algorithm

The CCI algorithm combines features of both the CCD and FCI algorithms. It can identify cyclic causal structures, similar to CCD, and can handle latent confounding, similar to FCI (Strobl, 2019). Employing a combined approach, however, comes at a cost; the algorithm requires more complex edge-endpoint inferences and lengthy orientation rules. For a detailed explanation of each step of the CCI algorithm, see Appendix C.

3.4.1 CCI Output Representation: Partial Ancestral Graph (PAG)

As with the other two algorithms, CCI generates a PAG that captures the common ancestral relationships among equivalent graphs. To account for the presence of latent confounding in the infinite causal graph space, as described in Section 3.3.1, CCI also uses bidirected edges ($\leftrightarrow$). Apart from that, the interpretation of the other types of edges in PAGs estimated by CCI is the same as that

Figure 7. Summary of FCI algorithm operation.



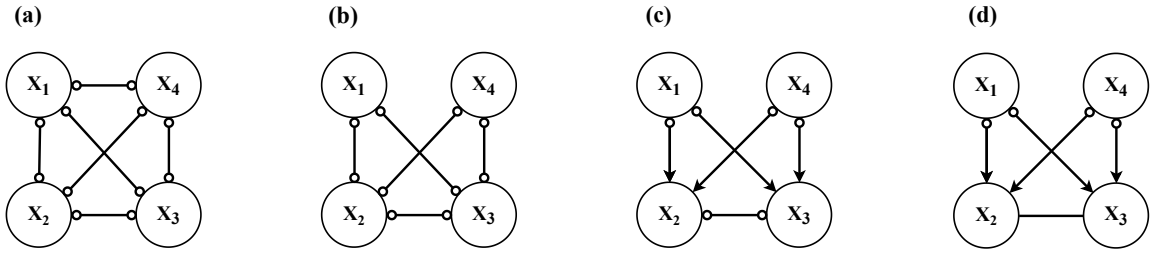
Note. Given the observed statistical independencies, FCI constructs a partial ancestral graph (PAG) that captures the common *ancestral* features of every directed mixed graph (DMG) in a Markov equivalence class. The PAG estimated by FCI has more circle endpoints than the one estimated by CCD in Figure 4, indicating a higher level of uncertainty about the causal relationships. This is because FCI accounts for the presence of latent confounders, which is represented by bidirected edges ($\leftrightarrow$) in the graph. As a result, the Markov equivalence class tends to be relatively large.

described in Section 3.2.1 for the CCD output. In the following section, we will briefly outline the steps of CCI with the same example DCG (from Figure 1(b)) that has been used throughout the paper and examine the interpretation of the resulting PAG.

3.4.2 Steps of CCI Algorithm

The CCI algorithm consists of seven steps in total, the first two of which are identical to those of the other algorithms (i.e., skeleton discovery and collider structure orientation), and the remaining steps are similar to the further orientation rules implemented in CCD and FCI. Same as the others, CCI initiates with a PAG that is fully connected with $\circ\!\!-\!\!\circ$ edges between every pair of variables (Figure 8(a)). After running the skeleton discovery procedure (i.e., step 1), the ancestral skeleton is estimated, as shown in Figure 8(b). Upon orienting the edges based on identified colliders (i.e., step 2), CCI produces the output shown in Figure 8(c). In the following orientation step (i.e., step 5), the edge between X_2 and X_3 is oriented utilizing **Supsets** — similar to step 4 of the CCD algorithm

Figure 8. Trace of CCI algorithm.



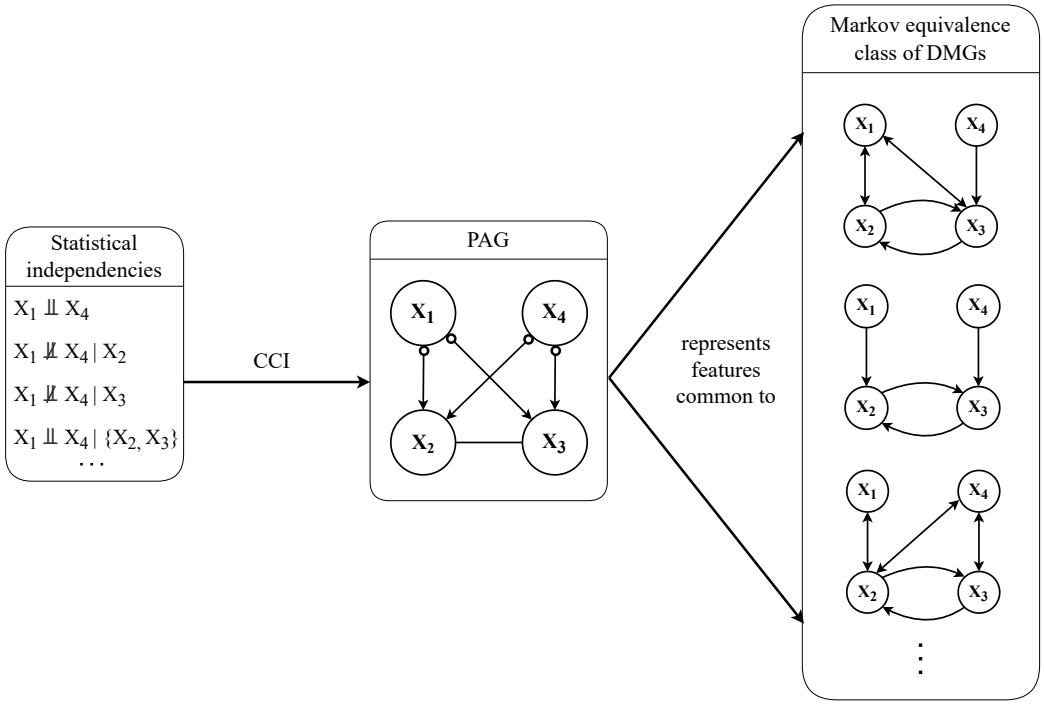
Note. (a) depicts the fully-connected PAG, which is the starting point of the algorithm. (b) depicts the *ancestral* skeleton estimated in the same way as the CCD and FCI algorithms. (c) depicts the state of the PAG after orienting the collider structures in step 2. (d) depicts the state of the PAG after applying the extra orientation rules in step 5. No further orientation takes place in the subsequent steps, leaving (d) as the final PAG.

— which results in Figure 8(d). Since no additional edges are oriented in the subsequent steps, Figure 8(d) remains the final PAG. From this PAG, we can read off the following:

1. X_2 and X_3 are not ancestors of X_1 and X_4 in every graph in $Equiv(\mathcal{G})$.
2. X_2 is an ancestor of X_3 and X_3 is an ancestor of X_2 in every graph in $Equiv(\mathcal{G})$, implying a cyclic relationship between them.

Similar to FCI, the PAG obtained from CCI has more circle endpoints compared to the one obtained from CCD, due to its consideration of latent confounding. Figure 9 summarizes the operation of CCI, which is similar to that of FCI shown in Figure 7, with a relatively large Markov equivalence class of graphs. However, unlike the PAG from FCI where cycles are implied by the fully-connected edge with circle endpoints (∘—∘), the PAG from CCI clearly indicates a cyclic relationship between X_2 and X_3 with an undirected edge (—), which implies the reciprocal ancestral relationship between the two variables.

Figure 9. Summary of CCI algorithm operation.



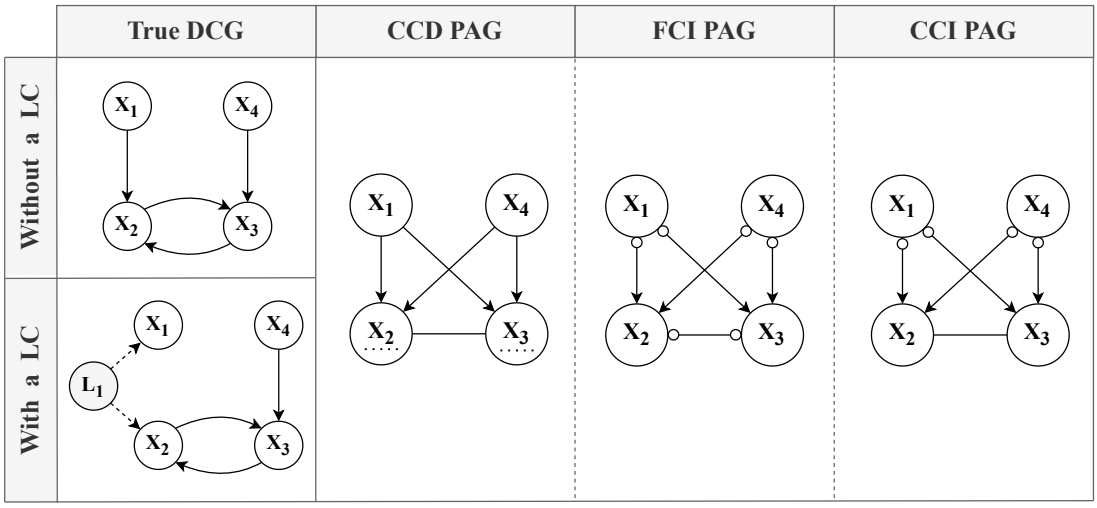
Note. CCI constructs a partial ancestral graph (PAG) based on the observed statistical independencies, which captures the *ancestral* features common to every directed mixed graph (DMG) in a Markov equivalence class. Similar to the FCI algorithm, CCI accounts for the potential presence of latent confounding, which is represented by bidirected edges ($\leftrightarrow$), leading to more circle endpoints in the PAG. This generally results in a larger Markov equivalence class, as illustrated in the right-side of the figure.

3.5 Overview of Algorithms

Thus far, we have introduced three constraint-based causal discovery algorithms for cyclic graphs and discussed their characteristics. Based on our example (DCG from Figure 1(b)), the CCD algorithm may appear to be the preferred option, as it has less uncertainty in its output and correspondingly returns a smaller set of equivalent graphs. However, recall that the main advantage of the other two algorithms — FCI and CCI — is their ability to handle the presence of latent confounding, which commonly occurs in psychological research in practice (Rohrer, 2018).

The left-most panel of Figure 10 depicts the DCGs of two distinct data-generating processes: In the first, the causal graph consists of four variables $X_1, \ldots, X_4$, while in the second, the causal graph consists of a fifth variable L_1, which we will consider to represent a latent confounder. Now consider what output we would expect each of the algorithms reviewed above to return if they were fit to data generated from either of these data-generating systems: All three algorithms happen to generate the same PAG for these two different DCGs, shown in the remaining panels of Figure 10. Suppose the true underlying causal structure is the DCG without a latent confounder (top left in

Figure 10. Comparison of the partial ancestral graphs (PAGs).



Note. Two different DCGs are displayed in the first column, with subsequent columns showing the corresponding PAGs estimated by each algorithm. In this particular case, all three algorithms happen to produce identical PAG estimates for both DCGs. *LC* = latent confounder; *DCG* = directed cyclic graph; *PAG* = partial ancestral graph.

Figure 10). In that case, all three PAGs correctly depict the ancestral features of the true DCG. However, the PAG output by CCD is by far the most*precise*, as it contains no circle endpoints, thus representing the smallest equivalent set of graphs. Now, suppose the underlying causal structure is the DCG with a latent confounder (bottom left in Figure 10). Then, the PAG generated by CCD contains errors (i.e., X_1 is not an ancestor of X_2 and X_3), while the PAGs estimated by FCI and CCI correctly represent the ancestral features of the true DCG. In the following section, we will conduct a simulation study to compare the performance of each algorithm under various conditions and investigate which algorithm is most suitable for situations likely to arise in psychological research.

4 Simulation

In this section, we present a simulation study to evaluate the performance of the three causal discovery algorithms introduced above under different conditions: namely, the sample size, number of variables in the graph, density, and presence of latent confounders. We investigate how these factors affect each algorithm's performance and whether any algorithm outperforms the others in specific ways. In the following, we will first discuss the data generation process, simulation design, and evaluation metrics, before presenting the results of the simulation study.

4.1 Data Generation

We simulate data from the different cyclic models, as shown in Figure 11, all of which are characterized by *linear* relations and *independent Gaussian* error terms. These types of models are often used in psychological research, and for such cyclic models, the global Markov property — the necessary condition for constraint-based causal discovery — also holds, as discussed in Section 2.2.

To generate data, we first define a coefficient matrix $\mathbf{B}$. A non-zero entry B_{ij} indicates an edge from X_j to X_i with a strength of B_{ij}. Thus, $X_1, ..., X_p$ can be generated according to the following equation:

$$X_i = \sum_{r=1}^{p} B_{ir} X_r + \varepsilon_i,$$

for $i = 1, ..., p$, where p is the number of vertices and ε are mutually independent $\mathcal{N}(0, 1)$ random variables. The variables $X_1, ..., X_p$ then have a multivariate Gaussian distribution with a mean vector of zero and a covariance matrix $\mathbf{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B})^{-T}$, where $\mathbf{I}$ denotes the identity matrix. Note that this data generation scheme is possible provided that $(\mathbf{I} - \mathbf{B})$ is invertible, which is the case when the eigenvalues of $\mathbf{B}$ are smaller than one in absolute value, $|\lambda| < 1$ (Eberhardt et al., 2010). While this is guaranteed if $\mathbf{B}$ defines an acyclic model (Drton & Maathuis, 2017), for cyclic models, this does not always hold. To satisfy this condition, cyclic relations need to be not too strong so that the dynamical system converges to equilibrium (Rothenhäusler et al., 2015). We set fixed values for $\mathbf{B}$ to make the simulation results easier to track and interpret. When specifying the $\mathbf{B}$ matrices, we choose values within a range that is deemed reasonable (e.g., restricting the strength of cyclic relations to relatively small such that they do not diverge) and verify that the eigenvalues satisfy the equilibrium condition.[7] When this condition is violated, we adjust the parameters until it is satisfied. In addition, to ensure that the simulation results are not dependent on the specific values of the coefficient matrix $\mathbf{B}$, we perform a sensitivity analysis. In this analysis, we randomly sample parameter values of $\mathbf{B}$ from a uniform distribution on $[-0.9, -0.1] \cup [0.1, 0.9]$ in each iteration. We then check whether the equilibrium condition is met, and if not, we re-sample the parameters until it is satisfied. In the following, we provide a detailed description of the simulation setup, which is replicated across both simulation studies.

4.2 Simulation Setup

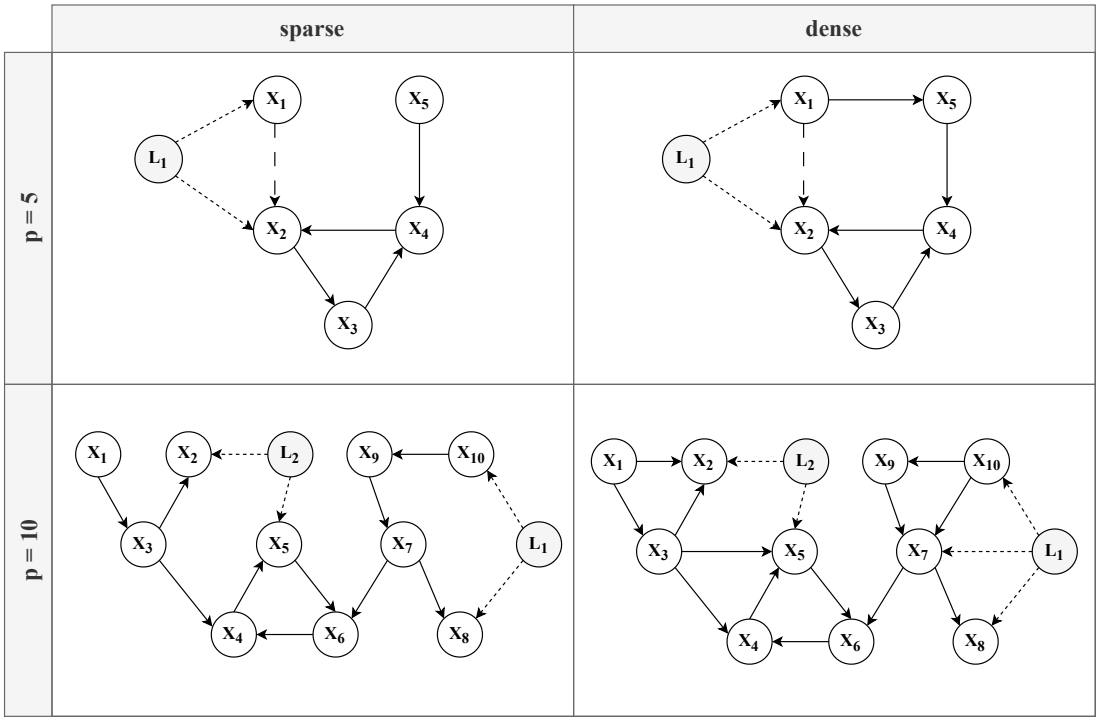
We test each algorithm under different conditions by varying the number of variables (rows in Figure 11) and the number of edges — the density (columns in Figure 11). We also explore the impact of unobserved confounding by introducing latent variables (L_1 and L_2 in Figure 11), as latent confounding is a common issue in psychological research, particularly when it comes to inferring causal relationships (Hallquist et al., 2021; Rohrer et al., 2022). Lastly, we vary the sample size

[7] For more details on the specific $\mathbf{B}$ matrices used in the simulations, please refer to Appendix D.

ranging from small to moderately large, $N \in \{50, 150, 500, 1000, 2000, 3000, 4000, 5000, 7500, 10000\}$, for every simulated cyclic model. This leads to a $2 \times 2 \times 2 \times 10$ design; number of variables $\times$ density $\times$ latent confounder (presence/absence) $\times$ sample size. We simulate each condition 500 times, estimating three PAGs per iteration using each algorithm. See Figure 11 for an overview of the different types of cyclic models used in the simulation. These models precisely depict the ones utilized in the simulation.

To test for conditional independence, we employ partial correlations since the variables in our simulated data have linear relations with additive Gaussian errors. In such cases, conditional independence is equivalent to zero partial correlation (Lawrance, 1976). Throughout the study, we use a fixed alpha level of 0.01 ($\alpha = 0.01$) for conducting the conditional independence tests, which is a commonly used value in causal discovery studies with moderate sample sizes (Malinsky & Danks, 2018). However, we acknowledge that following the fixed α convention is not straightforwardly justified (Strobl et al., 2017). To address this, we perform another sensitivity analysis where we

Figure 11. Simulation settings.



Note. In our simulation study, we vary several factors such as the number of variables: $p \in \{5, 10\}$, the density: sparse / dense, the influence of a latent confounder: absence / presence, and the sample size: $N \in \{50, 150, 500, 1000, 2000, 3000, 4000, 5000, 7500, 10000\}$. This results in a $2 \times 2 \times 2 \times 10$ simulation design, with each combination of factors (except for N) yielding a unique model structure. Note that the edge between X_1 and X_2 (long-dashed line $--$) in the 5-variable models (top row) is only present in the conditions without a latent confounder.

adjust the α level based on the sample size. Further elaboration on this analysis will be provided in Appendix H. All simulations are performed using R software version 4.2.3 (R Core Team, 2023).

4.3 Evaluation Metrics

As discussed in Section 3, the algorithms attempt to recover the ancestral graphs implied by the underlying DCG. In order to evaluate the performance of each algorithm, we compare how well it recovers the true ancestral graph. To do this, we first need to construct a true ancestral graph for each simulated condition, which can then be compared to the estimated graphs. The true ancestral graph for each DCG considered in our simulation study is shown in Figure 12. For those interested, a step-by-step procedure for deriving a true ancestral graph from a DCG is provided in Appendix E.

For assessment, we use both *local* and *global* evaluation metrics. At the local level, we look at the individual edge-endpoints and compare whether they match the corresponding endpoint of the true graph. At the global level, we look at the graph structure as a whole and measure the distance between the true and estimated graph, assessing how closely the estimated graph's structure matches the true graph's structure. As the local metrics, we utilize *precision*, *recall*, and *uncertainty rate*.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} \ + \ \text{False Positive}}$$

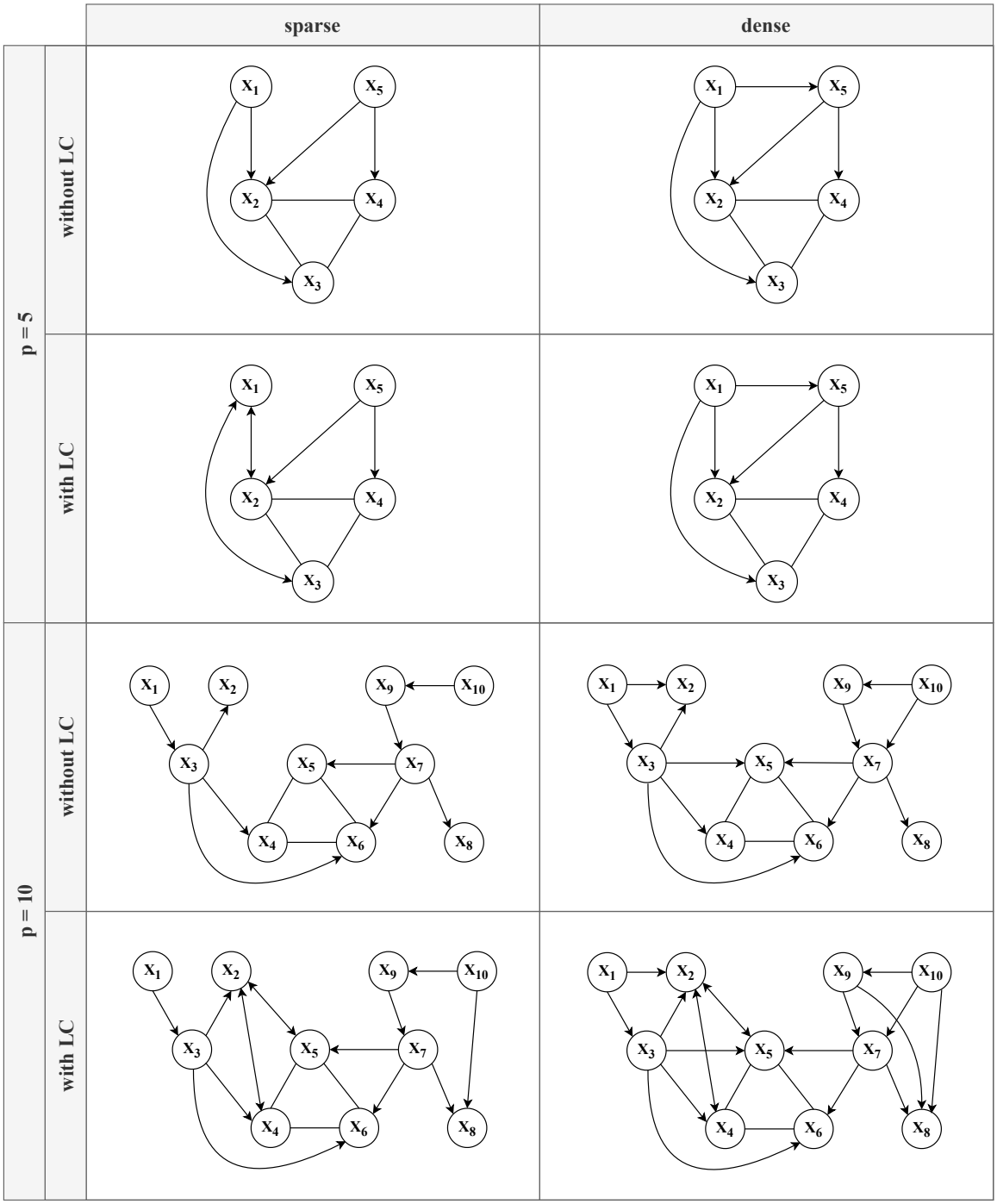
$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} \ + \ \text{False Negative}}$$

$$\text{Uncertainty rate} = \frac{\text{Number of circle endpoints } (\circ)}{\text{Total number of edge-endpoints}}$$

Precision reflects the prediction accuracy (i.e., out of all predicted cases, how many are correct), and recall reflects the retrieval rate (i.e., out of all true cases, how many are retrieved). Each edge-endpoint in a resulting graph can fall into one of four categories: no edge-endpoint (null), arrow head ($>$), arrow tail ($-$), and circle ($\circ$). Given that circle endpoints indicate that an algorithm is unsure of the direction of causal relations, the uncertainty rate is defined as the proportion of the circle endpoints present in the output. For example, consider the example PAG shown in Figure 13(b). The uncertainty rate for this PAG can be calculated as $\frac{3}{20}$, where the number of circle endpoints is 3, and the total number of edge-endpoints is 20 ($\binom{5}{2} \times 2 = 20$). For the other *non-circle* endpoints, we compute the precision and recall. As an example, suppose that Figure 13(a) is the true ancestral graph and Figure 13(b) is the estimated PAG output. We can then construct a confusion matrix of the estimated versus true edge-endpoints.[8] Based on the confusion matrix shown in Figure 13(c), we can compute the precision and recall for each type of non-circle endpoints. For the arrow head ($>$), for instance, the precision and recall are computed as: $\textit{precision} = \frac{4}{4+3+0}$ and $\textit{recall} = \frac{4}{4+0+0}$. Note

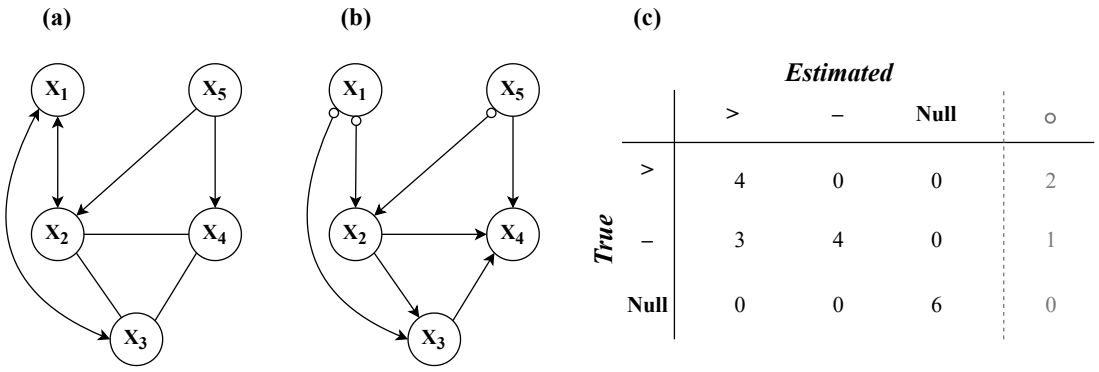
[8]This example is based on the 5-variable sparse condition with a latent confounder shown in the top left panel of Figure 11.

Figure 12. True ancestral graph for each condition.



Note. Each panel in the figure depicts the true ancestral graph for each simulated condition. Directed edges denote ancestral relationships ($A \longrightarrow B$ means A is an ancestor of B), bidirected edges indicate the presence of latent confounders ($A \longleftrightarrow B$ means there is a latent confounder between A and B), and undirected edges denote mutual ancestral relationships, thereby indicating cyclic relationships between the corresponding variables (A — B means A and B are ancestors of each other, implying the presence of a cyclic relationship between them). For more information on how to derive the true ancestral graph from a DCG, see Appendix E. p = number of variables; *LC* = latent confounder.

Figure 13. Example performance evaluation.



Note. For the purpose of illustration, we consider the 5-variable sparse condition with a latent confounder shown in the top left panel of Figure 11. Panel (a) displays the true ancestral graph, while panel (b) presents an example of the estimated PAG output. Panel (c) shows the confusion matrix for the estimated versus true edge-endpoint, where the true endpoints are presented in rows and the estimated endpoints in columns. There are in total four possible edge-endpoints that can occur in an output: arrow head (>), arrow tail (−), null (no endpoint), and circle (○). The circle endpoints, however, are not counted toward the calculation of *precision* and *recall*, and thus are greyed out in the table.

that the circle endpoints are not considered in the calculation of precision and recall and therefore they are greyed out in the confusion matrix below.

As the global metric, we use the *structural Hamming distance* (SHD) (de Jongh & Druzdzel, 2009). SHD quantifies the level of differences between two graphs by counting the number of edge insertions, deletions, and direction changes required to move from one graph (estimated graph $\hat{\mathcal{G}}$) to the other (true graph $\mathcal{G}$). It can be formulated as: $SHD = A + D + C$, where A, D, and C represent the number of added edges, deleted edges, and changes in edge-endpoints, respectively. A lower SHD value suggests that the estimated graph ($\hat{\mathcal{G}}$) is more closely aligned with the true graph ($\mathcal{G}$), indicating a better recovery of the true graph structure. For example, the SHD value for the PAG output in Figure 13(b) — provided that the true ancestral graph is Figure 13(a) — is 6, which is calculated by summing: 0 (A) + 0 (D) + 6 (C). As such, in contrast to the precision and recall metrics, which omit circle endpoints in their computation, and the uncertainty metric, which directly counts only the degree of circle endpoints, the SHD metric treats circle endpoints as equivalent to any other type of endpoint error.

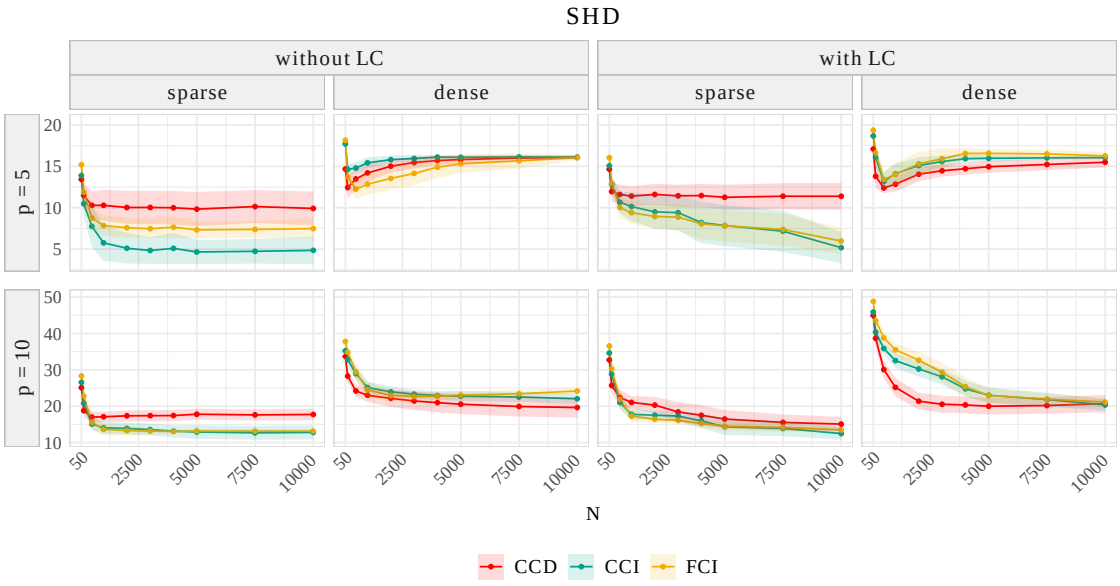
4.4 Simulation Results

In this section, we will report the results of the simulation study using the fixed values of **B** matrices. First, we will present the overall performance of the considered algorithms based on the structural Hamming distance (SHD), and then we will take a more detailed look into their performance using precision, recall, and uncertainty rate.

4.4.1 Performance on Global Metric

The SHD values for each algorithm per condition are shown in Figure 14, where the points and shaded area represent the average SHD values of 500 iterations and the corresponding interquartile range (IQR), respectively. As discussed in Section 4.3, a lower SHD value indicates a better performance, as it means that the estimated graph is closer to the true graph. Overall, the FCI and CCI algorithms perform better in sparse conditions, while the CCD algorithm outperforms the others in dense conditions. Additionally, the performance of all three algorithms is generally worse in dense conditions than in sparse conditions. Interestingly, in small models ($p = 5$) with high density (second and fourth columns from the top row of Figure 14), the SHD values momentarily decrease and then start increasing again as the sample size (N) becomes larger. This is partly against our expectation that the SHD values would decrease monotonically with increasing sample size. Furthermore, we do not observe any significant contrasting patterns between conditions with and without latent confounders. This is contrary to our expectation that CCD would perform relatively better in conditions without latent confounders, while FCI and CCI would perform better in conditions involving latent confounders as they can handle latent confounding.

Figure 14. Structural Hamming distance (SHD).



Note. The sample size (N) is shown on the x-axis, and the SHD values are shown on the y-axis. Each point represents the average SHD value across 500 simulations, while the shaded area represents the interquartile range (IQR). p = number of variables; LC = latent confounder.

4.4.2 Performance on Local Metrics

In what follows, we will further examine the performance of the algorithms using local metrics, while addressing some of the unexpected findings discussed previously in Section 4.4.1. These include (*i*) the observation that CCI and FCI outperform CCD in sparse conditions, even when latent confounders are not present and (*ii*) the SHD values increase instead of decreasing with increasing sample size in the *5-variable dense* conditions.

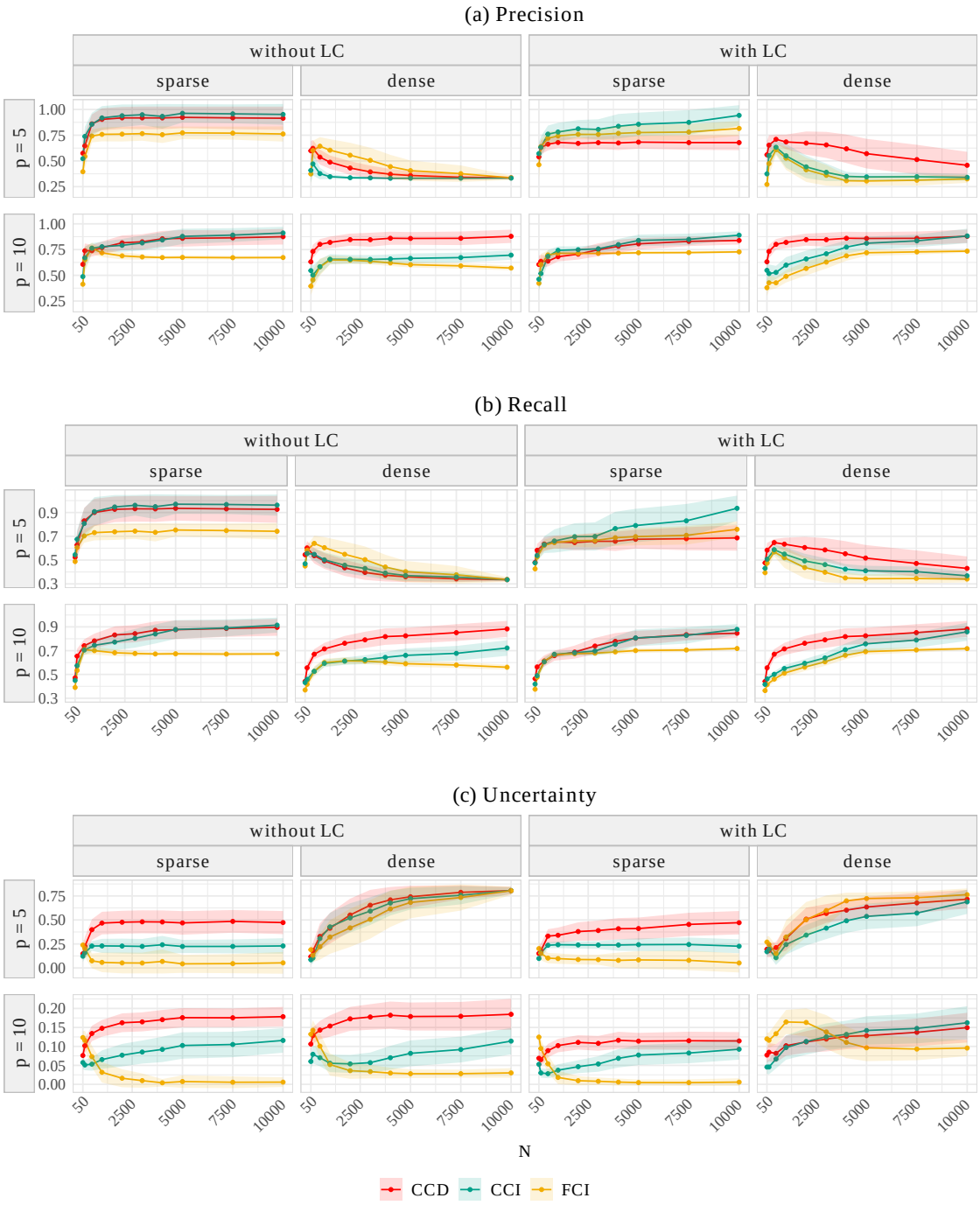
The results for precision, recall, and uncertainty rate for each algorithm are summarized in Figure 15, where each point represents the average value of the metrics over 500 iterations in the corresponding condition, and the shaded area represents the interquartile range (IQR) of the values obtained. As with the SHD results, we observe that the CCI algorithm generally exhibits high precision and recall in sparse conditions. On the other hand, the FCI algorithm records low precision and recall, even lower than CCD. This might seem contradictory at first glance, given that FCI has better SHD values than CCD in most conditions (except for a few dense cases). However, this can be explained by the overall low uncertainty rate of FCI as opposed to the high uncertainty rate of CCD, as shown in Figure 15(c). FCI tends to guess directions without necessarily outputting circles, while CCD tends to produce circles rather than guessing directions. This contrasting behavior consequently leads to FCI having better SHD values than CCD. To illustrate this point further, we examine the most frequently estimated PAGs from each algorithm in the *5-variable sparse* condition without a latent confounder presented in Figure 16.[9] In Figure 16(c), we can observe the typical orienting behavior of FCI, where it guesses the directions for every edge endpoint (i.e., not conservative), in contrast to CCD, which often produces circles (i.e., conservative), as shown in Figure 16(b). Based on the true graph shown in Figure 16(a), we can calculate the SHD: 9 for CCD (Figure 16(b)), 7 for FCI (Figure 16(c)), and 4 for CCI (Figure 16(d)). As such, FCI achieves a lower SHD value than CCD, despite having relatively lower precision and recall values due to some incorrectly predicted edge-endpoints.[10] This highlights the distinct properties of the metrics used in our study. SHD is agnostic about whether an algorithm is *conservative* or *less conservative*, since it measures all differences between the estimated and true graphs, including the circle marks. As a result, it penalizes an algorithm for being conservative as much as for making incorrect predictions. On the other hand, precision and recall only consider non-circle endpoints, while disregarding the circle marks. Therefore, they penalize an algorithm for making incorrect predictions but not for being conservative (i.e., producing circles).

In addition, the comparison between CCI and CCD based on precision and recall in Figure 15 reveals only a subtle difference in performance, even in the *5-variable sparse* conditions where CCI outperformed CCD significantly in terms of SHD. This can be again attributed to CCD having

[9] These PAGs are constructed by assigning each edge-endpoint to the endpoint type that appears most frequently across all iterations.

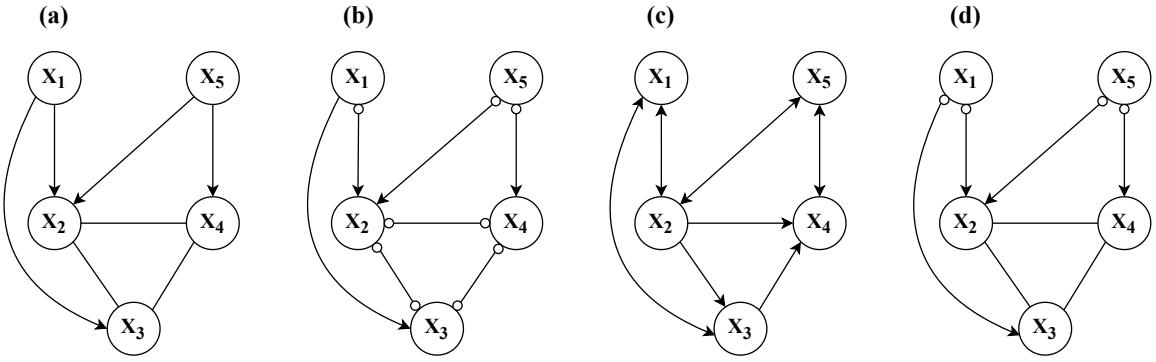
[10] Similar patterns are observed in the other sparse conditions, as shown in the first and third columns of Figure 15.

Figure 15. Precision, recall, and uncertainty rate.



Note. The sample size (N) is shown on the x-axis, and the corresponding metric values are shown on the y-axis. Each point on the graph represents the average of iteration-specific values of precision (top panel), recall (middle panel), and uncertainty rate (bottom panel) for each condition, with the shaded area indicating the interquartile range (IQR). p = number of variables; *LC* = latent confounder.

Figure 16. Frequently estimated PAGs in the 5-variable sparse condition without a latent confounder.

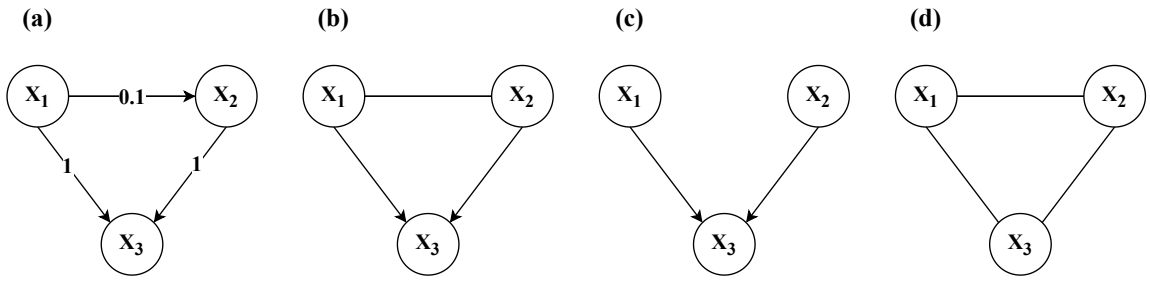


Note. Panel (a) shows the true ancestral graph of the 5-variable sparse condition without a latent confounder. Panels (b), (c), and (d) present the most frequently occurring PAGs in the 5-variable sparse condition without a latent confounder resulted from the CCD, FCI, and CCI algorithms, respectively. They were obtained by picking the most frequent type of edge-endpoints produced by each algorithm from 500 simulations with a sample size of 1000.

a comparatively high uncertainty rate. From the PAG estimated by CCD in Figure 16(b), it can be observed that all of the directions predicted by CCD are correct, but the PAG contains a fairly large number of circle endpoints. These circle endpoints lead to a poorer SHD value for CCD, even though it exhibits comparably high precision and recall for correctly predicting directions. Moreover, further examination of the resulting PAGs shows that CCI performs better than CCD in detecting mutual ancestral relationships in cyclic structures. For instance, in Figure 16(b), CCD assigned all circle marks to the cycle involving X_2, X_3, and X_4, while CCI correctly identified the mutual ancestral relationships in the cycle with undirected edges (—), as shown in Figure 16(d). This superiority of CCI in recovering mutual ancestry in cycles was consistently observed in other conditions as well, suggesting that this property is a significant factor contributing to CCI's overall better performance compared to other algorithms, even in the absence of latent confounding.

With respect to the unexpected increase in SHD values in the *5-variable dense* conditions, we also observe an unusual decrease in precision and recall as the sample size increases (second and fourth panels from the top rows of Figure 15(a) and Figure 15(b)), along with an increase in the uncertainty rate (second and fourth panels from the top row of Figure 15(c)), indicating that more circle endpoints appear in the resulting PAGs as the sample size grows. Learning a dense causal structure is generally more challenging, because there is less information available about the conditional independence relations when a large number of vertices are connected by edges (i.e., almost everything is dependent on everything else). This problem is further compounded in the presence of cycles, as all variables that are part of cycles are completely dependent on each other, making it even

Figure 17. An example dense graph with a weak edge.



Note. Panel (a) shows the true graph. Panel (b) displays the desired output where the collider structure is correctly oriented while the weak edge between X_1 and X_2 is also identified. Panel (c) presents the output graph when an algorithm fails to detect the weak edge between X_1 and X_2 but correctly orients the collider structure. Panel (d) displays the output graph when an algorithm detects the weak edge between X_1 and X_2. This results in a denser graph with less information on independencies, leading to the loss of orientation of the collider structure.

more challenging to obtain information about independence relations. Therefore, algorithms tend to fail to orient any edges and mostly output only circle endpoints in dense cyclic graphs. The decline in precision and recall, along with the rapid increase in uncertainty rate observed in Figure 15 for the *5-variable dense* conditions, are primarily due to this challenge in high-density situations — as the sample size becomes larger, relatively weak edges start getting picked up by the algorithms, leading to denser structures that make causal discovery exceedingly difficult.

However, interestingly, we also note a slight increase in both precision and recall for the *5-variable dense* conditions when the sample size is relatively small, accompanied by a small drop in the SHD (second and fourth panels from the top row of Figure 14). This might seem counterintuitive, as we typically expect the algorithms to learn causal structures more accurately with larger sample sizes. But, in fact, it is possible to lose the correct edge orientation when the sample size is large, particularly in the case of a dense graph with weak edges (Eigenmann et al., 2017). Figure 17 illustrates a simple example case where the inclusion of a weak edge leads to the loss of edge orientation. With a large sample size, algorithms are likely to pick the weak edge between X_1 and X_2, resulting in an uninformative undirected graph, as shown in Figure 17(d). However, with a small sample size, algorithms are more likely to miss the weak edge and identify the collider X_3, thereby obtaining the correct edge orientation, as shown in Figure 17(c). In our *5-variable dense* scenarios, we run into a similar situation. There is a relatively weak edge between X_2 and X_5, which is not detected by algorithms when the sample size is small, resulting in some correctly oriented edges. However, as the sample size becomes larger, the weak edge gets detected, and the algorithms fail to orient any edges, thus yielding completely undirected graphs (see Appendix F for a more detailed explanation of the results in our *5-variable dense* cases). This explains the small dips at

the beginning of the SHD line graphs (Figure 14) and brief spikes in the precision and recall graphs (Figure 15) for the *5-variable dense* conditions with relatively small sample sizes.

To summarize, both CCD and CCI algorithms show good performance, with CCI slightly outperforming in sparse conditions and CCD in dense conditions. However, CCD tends to be more conservative in terms of edge orientation and produces more circle endpoints than CCI, which often results in higher SHD values. In contrast, FCI demonstrates poor performance across most conditions, which is not immediately obvious when considering only the global metric, SHD. Further analysis of local metrics reveals that FCI often makes quick directional inferences without producing circle marks, resulting in comparably good SHD values but low precision and recall values.

In addition, to investigate the robustness of our findings against the specific parameter values chosen in this simulation, including the coefficients of **B** matrices and the value of α, we conducted additional simulations as part of sensitivity analyses. In one of these simulations, we randomly sampled parameter values for **B** at each iteration. Our results were similar to the original simulation, with a few minor differences; we do not find any unusual kinks or dips in the performance curves with small sample sizes in the *5-variable dense* conditions (see Figure 18). Instead, we find that the algorithms' performance steadily improves with increasing sample size, which aligns with our initial expectation. In another simulation, we varied the α level according to the sample size; decreasing α as the sample size (N) increased such that $\alpha_N \longrightarrow 0$ as $N \longrightarrow \infty$ at a suitable rate.[11] It is commonly suggested to adjust α according to the sample size in order to ensure consistent results of conditional independence tests (Mooij et al., 2020; Colombo et al., 2012).[12] Overall, the results show no significant differences in the observed patterns across all conditions, indicating that the effect of adjusting α is negligible. The detailed results of these sensitivity analyses can be found in Appendix G and Appendix H, respectively.

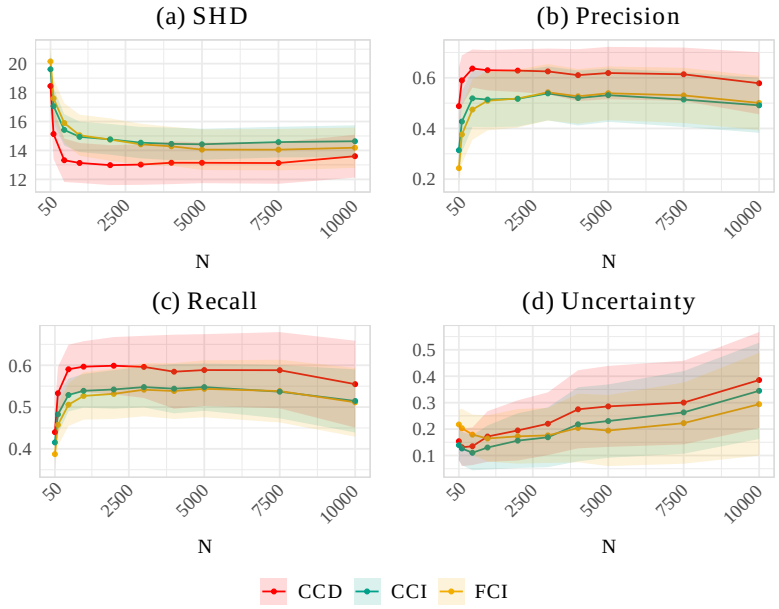
5 Empirical Example

In this section, we present an empirical example where we apply the three causal discovery algorithms studied above to real psychological data to assess their practical applicability. We begin by introducing the dataset and then present the output of each algorithm, along with the corresponding statistical network model. Our goal is to assess whether the output PAGs carry meaningful information about the causal structure and to compare the insights gained from these PAGs with those obtained from the statistical network model. To aid researchers with applying these methods in their own research, we provide an annotated script to reproduce our empirical analysis in

[11] When setting the α level, we adopted a heuristic value of $\alpha = \frac{1}{\sqrt{N}}$, given that the partial correlation can decay as $N^{-1/2+\epsilon}$ for any $0 < \epsilon < 1/2$ in multivariate Gaussian cases (Kalisch & Buehlmann, 2005).

[12] If alpha is not adjusted appropriately based on the sample size, it can lead to either accepting the null hypothesis of independence ($H_0 : X_i \perp\!\!\!\perp X_j \mid \mathbf{Y}$, where a set $\mathbf{Y} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$) with small sample sizes, falsely declaring everything as independent, or rejecting it with large sample sizes, failing to find any independence relations (Magliacane et al., 2017).

Figure 18. Performance in the 5-variable dense condition without a latent confounder.



Note. This figure illustrates the performance of each algorithm in the 5-variable dense condition without a latent confounder, where the coefficients of **B** were randomly sampled. The sample size (N) is shown on the x-axis, and the corresponding metric values are shown on the y-axis. Each point represents the mean value of each metric across 500 simulations, and the shaded area represents the interquartile range (IQR). Note that the performance under the *5-variable dense* condition *with* a latent confounder shows more or less the same patterns (see Appendix G for details).

the reproducibility archive of this paper: https://github.com/KyuriP/Discovering_CCM/tree/main/empirical_example.

5.1 Data and Model Fitting

The dataset used in our empirical example is from McNally et al. (2017), where the authors focused on examining the causal relationships between symptoms of obsessive-compulsive disorder (OCD) and depression.[13] Notably, the authors expressed a particular interest in investigating cyclic causal relationships. However, the causal discovery methods they were familiar with were only applicable to acyclic causal structures, leading them to use a statistical network model. Although they acknowledged that the network captures statistical rather than causal relationships, they argued that statistical network analysis addresses a key limitation of the acyclic causal discovery algorithm, which is excluding the possibility of feedback loops (cycles). Their explicit interest in cyclic causal discovery makes this dataset an ideal candidate for testing the cyclic causal discovery algorithms discussed in the current paper.

[13]The data set is publicly available on the Psychological Medicine Journal webpage.

Table 2. Summary of depression symptoms.

Symptom (*abbreviation*)	Mean (SD)
1. Sleep-onset insomnia (*ons*)	1.20 (1.07)
2. Middle insomnia (*mdd*)	1.44 (1.07)
3. Early morning awakening (*lat*)	0.81 (1.07)
4. Hypersomnia (*hyp*)	1.01 (0.99)
5. Sadness (*sad*)	1.55 (0.94)
6. Decreased appetite (*dcp*)	0.49 (0.72)
7. Increased appetite (*inc*)	0.44 (0.87)
8. Weight loss (*wghtl*)	0.50 (0.94)
9. Weight gain (*wghtg*)	0.67 (1.04)
10. Concentration impairment (*cnc*)	1.48 (0.87)
11. Guilt and self-blame (*glt*)	1.56 (1.17)
12. Suicidal thoughts (*scd*)	0.63 (0.82)
13. Anhedonia (*anh*)	1.27 (1.05)
14. Fatigue (*ftg*)	1.33 (0.95)
15. Psychomotor retardation (*rtr*)	0.66 (0.81)
16. Agitation (*agt*)	1.10 (0.93)

Note. SD = standard deviation.

The dataset consists of 408 observations and 26 variables, comprising 16 depression symptoms and 10 OCD symptoms, all with no missing values. The severity of each symptom was measured using a four-point Likert scale, with 0 indicating no symptoms and 3 indicating extreme symptoms. The participants' age ranged from 18 to 69 years, with a mean of 31.1 and a standard deviation of 12.2. In this paper, we focus our analysis on a subset of the variables — depression symptoms — in order to keep the model size manageable and facilitate the interpretation of the estimated PAGs. For a summary of the depression symptom variables, please refer to Table 2.

To compare the results of the cyclic causal discovery algorithms with those of network analysis, we estimated a Gaussian graphical model (GGM). In a GGM, the edges signify partial correlations between pairs of nodes, controlling for the rest of nodes in the network with the assumption that the variables follow a normal distribution. (Epskamp et al., 2018b). To obtain a sparse network, we used the *graphical lasso* (glasso) method to regularize partial correlations, such that small partial correlations are driven to zero and therefore do not appear in the network (Friedman et al., 2008). For the causal models, we ran all three algorithms (CCD, FCI, and CCI) while setting the alpha level to 0.01 ($\alpha = 0.01$) as a rough way to correct for spurious edges resulting from a relatively

large set of variables (Zhang et al., 2012). As in the simulation study, we used partial correlations to test for conditional independencies. In the following section, we will present the estimated models and interpret the findings from each model while comparing them to one another.

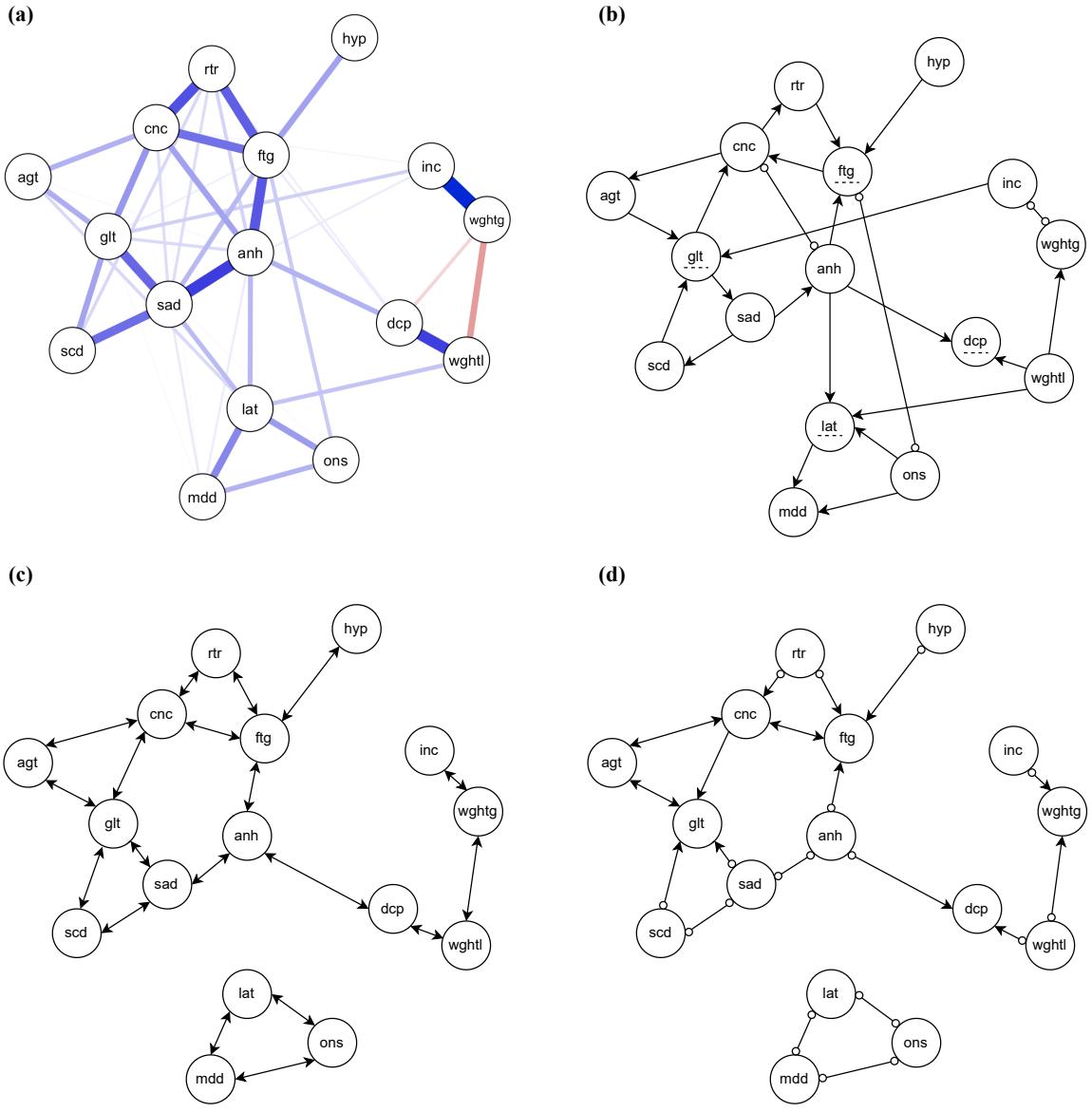
5.2 Empirical Analysis Results

The causal models estimated by each algorithm (i.e., PAGs) and the GGM are depicted in Figure 19. Upon examining the GGM presented in Figure 19(a), we can identify three symptom clusters that are highly interconnected within themselves but less connected with the rest. On the right-hand side of the network, we see symptoms related to physical weight and appetite: weight gain (*wghtg*), weight loss (*wghtl*), increased appetite (*inc*), and decreased appetite (*dcp*), with particularly strong partial correlations between *inc* – *wghtg* and *dcp* – *wghtl*. Towards the bottom of the network, we see three symptoms relating to insomnia: early morning awakening (late), sleep onset insomnia (ons), and middle insomnia (mdd). The remaining items towards the left side of the network form a single cluster. Here, we observe strong connections among several symptoms, including psychomotor retardation (*rtr*) – concentration impairment (*cnc*) – fatigue (*ftg*) – anhedonia (*anh*) – sadness (*sad*) – suicidal thoughts (*scd*). Especially, fatigue, anhedonia, and sadness emerge as the most central symptoms based on the number and strength of connections they have with the other nodes in the network. As McNally et al. (2017) note, this is in accordance with the general clinical observations where sadness and anhedonia, as the two gateway symptoms, are necessary for a diagnosis of depression.

Looking at the PAG from the CCD algorithm shown in Figure 19(b), we can read-off information about possible directions of causal relations between these variables. Overall, the estimated structure is not too different from that of the GGM, in terms of the presence or absence of edges, but it is more sparse. The estimated PAG from CCD reveals that fatigue and anhedonia still remain central nodes. However, in the causal graph, we can now attribute their centrality to the fact that fatigue is the (indirect) effect of many variables, and anhedonia is the (indirect) cause of many variables — the directional information that could not be obtained from the statistical network model above. The three-cluster structure observed in the GGM is also present in this PAG, with only a few edges linking the separate clusters. For example, on the right side of the graph, we can see that increased appetite (*inc*) is an ancestor of guilt (*glt*), and weight loss (*wghtl*) is an ancestor of early morning awakening (*lat*). Anhedonia (*anh*) also plays an important role in linking these separate symptom clusters, acting as an ancestor of both early morning waking (*lat*) and decreased appetite (*dcp*). On the left side of the graph, we observe an intricate network where most of the variables are involved in a web of cyclic ancestral relationships.[14] While it is tempting to interpret these ancestral

[14]The cyclic structures include chains such as fatigue (*ftg*) ⟶ concentration (*cnc*) ⟶ psychomotor retardation (*rtr*) ⟶ fatigue (*ftg*), or longer chains like concentration (*cnc*) ⟶ agitation (*agt*) ⟶ guilt (*glt*) ⟶ sadness (*sad*) ⟶ suicidal thoughts (*scd*) ⟶ guilt (*glt*) ⟶ concentration (*cnc*).

Figure 19. A statistical network model and PAGs estimated from empirical data.



Note. Panel (a) shows the statistical network model (i.e., Gaussian graphical model) estimated from the empirical data on depression symptoms. Panels (b), (c), and (d) show the PAG estimated by the CCD, CCI, and FCI algorithm, respectively. *ons* = sleep onset insomnia; *mdd* = middle insomnia; *lat* = late (early morning awakening); *hyp* = hypersomnia; *sad* = sad; *dcp* = decreased appetite; *inc* = increased appetite; *wghtl* = weight loss; *wghtg* = weight gain; *cnc* = concentration impairment; *glt* = guilt; *scd* = suicidal thoughts; *anh* = anhedonia; *ftg* = fatigue; *rtr* = psychomotor retardation; *agt* = agitation.

patterns as those of direct cyclic causal relationships, readers should exercise caution. As previously demonstrated in Figure 4 from Section 3.2.1, cyclic causal relationships would be represented by undirected edges (—) in a PAG, not by patterns of directed ancestral relationships. In addition, the dotted-underlining of the variables such as fatigue (*ftg*) and guilt (*glt*) likely indicates that not all the ancestral relationships shown in this PAG can be mapped onto direct causal relationships in the corresponding DCG. Hence, the structures that appear as cycles in this PAG may not necessarily imply the same direct causal cycles in the underlying DCG. Despite the rather complicated nature of interpreting the PAG, we believe that it provides more insightful information on the causal dynamics of depression symptoms when compared to the GGM, given that the PAG provides directional information, indicating the causal flow of variables, which cannot be obtained from the GGM.

As we have outlined previously, a key limitation of the CCD algorithm is its reliance on the assumption of causal sufficiency (i.e., the absence of unobserved confounders). If this assumption is violated, the PAG will not accurately reflect the causal relationships. This limitation is addressed by the CCI and FCI algorithms, which allow for the presence of latent confounders. Figure 19(c) and Figure 19(d) show the estimated PAGs from CCI and FCI, respectively. The overall structure is reminiscent of the PAG obtained from CCD, with three distinct clusters. However, the PAGs generated by CCI and FCI contain fewer edges, resulting in more independent clusters. Also, these PAGs feature bidirected edges ($\leftrightarrow$), indicating the presence of latent confounders. The CCI algorithm, in particular, produced a significant number of bidirected edges, suggesting the presence of latent confounders between almost every variable (see Figure 19(c)). On the other hand, the PAG generated by FCI shown in Figure 19(d) has only a few bidirected edges but more circle endpoints. Both CCI and FCI identified some common bidirected edges, including $cnc \leftrightarrow ftg$, $agt \leftrightarrow cnc$, and $agt \leftrightarrow glt$, indicating that there are likely to be latent confounders present between these variables. This is consistent with previous research that has shown, for example, chronic physical illness to be a potential contributing factor to symptoms such as concentration impairment (*cnc*) and fatigue (*ftg*) (Menzies et al., 2021; Goërtz et al., 2021; de Ridder et al., 2008). In addition, both CCD and FCI predicted some common directional features, such as $rtr \ast\!\!\rightarrow ftg$, $hyp \ast\!\!\rightarrow ftg$, $scd \ast\!\!\rightarrow glt$, and $anh \ast\!\!\rightarrow dcp$. For example, $anh \ast\!\!\rightarrow dcp$ means that anhedonia may or may not cause decreased appetite, but decreased appetite does not cause anhedonia, which also partly aligns with a previous study, where appetite loss is identified as one of the features of anhedonia (Coccurello, 2019).

Overall, this example highlights that the PAGs produced by cyclic causal discovery algorithms provide unique insights into possible causal relationships that cannot be gained through network analysis alone. For instance, interpreting the CCD output (Figure 19(b)) provides some directions for investigating potential intervention targets, a type of inference often of interest in psychological network analysis (Ryan et al., 2022; Ryan & Hamaker, 2022; Rodebaugh et al., 2018; Bringmann et al., 2019). Anhedonia (*anh*) could be a promising intervention target as it (indirectly) causes many other symptoms and acts as a bridge that connects different clusters of symptoms. Guilt (*glt*) also acts as a bridge to many sub-clusters of symptoms, and so might be effective in breaking ties

and deactivating the overall system. These causal inferences are challenging to make in network analysis, which lacks directionality, making it difficult to determine the driving node. By solely looking at Figure 19(a), one might assume that fatigue should be the target node due to its numerous thick edges, but in the causal structure, this may not be accurate as it is primarily an effect node without direct causal influence on other symptoms. In addition, our findings show that the algorithms generally agree on many features, including the presence or absence of causal relations and some of the causal directions, which are in line with prior research, further enhancing their plausibility. However, the choice of which output PAG to pay most attention to depends on our degree of confidence in the assumptions underlying each algorithm. While the most informative PAG is generated by CCD, it makes strong assumptions about the absence of unobserved confounding and the linearity of causal relations. On the other hand, CCI suggests that most or all relationships between symptoms are likely influenced by latent confounding. FCI also flags the possibility of latent confounding but recovers different directed ancestral relations than CCI, as it makes use of different orientation rules. In general, FCI may be preferred over CCI when the linearity assumption does not hold, but CCI may be more appropriate in cases where the linearity assumption holds, as suggested by our simulation study.

Finally, to ensure the reliability of our findings, we conducted an additional stability analysis by running the algorithms on multiple random subsets of the data and retaining only the causal relations that were consistently discovered. This analysis confirmed that the identified causal relations from our original analysis are reliable such that they are not undermined by small variations in the data. Further details regarding this analysis can be found in Appendix I.

6 Discussion

In this paper, we studied constraint-based cyclic causal discovery algorithms in typical psychological research settings, with a focus on identifying an effective algorithm for studying the underlying cyclic causal structure. We provided a comprehensive overview and didactic treatment of cyclic causal discovery by outlining the properties of three specific algorithms: CCD, FCI, and CCI. We assessed the performance of these algorithms under varying conditions through a simulation study, and we also demonstrated their practical applicability in psychological research by applying them to empirical data. Our results suggested that the CCI algorithm generally performed well, particularly in sparse conditions, and the CCD algorithm tended to outperform the others in dense conditions. The FCI algorithm performed poorly across all conditions, mostly guessing the directions of edges, though notably, our simulation was limited to studying systems with linear causal relations. Our empirical example showed that causal discovery methods provided more detailed and richer insights into the underlying causal dynamics of depression than the statistical network model, which in fact was found to contain numerous spurious edges, rendering it unsuitable for serving as a causal skeleton (Ryan et al., 2022).

With this paper, we aimed to provide empirical researchers with guidance on selecting an appropriate cyclic causal discovery algorithm for studying causal relationships. Our findings indicate that no single algorithm is suitable for all cases, and the choice of algorithm should be based on the characteristics of the causal system of interest. If the causal system is believed to be relatively sparse, then the CCI algorithm may be preferred, as it performs well under such conditions. Conversely, if the system is considered to be comparatively dense, researchers may opt for the CCD algorithm. Also, researchers need to consider their priorities when selecting an algorithm. If avoiding incorrect edge orientations is a priority, then the CCD algorithm, which is more conservative in edge orientation, would be the preferred choice. However, if acquiring more insights into causal relationships is a priority, even if it means accepting some incorrect edge orientations, then CCI would be the more suitable option. Although the impact of violating assumptions was not entirely clear from our simulation study, researchers should assess which assumptions are relevant and critical for the particular nature of the causal system they seek to study, as each algorithm makes different assumptions. Despite the complexity involved in using cyclic causal discovery methods, we emphasize that these techniques are more informative than statistical network analysis when exploring the underlying causal structure.

However, we acknowledge that causal discovery in cyclic settings entails theoretical and practical challenges, and that much future work remains to be done to gain a better understanding of the behavior of causal discovery algorithms. In our simulation study, we restricted ourselves to a set of fixed causal structures with fixed weights instead of randomly sampling graph structures, which is a more typical approach for simulating models (Mooij et al., 2020; Strobl, 2019; Colombo et al., 2012). We chose fixed structures to prioritize the explainability and interpretability of the results, as using randomly sampled structures would not have allowed us to assess individual configurations and edges as thoroughly as we did with the fixed structures. However, our use of fixed structures limits the generalizability of our findings, as we only studied a small number of structures. This limitation also made it challenging to assess algorithm performance in scenarios with varying latent variables. Within our fixed structure, which involved a relatively small number of variables, latent variables had a limited overall influence, while the (high) density primarily impacted the algorithms' performance. Consequently, our results did not clearly reveal the anticipated differences in performance between CCD and FCI/CCI in the presence of latent variables. Although our sensitivity analysis showed some level of robustness to variations in the weights of the causal structure, our limited exploration of structural variability remains a substantial constraint in our study. One possible approach to improve the generalizability of our findings is to expand the simulation settings by incorporating a random graph structure at each iteration. However, this approach would be computationally intensive, as it requires verifying the cyclicity and equilibrium condition for each structure and iterating until satisfied. Future studies can explore the feasibility of this approach in more detail.

Furthermore, certain operational details of the algorithms were not considered in the assessment

of their performance in our simulation study. For instance, the CCD algorithm generates additional underlinings that convey more information about the allowable patterns of direct causal relationships in the equivalence class of DCGs. However, this additional information is not provided by the CCI and FCI algorithms, and to make the algorithm outputs directly comparable, we did not account for this in our simulation study. While we examined the uncertainty rate to evaluate the informativeness of the resulting PAGs for each algorithm — more circle endpoints indicate greater uncertainty in the inferred causal relations, thereby implying a less informative PAG — it is possible that the extra underlinings from the CCD algorithm may have made its output PAG significantly more informative. To further investigate the usefulness of this extra information, future studies could explore alternative performance metrics, such as directly assessing the size of the equivalence class implied by each algorithm output. This would determine the extent to which the PAGs are informative in inferring the causal structure in general, with a smaller equivalence class indicating a more informative PAG. Though conceptually promising, it could be practically challenging as the search space can expand exponentially with a larger model, and it may not be straightforward to derive the size of the equivalence class when latent confounders are involved. Nonetheless, it can be a crucial validity check on our simulation results, where we found that CCD consistently generated less informative PAGs by producing more circle endpoints compared to the other algorithms.

Our simulation was also limited in scope as we only considered Gaussian linear cases, which had the advantage of satisfying the global Markov condition and allowing the use of partial correlations to test conditional independencies. Although Gaussian linear processes are commonly assumed in psychological research (Pek et al., 2018; Beller & Baier, 2013), this assumption may be oversimplified and too strict in practice. In fact, the data used in our empirical example, which is typical of those used in psychological network research measured on a Likert scale, deviated from the Gaussian distribution.[15] Despite this, we proceeded with testing conditional independencies using partial correlations. Although the exact impact of this misspecification on our results remains unclear, it may have led to misleading findings (Baba et al., 2004), which requires further investigation. As real-world applications often involve non-linear and non-Gaussian processes, future studies should explore more general scenarios beyond linear Gaussian cases to enhance the practical applicability and to gain a more comprehensive understanding of cyclic causal discovery algorithms. Various flexible conditional independence (CI) tests have already been developed to accommodate such cases (Li & Fan, 2020; Canonne et al., 2018), including non-parametric discretization-based CI tests (Huang, 2010) and kernel-based CI tests (Zhang et al., 2012), which can be applied without assuming a functional form between the variables or the data distribution. However, these testing methods are often more complex and require larger sample sizes, making their practical application challenging. Hence, it would be an interesting extension for future research to investigate the feasibility of implementing these methods and their effectiveness in real-world scenarios.

[15] See Appendix J for the distributions of all depression symptom variables.

The network theory of psychopathology, which posits that mental disorders arise from direct causal interactions between symptoms and sustain themselves through feedback loops, has greatly contributed to the understanding of psychopathology (Borsboom, 2017). However, despite evidence demonstrating that network models are inadequate as causal discovery tools, many empirical researchers in psychology have attempted to generate causal hypotheses from estimated network models (Ryan et al., 2022). Alternatively, some researchers have tried to estimate a DAG even though the theory strongly suggests the presence of cycles/feedback loops (McNally et al., 2017; Briganti et al., 2022). In this paper, we propose a promising alternative, cyclic causal discovery methods, which are explicitly designed to recover causal structure with cycles. Although we acknowledge that these methods come with caveats, such as complicated algorithm steps, restrictive assumptions, and somewhat difficult output interpretations, we believe that these methods can offer a valid approach to understanding the causal mechanisms underlying dynamics of mental disorders. To advance the network approach to studying the dynamics of psychopathology, we argue that it is essential for researchers to gain familiarity with these methods and explore their potential applications in psychological research. We hope that the current paper will remove some barriers to entry and encourage a more widespread adoption of cyclic causal discovery methods in psychological science.

Materials

The reproducibility archive for this paper, including all R-code, results, and figures presented, is available on https://github.com/KyuriP/Discovering_CCM. The supplementary material for this paper can be found online at: https://kyurip.quarto.pub/discovering-cyclic-causal-models/.[16]

[16] Alternatively, you can access all the materials in the Open Science Framework (OSF) repository.

References

Baba, K., Shibata, R., & Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, *46*(4), 657–664. https://doi.org/10.1111/j.1467-842X.2004.00360.x

Beller, J. & Baier, D. (2013). Differential effects: Are the effects studied by psychologists really linear and homogeneous? *Europe's Journal of Psychology*, *9*(2), 378–384. https://doi.org/10.5964/ejop.v9i2.528

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Bollen, K. A. & Long, J. S. (1993). *Testing structural equation models* (Vol. 154). Sage.

Bongers, S., Blom, T., & Mooij, J. (2022). Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*. https://doi.org/https://doi.org/10.48550/arXiv.1803.08784

Bongers, S., Forré, P., Peters, J., & Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, *49*(5), 2885–2915. https://doi.org/10.1214/21-AOS2064

Bongers, S., Peters, J., Schölkopf, B., & Mooij, J. M. (2018). Theoretical aspects of cyclic structural causal models. *arXiv preprint arXiv:1611.06221*.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375

Borsboom, D. & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1–18. https://doi.org/10.1038/s43586-021-00055-w

Briganti, G., Scutari, M., & McNally, R. J. (2022). A tutorial on bayesian networks for psychopathology researchers. *Psychological Methods*. https://doi.org/10.1037/met0000479

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of abnormal psychology*, *128*(8), 892. https://doi.org/10.1037/abn0000446

Canonne, C. L., Diakonikolas, I., Kane, D. M., & Stewart, A. (2018). Testing conditional independence of discrete distributions. *arXiv preprint arXiv:1711.11560*. https://doi.org/10.48550/arXiv.1711.11560

Coccurello, R. (2019). Anhedonia in depression symptomatology: Appetite dysregulation and defective brain reward processing. *Behavioural Brain Research*, *372*, 112041. https://doi.org/10.1016/j.bbr.2019.112041

Colombo, D., Maathuis, M. H., Kalisch, M., & Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, *40*(1),

294–321. https://doi.org/10.1214/11-AOS940

Dablander, F. & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific Reports*, *9*(1), 6846. https://doi.org/10.1038/s41598-019-43033-9

Dash, D. (2005). Restructuring dynamic causal systems in equilibrium, In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics* (pp. 81–88). Proceedings of Machine Learning Research.

de Jongh, M. & Druzdzel, M. J. (2009). A comparison of structural distance measures for causal bayesian network models. *Recent Advances in Intelligent Information Systems*, 443–456.

de Ridder, D., Geenen, R., Kuijer, R., & van Middendorp, H. (2008). Psychological adjustment to chronic disease. *The Lancet*, *372*(9634), 246–255. https://doi.org/10.1016/S0140-6736(08)61078-8

Drton, M. & Maathuis, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, *4*(1), 365–393. https://doi.org/10.1146/annurev-statistics-060116-053803

Eberhardt, F., Hoyer, P., & Scheines, R. (2010). Combining experiments to discover linear cyclic models with latent variables, In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 185–192). JMLR Workshop and Conference Proceedings.

Eigenmann, M. F., Nandy, P., & Maathuis, M. H. (2017). Structure learning of linear Gaussian structural equation models with weak edges. *arXiv preprint arXiv:1707.07560*. https://doi.org/10.48550/arXiv.1707.07560

Epskamp, S., Borsboom, D., & Fried, E. I. (2018a). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S. & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617–634. https://doi.org/10.1037/met0000167

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018b). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Forré, P. & Mooij, J. M. (2017). Markov properties for graphical models with cycles and latent variables. *arXiv preprint arXiv:1710.08775*. https://doi.org/10.48550/arXiv.1710.08775

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Geiger, D. & Pearl, J. (1990). On the logic of causal models. In Shachter, R. D., Levitt, T. S., Kanal, L. N., & Lemmer, J. F. (Eds.). *Machine intelligence and pattern recognition* (Vol. 9, pp. 3–14). North-Holland. https://doi.org/10.1016/B978-0-444-88650-7.50006-8

Geiger, D., Verma, T., & Pearl, J. (1990). d-Separation: From theorems to algorithms. In Henrion, M., Shachter, R. D., Kanal, L. N., & Lemmer, J. F. (Eds.). *Machine intelligence and pattern recognition* (Vol. 10, pp. 139–148). North-Holland. https://doi.org/10.1016/B978-0-444-88738-2.50018-X

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524. https://doi.org/10.3389/fgene.2019.00524

Goërtz, Y. M. J., Braamse, A. M. J., Spruit, M. A., Ebadi, Z., Van Herck, M., Burtin, C., Peters, J. B., Lamers, F., Geerlings, S. E., Vaes, A. W., van Beers, M., & Knoop, H. (2021). Fatigue in patients with chronic disease: results from the population-based Lifelines Cohort Study. *Scientific Reports*, *11*(1), 20977. https://doi.org/10.1038/s41598-021-00337-z

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2021). Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory. *Multivariate Behavioral Research*, *56*(2), 199–223. https://doi.org/10.1080/00273171.2019.1640103

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). The sum of all fears: Comparing networks based on symptom sum-scores. *Psychological Methods*, *27*(6), 1061–1068. https://doi.org/10.1037/met0000418

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*. https://doi.org/10.1037/met0000303

Haslbeck, J. M. B. & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, *50*(2), 853–861. https://doi.org/10.3758/s13428-017-0910-x

Huang, M., Sun, Y., & White, H. (2016). A flexible nonparametric test for conditional independence. *Econometric Theory*, *32*(6), 1434–1482. https://doi.org/10.1017/S0266466615000286

Huang, T. M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, *38*(4). https://doi.org/10.1214/09-AOS770

Hyttinen, A., Hoyer, P. O., Eberhardt, F., & Jarvisalo, M. (2013). Discovering cyclic causal models with latent variables: A general SAT-based procedure. *arXiv preprint arXiv:1309.6836*. https://doi.org/10.48550/arXiv.1309.6836

Iwasaki, Y. & Simon, H. A. (1994). Causality and model abstraction. *Artificial intelligence*, *67*(1), 143–194. https://doi.org/https://doi.org/10.1016/0004-3702(94)90014-0

Kalisch, M. & Buehlmann, P. (2005). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *arXiv preprint arXiv:math/0510436*. https://doi.org/10.48550/arXiv.math/0510436

Kossakowski, J., Waldorp, L. J., & van der Maas, H. L. J. (2021). The search for causality: A comparison of different techniques for causal inference graphs. *Psychological Methods*, *26*(6), 719–742. https://doi.org/10.1037/met0000390

Lauritzen, S. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Lauritzen, S. (2001). Causal inference from graphical models. *Monographs on Statistics and Applied Probability*, *87*, 63–108.

Lawrance, A. J. (1976). On conditional and partial correlation. *The American Statistician*, *30*(3), 146–149. https://doi.org/10.2307/2683864

Li, C. & Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, *12*(3), e1489. https://doi.org/10.1002/wics.1489

Magliacane, S., Claassen, T., & Mooij, J. M. (2017). Ancestral causal inference. *arXiv preprint arXiv:1606.07035*. https://doi.org/10.48550/arXiv.1606.07035

Malinsky, D. & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470. https://doi.org/10.1111/phc3.12470

McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (2017). Co-morbid obsessive-compulsive disorder and depression: a Bayesian network approach. *Psychological Medicine*, *47*(7), 1204–1214. https://doi.org/10.1017/S0033291716003287

Menzies, V., Kelly, D. L., Yang, G. S., Starkweather, A., & Lyon, D. E. (2021). A systematic review of the association between fatigue and cognition in chronic noncommunicable diseases. *Chronic illness*, *17*(2), 129–150. https://doi.org/10.1177/1742395319836472

Mooij, J. M. & Claassen, T. (2020). Constraint-based causal discovery using partial ancestral graphs in the presence of cycles, In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence* (pp. 1159–1168). Proceedings of Machine Learning Research.

Mooij, J. M., Janzing, D., & Schölkopf, B. (2013). From Ordinary Differential Equations to Structural Causal Models: The Deterministic Case, In *Uncertainty in Artificial Intelligence* (pp. 440–449). https://doi.org/https://doi.org/10.48550/arXiv.1304.7920

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*. https://doi.org/10.48550/arXiv.1611.10351

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02104

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(5), 947–1012. https://doi.org/10.1111/rssb.12167

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Richardson, T. (1996a). *Discovering cyclic causal structure* (Publication No. CMU-PHIL-68). Carnegie Mellon University, Department of Philosophy.

Richardson, T. (1996b). A discovery algorithm for directed cyclic graphs, In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence* (pp. 454–461). Morgan Kaufmann.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, *30*(1), 145–157. https://doi.org/10.1111/1467-9469.00323

Richardson, T. & Spirtes, P. (1999). Automated discovery of linear feedback models. In Glymour, C. & Cooper, G. F. (Eds.). *Computation, causation, and discovery* (pp. 253–302). MIT Press. https://doi.org/10.7551/mitpress/2006.001.0001

Richardson, T. & Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, *30*(4), 962–1030. https://doi.org/10.1214/aos/1031689015

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Rodebaugh, T. L., Tonge, N. A., Piccirillo, M. L., Fried, E., Horenstein, A., Morrison, A. S., Goldin, P., Gross, J. J., Lim, M. H., Fernandez, K. C., et al. (2018). Does centrality in a cross-sectional network suggest intervention targets for social anxiety disorder? *Journal of consulting and clinical psychology*, *86*(10), 831. https://doi.org/10.1037/ccp0000336

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095827

Rothenhäusler, D., Heinze, C., Peters, J., & Meinshausen, N. (2015). backShift: Learning causal cyclic graphs from unknown shift interventions. *arXiv preprint arXiv:1506.02494*. https://doi.org/10.48550/arXiv.1506.02494

Ryan, O., Bringmann, L. F., & Schuurman, N. K. (2022). The challenge of generating causal hypotheses using network models. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(6), 953–970. https://doi.org/10.1080/10705511.2022.2056039

Ryan, O. & Dablander, F. (2022). Equilibrium causal models: Connecting dynamical systems modeling and cross-sectional data analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/q4d9g

Ryan, O. & Hamaker, E. L. (2022). Time to intervene: A continuous-time approach to network analysis and centrality. *Psychometrika*, *87*(1), 214–252. https://doi.org/10.1007/s11336-021-09767-0

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, *33*(1), 65–117. https://doi.org/10.1207/s15327906mbr3301_3

Spirtes, P. (1994). *Conditional independence in directed cyclic graphical models for feedback* (Publication No. CMU-PHIL-53). Carnegie Mellon University, Department of Philosophy.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models, In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 491–498). Morgan Kaufmann.

Spirtes, P. & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, *9*(1), 62–72. https://doi.org/10.1177/089443939100900106

Spirtes, P., Glymour, C., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., Meek, C., & Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias, In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. UAI'95 (pp. 499–506). Morgan Kaufmann.

Strobl, E. V. (2019). A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, *8*(1), 33–56. https://doi.org/10.1007/s41060-018-0158-2

Strobl, E. V., Spirtes, P. L., & Visweswaran, S. (2017). Estimating and controlling the false discovery rate for the PC algorithm using edge-specific p-values. *arXiv preprint arXiv:1607.03975*. https://doi.org/10.48550/arXiv.1607.03975

Strotz, R. H. & Wold, H. O. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis (part i of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, *28*(2), 417–427. https://doi.org/https://doi.org/10.2307/1907731

Tennant, P. W., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., et al. (2021). Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, *50*(2), 620–632.

Tian, J., Paz, A., & Pearl, J. (1998). *Finding minimal d-separators* (Publication No. R-254). University of California, Los Angeles, Department of Computer Science.

Versteeg, P., Mooij, J., & Zhang, C. (2022). Local constraint-based causal discovery under selection bias, In *Proceedings of the First Conference on Causal Learning and Reasoning* (pp. 840–860). Proceedings of Machine Learning Research.

Weinberger, N. (2021). Intervening and letting go: On the adequacy of equilibrium causal models. *Erkenntnis*, 1–25. https://doi.org/https://doi.org/10.1007/s10670-021-00463-0

Wittenborn, A. K., Rahmandad, H., Rick, J., & Hosseinichimeh, N. (2016). Depression as a systemic syndrome: Mapping the feedback loops of major depressive disorder. *Psychological Medicine*, *46*(3), 551–562. https://doi.org/10.1017/S0033291715002044

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, *172*(16), 1873–1896. https://doi.org/10.1016/j.artint.2008.08.001

Zhang, J. & Spirtes, P. (2005). *A characterization of Markov equivalence classes for ancestral graphical models* (Publication No. CMU-PHIL-168). Carnegie Mellon University, Department of Philosophy.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*. https://doi.org/10.48550/arXiv.1202.3775

Appendix A CCD Details

Algorithm 1 Cyclic Causal Discovery (CCD)

Input: A conditional independent oracle for a distribution $\mathcal{P}$, satisfying global directed Markov property and faithfulness conditions with respect to a directed graph $\mathcal{G}$ with vertex set $\mathcal{V}$.

Output: A PAG Ψ for the Markov equivalence class of DCGs, $Equiv(\mathcal{G})$.

Step 1. Form a complete graph (Ψ) with the edge $\circ\!\!-\!\!\circ$ between every pair of vertices in $\mathcal{V}$.
$n = 0$
repeat
  repeat
    Select an ordered pair of variables X and Y that are adjacent in Ψ such that the number of vertices in **Adjacent**$(\Psi, X)\backslash\{Y\} \geq n$, and select a subset $\mathcal{S}$ of **Adjacent**$(\Psi, X)\backslash\{Y\}$ with n vertices.
    If $X \perp\!\!\!\perp Y \mid \mathcal{S}$, then delete the edge $X\circ\!\!-\!\!\circ Y$ and record $\mathcal{S}$ in **Sepset**$\langle X, Y\rangle$ and **Sepset**$\langle X, Y\rangle$.
  until all pairs of adjacent variables X and Y such that the number of vertices in **Adjacent**$(\Psi, X)\backslash\{Y\} \geq n$ and all sets $\mathcal{S}$ such that the number of vertices in $\mathcal{S} = n$ have been tested.
  $n = n + 1$;
until for all ordered pairs of adjacent vertices X and Y, **Adjacent**$(\Psi, X)\backslash\{Y\} < n$.
Step 2. For each triple of vertices A, B, C such that each of the pair of A, B and the pair B, C are adjacent in Ψ but the pair A, C are not adjacent in Ψ, then:
(i) orient $A\ast\!\!-\!\!\ast B\ast\!\!-\!\!\ast C$ as $A\rightarrow B\leftarrow C$ *iff* $B \notin$ **Sepset**$\langle A, C\rangle$.
(ii) orient $A\ast\!\!-\!\!\ast B\ast\!\!-\!\!\ast C$ as $A\ast\!\!-\!\!\ast \underline{B}\ast\!\!-\!\!\ast C$ *iff* $B \in$ **Sepset**$\langle A, C\rangle$.
Step 3. For each triple of vertices A, X, Y in Ψ such that (i) A is not adjacent to X or Y, (ii) X and Y are adjacent, (iii) $X \notin$ **Sepset**$\langle A, Y\rangle$, then orient $X\ast\!\!-\!\!\ast Y$ as $X\leftarrow Y$ if $A \not\perp\!\!\!\perp X \mid$ **Sepset**$\langle A, Y\rangle$.
Step 4. For each vertex V in Ψ form the following set: $X \in$ **Local**(Ψ, V) *iff* X is adjacent to V in Ψ, or there is a vertex Y such that $X\rightarrow Y\leftarrow V$ in Ψ.
$m = 0$
repeat
  repeat
    Select an ordered triple $\langle A, B, C\rangle$ such that $A\rightarrow B\leftarrow C$, A and C are not adjacent, and **Local**$(\Psi, A)\backslash\{B, C\}$ has $\geq m$ vertices.
    Select a set $T \subseteq$ **Local**$(\Psi, A)\backslash\{B, C\}$ with m vertices. If $A \perp\!\!\!\perp C \mid T \cup \{B\}$, then orient $A\rightarrow B\leftarrow C$ as $A\rightarrow \underline{B}\leftarrow C$ and record $T \cup \{B\}$ in **Supset**$\langle A, B, V\rangle$.

until for all triples such that $A \rightarrow B \leftarrow C$ (not $A \rightarrow \underset{\cdot}{\mathrm{B}} \leftarrow C$), A and C are not adjacent, **Local**$(\Psi, A) \backslash \{B\}$ has $\geq m$ vertices, every subset T with m vertices has been considered.

$m = m + 1$;

until all ordered triples $\langle A, B, C \rangle$ such that $A \rightarrow B \leftarrow C$, A and C are not adjacent, are such that **Local**$(\Psi, A) \backslash \{B\}$ have $< m$ vertices.

Step 5. If there is a quadruple $\langle A, B, C, D \rangle$ of distinct vertices in Ψ such that:

(i) $A \rightarrow \underset{\cdot}{\mathrm{B}} \leftarrow C$,

(ii) $A \rightarrow D \leftarrow C$ or $A \rightarrow \underset{\cdot}{\mathrm{D}} \leftarrow C$,

(iii) B and D are adjacent,

then orient $B \ast\!\!-\!\!\ast D$ ad $B \rightarrow D$ in Ψ if $D \notin$ **Subset**$\langle A, B, C \rangle$. Else orient $B \ast\!\!-\!\!\ast D$ as $B \ast\!\!- D$ in Ψ.

Step 6. For each quadruple A, B, C, D in Ψ of distinct vertices such that:

(i) D is not adjacent to both A and C,

(ii) $A \rightarrow \underset{\cdot}{\mathrm{B}} \leftarrow C$,

if $A \not\perp\!\!\!\perp D \mid$ **Supset**$\langle A, B, C \rangle \cup \{D\}$, then orient $B \ast\!\!-\!\!\ast D$ as $B \rightarrow D$ in Ψ.

Appendix B FCI Details

Algorithm 2 Fast Causal Inference (FCI)

Input: A conditional independent oracle for a distribution $\mathcal{P}$, satisfying global directed Markov property and faithfulness conditions with respect to a directed graph $\mathcal{G}$ with vertex set $\mathcal{V}$.

Output: A PAG $\hat{\mathcal{G}}'$ for the Markov equivalence class of DMGs, $Equiv(\mathcal{G})$.

Step 1. Form the complete undirected graph Q on the vertex set $\mathcal{V}$.

$n = 0$

repeat

  repeat

    Select an ordered pair of variables X and Y that are adjacent in Q such that the number of vertices in **Adjacent**$(Q, X) \backslash \{Y\} \geq n$, and select a subset $\mathcal{S}$ of **Adjacent**$(Q, X) \backslash \{Y\}$ with n vertices.

    If $X \perp\!\!\!\perp Y \mid \mathcal{S}$, then delete the edge $X \circ\!\!-\!\!\circ Y$ and record $\mathcal{S}$ in **Sepset**$\langle X, Y \rangle$ and **Sepset**$\langle X, Y \rangle$.

  until all pairs of adjacent variables X and Y such that the number of vertices in **Adjacent**$(Q, X) \backslash \{Y\} \geq n$ and all sets $\mathcal{S}$ such that the number of vertices in $\mathcal{S} = n$ have been tested.

  $n = n + 1$;

7: **until** for all ordered pairs of adjacent vertices X and Y, $\mathbf{Adjacent}(Q, X)\backslash\{Y\} < n$.
8: ***Step 2.*** Let Q' be the undirected graph resulting from step 1.
9: For each triplet $\langle A, B, C\rangle$ such that each of the pair of A, B and the pair B, C are adjacent in Q' but the pair A, C are not adjacent in Q', then orient $A{*\!\!-\!\!*}B{*\!\!-\!\!*}C$ as $A{*\!\!\rightarrow}B{\leftarrow\!\!*}C$ *iff* $B \notin \mathbf{Sepset}\langle A, B\rangle$.
10: ***Step 3.*** For each pair of variables A and B adjacent in Q', if A and B are d-separated given any subset S of $\mathbf{Possible\text{-}d\text{-}sepset}\langle A, B\rangle)\backslash\{A, B\}$ or any subset S of $\mathbf{Possible\text{-}d\text{-}sepset}\langle B, A\rangle)\backslash\{A, B\}$ in Q', then remove the edge between A and B, and record $\mathcal{S}$ in $\mathbf{Sepset}\langle A, B\rangle$ and $\mathbf{Sepset}\langle B, A\rangle$.
11: ***Step 4.*** Execute the following orientation rules iteratively until none applies:
(i) If $A{*\!\!\rightarrow}B{\circ\!\!-\!\!*}C$, and A and C are not adjacent, then orient the triple as $A{*\!\!\rightarrow}B{\rightarrow}C$.
(ii) If $A{\rightarrow}B{*\!\!\rightarrow}C$ or $A{*\!\!\rightarrow}B{\rightarrow}C$, and $A{*\!\!-\!\!\circ}C$, then orient $A{*\!\!-\!\!\circ}C$ as $A{*\!\!\rightarrow}C$.
(iii) If $A{*\!\!\rightarrow}B{\leftarrow\!\!*}C$, $A{*\!\!-\!\!\circ}D{\circ\!\!-\!\!*}C$, A and C are not adjacent, and $D{*\!\!-\!\!\circ}B$, then orient $D{*\!\!-\!\!\circ}B$ ad $D{*\!\!\rightarrow}B$.
(iv) If $u = \langle D, ..., A, B, C\rangle$ is a discriminating path between D and C for B, and $B{\circ\!\!-\!\!*}C$; then if $B \in \mathbf{Sepset}\langle D, C\rangle$, orient $B{\circ\!\!-\!\!*}C$ as $B{\rightarrow}C$; otherwise orient the triple $\langle A, B, C\rangle$ as $A{\leftrightarrow}B{\leftrightarrow}C$.
(v) For every (remaining) $A{\circ\!\!-\!\!\circ}B$, if there is an uncovered circle path $p = \langle A, C, ..., D, B\rangle$ between A and B such that A, D are not adjacent and B, C are not adjacent, then orient $A{\circ\!\!-\!\!\circ}B$ and every edge on p as undirected edges (—).
(vi) If A — $B{\circ\!\!-\!\!*}C$ (A and C may or may not be adjacent), then orient $B{\circ\!\!-\!\!*}C$ as $B{-\!\!*}C$.
(vii) If $A{-\!\!\circ}B{\circ\!\!-\!\!*}C$ and A, C are not adjacent, then orient $B{\circ\!\!-\!\!*}C$ ad $B{-\!\!*}C$.
(viii) If $A{\rightarrow}B{\rightarrow}C$ or $A{-\!\!\circ}B{\rightarrow}C$, and $A{\circ\!\!\rightarrow}C$, orient $A{\circ\!\!\rightarrow}C$ as $A{\rightarrow}C$.
(ix) If $A{\circ\!\!\rightarrow}C$, and $p = \langle A, B, D, ..., C\rangle$ is an uncovered potentially directed path from A to C such that C and B are not adjacent, then orient $A{\circ\!\!\rightarrow}C$ as $A{\rightarrow}C$.
(x) Suppose $A{\circ\!\!\rightarrow}C$, $B{\rightarrow}C{\leftarrow}D$, p_1 is an uncovered potentially directed path from A to B, and p_2 is an uncovered potentially directed path from A to D. Let m be the vertex adjacent to A on p_1 (m could be B), and w be the vertex adjacent to A on p_2 (w could be D). If m and w are distinct, and are not adjacent, then orient $A{\circ\!\!\rightarrow}D$ as $A{\rightarrow}D$.

Appendix C CCI Details

Algorithm 3 Cyclic Causal Inference (CCI)

Input: A conditional independent oracle for a distribution $\mathcal{P}$, satisfying global directed Markov property and faithfulness conditions *w.r.t.* a directed graph $\mathcal{G}$ with vertex set $\mathcal{V}$. ($\mathcal{V} = \mathbf{O} \cup \mathbf{L} \cup \mathbf{S}$), where **O**, **L**, and **S** refer to the sets of observable, latent, and selection variables, respectively.

Output: A PAG $\hat{\mathcal{G}}'$ for the Markov equivalence class of DMGs, $Equiv(\mathcal{G})$.

Step 1. Run FCI's skeleton discovery procedure.

Step 2. Run FCI's collider structure (v-structure) orientation procedure.

Step 3. For any triplet $\langle O_i, O_k, O_j \rangle$, such that we have $O_k \circ\!\!-\!\!* O_i$, if $O_i \perp\!\!\!\perp O_j \mid \mathbf{Sepset}\langle O_i, O_j \rangle \cup \mathbf{S}$, where $\mathbf{Sepset}\langle O_i, O_j \rangle$ is a separating set discovered in step 1, $O_k \notin \mathbf{Sepset}\langle O_i, O_j \rangle$, $O_i \not\perp\!\!\!\perp O_k \mid \mathbf{Sepset}\langle O_i, O_j \rangle \cup \mathbf{S}$ and $O_j \not\perp\!\!\!\perp O_k \mid \mathbf{Sepset}\langle O_i, O_j \rangle \cup \mathbf{S}$, then orient $O_k \circ\!\!-\!\!* O_i$ as $O_k \leftarrow\!\!* O_i$.

Step 4. Find additional non-minimal d-separating sets.

$m = 0$

repeat

  repeat

    Select an ordered triplet $\langle O_i, O_j, O_k \rangle$ with the collider structure $O_i *\!\!\rightarrow O_j \leftarrow\!\!* O_k$ such that $|\mathbf{PD\text{-}Sep}(O_i)| \geq m$

    repeat

      Select a subset
      $\mathbf{W} \subseteq \mathbf{PD\text{-}Sep}(O_i) \backslash \{\mathbf{Sepset}\langle O_i, O_k \rangle \cup \{O_j, O_k\}\}$ with m vertices
      $\mathbf{T} = \mathbf{W} \cup \mathbf{Sepset}\langle O_i, O_k \rangle \cup O_j$
      if O_i and O_k are d-separated given $\mathbf{T} \cup \mathbf{S}$, then record the set **T** in $\mathbf{Supset}\langle O_i, O_j, O_k \rangle$

    until all subsets
      $\mathbf{W} \subseteq \mathbf{PD\text{-}Sep}(O_i) \backslash \{\mathbf{Sepset}\langle O_i, O_k \rangle \cup \{O_j, O_k\}\}$ have been considered or
      a d-separating set of O_i and O_k has been recorded in $\mathbf{Supset}\langle O_i, O_j, O_k \rangle$;

  until all triplets $\langle O_i, O_j, O_k \rangle$ with the collider structure $O_i *\!\!\rightarrow O_j \leftarrow\!\!* O_k$ and $|\mathbf{PD\text{-}Sep}(O_i)| \geq m$ have been selected;

until all ordered triplets $\langle O_i, O_j, O_k \rangle$ with the collider structure $O_i *\!\!\rightarrow O_j \leftarrow\!\!* O_k$ have $|\mathbf{PD\text{-}Sep}(O_i)| < m$;

Step 5. Find all quadruples of vertices $\langle O_i, O_j, O_k, O_l \rangle$ such that O_i and O_k are non-adjacent, $O_i *\!\!\rightarrow O_l \leftarrow\!\!* O_k$, and $O_i \perp\!\!\!\perp O_k \mid \mathbf{W} \cup \mathbf{S}$ with $O_j \in \mathbf{W}$ and $\mathbf{W} \subseteq \mathbf{O} \backslash \{O_i, O_k\}$. If $O_l \notin \mathbf{W} = \mathbf{Sepset}\langle O_i, O_k \rangle$ as discovered in step 2, then orient $O_j *\!\!-\!\!\circ O_l$ as $O_j *\!\!\rightarrow O_l$. If we also have $O_i *\!\!\rightarrow O_j \leftarrow\!\!* O_k$ and $O_l \in \mathbf{W} = \mathbf{Supset}\langle O_i, O_j, O_k \rangle$ as discovered in step 4, then orient $O_j *\!\!-\!\!\circ O_l$ as $O_j *\!\!-\!\!* O_l$.

15: ***Step 6.*** For any two vertices O_i and O_k, if we have $O_i \perp\!\!\!\perp O_k \mid \mathbf{W} \cup \mathbf{S}$ for some $\mathbf{W} \subseteq \mathbf{O} \backslash \{O_i, O_k\}$ discovered in step 1 or step 4 with $O_j \in \mathbf{W}$ but we have $O_i \not\perp\!\!\!\perp O_k \mid O_l \cup \mathbf{W} \cup \mathbf{S}$, then orient $O_l \circ\!\!-\!\!* O_j$ as $O_l \leftarrow\!\!* O_j$.

16: ***Step 7.*** Execute orientation rules until no more endpoints can be oriented. The orientation rules are as follows:

(i) If we have $O_i *\!\!\rightarrow O_j \circ\!\!-\!\!* O_k$ with O_i and O_k non-adjacent, then orient $O_j \circ\!\!-\!\!* O_k$ as $O_j \!-\!\!* O_k$. Furthermore, if $O_i *\!\!\rightarrow O_j$ is not potentially 2-triangulated *w.r.t.* O_k, then orient $O_j \!-\!\circ O_k$ as $O_j \rightarrow O_k$.

(ii) If we have $O_i \!-\!\!* O_j \circ\!\!-\!\!* O_k$ with O_i and O_k non-adjacent, and $O_j \circ\!\!- O_k$ is not potentially 2-triangulated *w.r.t.* O_i, then orient $O_j \circ\!\!-\!\!* O_k$ as $O_j \!-\!\!* O_k$.

(iii) Suppose we have $O_i *\!\!\rightarrow O_j \!-\! O_k$ with O_i and O_k non-adjacent, and $O_i *\!\!\rightarrow O_j$ is potentially 2-triangulated *w.r.t.* O_k. If $O_i *\!\!\rightarrow O_j$ can be potentially 2-triangulated *w.r.t.* O_k using only one vertex O_l in the triangle involve $\{O_i, O_j, O_l\}$, then orient $O_i *\!\!-\!\circ O_l$ as $O_i *\!\!\rightarrow O_l$, $O_j \circ\!\!-\!\!* O_l$ as $O_j *\!\!\rightarrow O_l$ and/or $O_j *\!\!-\!\circ O_l$ as $O_j *\!\!- O_l$. Next, if there exists only one potentially undirected path $\prod_{O_l, O_k}$ between O_l and O_k, then substitute all circle endpoints on $\prod_{O_l, O_k}$ with tail endpoints (−).

(iv) If $O_i *\!\!\rightarrow O_j \!-\!\!* O_k$, there exists a path $\prod = \langle O_k, ..., O_i \rangle$ with at least $n \geq 3$ vertices such that we have $O_h \!-\!\!* O_{h+1}$ for all $1 \leq i \leq n-1$, and we have $O_1 \circ\!\!-\!\!* O_n$, then orient $O_1 \circ\!\!-\!\!* O_n$ as $O_1 \!-\!\!* O_n$.

(v) If we have the sequence of vertices $\langle O_1, ..., O_n \rangle$ such that $O_i \!-\!\!* O_{i+1}$ with $1 \leq i \leq n-1$, and we have $O_1 \circ\!\!-\!\!* O_n$, then orient $O_1 \circ\!\!-\!\!* O_n$ as $O_1 \!-\!\!* O_n$.

(vi) If we have $O_k *\!\!-\!\circ O_i$, there exists a non-potentially 2-triangulated path $\prod = \langle O_i, O_j, O_l, ..., O_k \rangle$ such that $O_k *\!\!-\!\circ O_i$ is not potentially 2-triangulated *w.r.t.* O_j, and $O_j *\!\!-\!\!* O_i *\!\!-\!\!* O_k$ is an unshielded non-v-structure, then orient $O_k *\!\!-\!\circ O_i$ as $O_k *\!\!- O_i$.

(vii) Suppose we have $O_i \circ\!\!- O_k$, $O_i \!-\!\!* O_k *\!\!- O_l$, a non-potentially 2-triangulated path $\prod_1$ from O_i to O_j, and a non-potentially 2-triangulated path $\prod_2$ from O_i to O_l. Let O_m be a vertex adjacent to O_i on $\prod_1$, and let O_n be the vertex adjacent to O_i on $\prod_2$. If further $O_m *\!\!-\!\!* O_i *\!\!-\!\!* O_n$ is an unshielded non-v-structure and $O_i \circ\!\!-\!\!* O_k$ is not potentially 2-triangulated *w.r.t.* both O_n and O_m, then orient $O_i \circ\!\!-\!\!* O_k$ as $O_i \!-\!\!* O_k$.

Appendix D Coefficient Matrices (B) Used in Simulation

In this appendix, we present the weight matrices **B** that were used in the simulation. There are eight matrices in total, with each matrix corresponding to a specific condition (*5p sparse* = 5-variable low density, *5p dense* = 5-variable high density, *5p sparse LC* = 5-variable low density with a latent confounder, *5p dense LC* = 5-variable high density with a latent confounder, *10p sparse* = 10-variable low density, *10p dense* = 10-variable high density, *10p sparse LC* = 10-variable low density with latent confounders, *10p dense LC* = 10-variable high density with latent confounders).

$$\mathbf{B}_{5p\ sparse} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0.7 & 0 & 0 & 1.5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{B}_{5p\ dense} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0.7 & 0 & 0 & 1.5 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{B}_{5p\ sparse\ LC} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & LC_1 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.8 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \qquad \mathbf{B}_{5p\ dense\ LC} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & LC_1 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.8 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 1.5 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{B}_{10p\ sparse} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{B}_{10p\ dense} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{B}_{10p\ sparse\ LC} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} & LC_1 & LC_2 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$\mathbf{B}_{10p\ dense\ LC} = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} & LC_1 & LC_2 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.8 & 0.5 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Appendix E True Ancestral Graph Derivation

This appendix provides a more detailed description of how to obtain the true ancestral graph from a DCG. The process consists of two steps. The first step step involves checking whether there exists an *inducing path* connecting a pair of observed vertices. If such a path exists, we make them adjacent by adding an edge between them. An inducing path is defined as follows.

Definition 1 (Inducing path) *An undirected path* $\mathbf{U}$ *between observed vertices* O_i *and* O_j *is called an inducing path if and only if every observed vertex on* $\mathbf{U}$ *is a collider except for the endpoints* O_i *and* O_j*, every collider on* $\mathbf{U}$ *is an ancestor of either* O_i *and* O_j*, and every non-collider on* $\mathbf{U}$ *except for the endpoints is latent variable* (Spirtes et al., 2000).

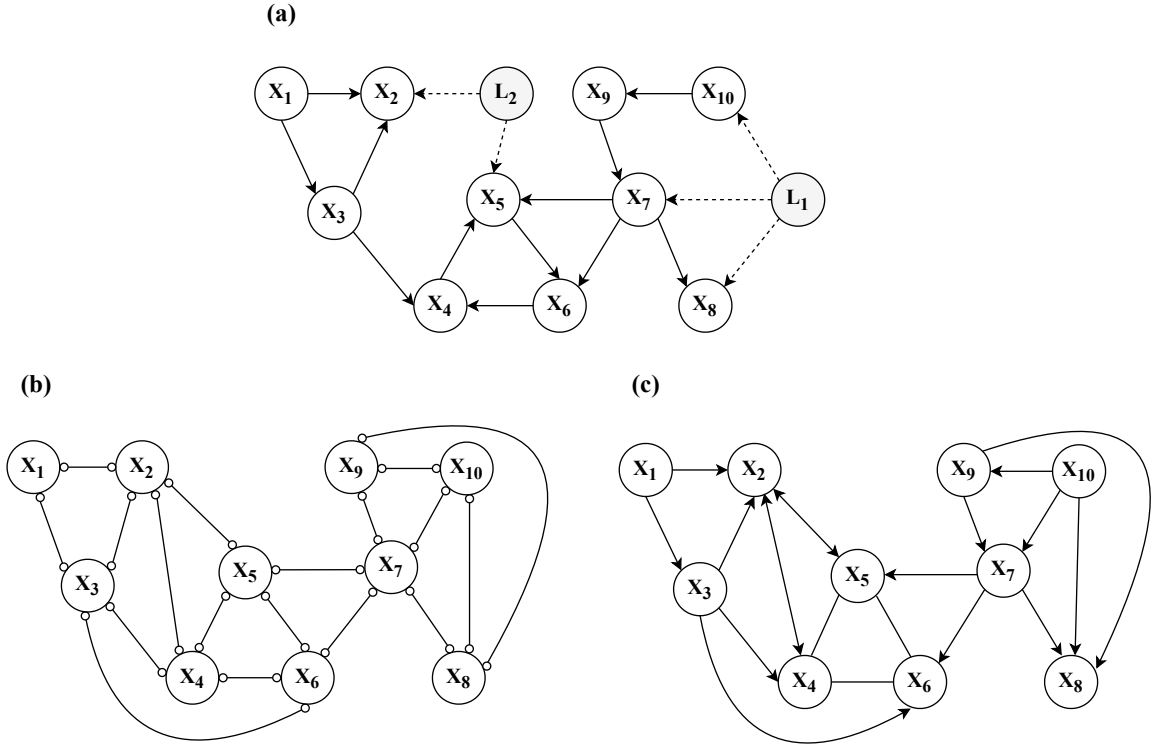
In the second step, we place either an arrow head (>) or tail (–) for every edge-endpoint based on the ancestral relationships between the corresponding vertices, as explained in Section 3.2.1.

To illustrate the procedure, we will use an example. Consider the DCG presented in Figure E1(a). By applying the previously defined concept of inducing path, we can identify the following inducing paths between the pairs of variables:

$$\begin{cases} X_2 \sim X_5 : X_2 \leftarrow L_2 \rightarrow X_5 \\ X_2 \sim X_4 : X_2 \leftarrow L_2 \rightarrow X_5 \leftarrow X_4 \\ X_3 \sim X_6 : X_3 \rightarrow X_4 \leftarrow X_6 \\ X_5 \sim X_7 : X_5 \rightarrow X_6 \leftarrow X_7 \\ X_7 \sim X_{10} : X_7 \leftarrow L_1 \rightarrow X_{10} \\ X_8 \sim X_{10} : X_8 \leftarrow L_1 \rightarrow X_{10} \\ X_9 \sim X_8 : X_9 \rightarrow X_7 \leftarrow L_1 \rightarrow X_8. \end{cases}$$

Based on the list of identified inducing paths above, we can construct the ancestral skeleton, as shown in Figure E1(b). Then, the true ancestral graph can be derived by examining the ancestral relationships between pairs of vertices. For example, orient $X_1 \rightarrow X_2$ when X_1 is an ancestor of X_2, orient $X_2 \leftrightarrow X_5$ when X_2 and X_5 are not ancestors of each other, and orient X_4 — X_5 when X_4 and X_5 are ancestors of each other. Recall that bidirected edges ($\leftrightarrow$) denote mutual non-ancestral relations between vertices indicating the presence of latent confounders, and undirected edges (—) indicate mutual ancestral relations implying cyclic relations between vertices. Once all the edge endpoints have been oriented, we can obtain the true ancestral graph, as shown in Figure E1(c).

Figure E1. Derivation of the true ancestral graph.



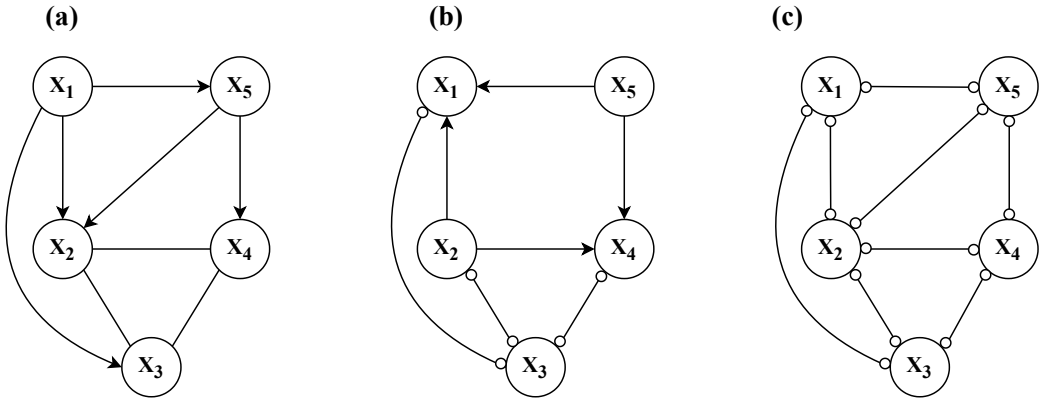
Note. Panel (a) shows an example DCG with two latent confounders L_1 and L_2. Panel (b) depicts the ancestral skeleton of the DCG shown in panel (a). Panel (c) depicts the true ancestral graph of the DCG shown in panel (a).

Appendix F 5-variable Dense Case Examination

In Section 4.4, we reported an unexpected finding that the performance of the algorithms deteriorated (e.g., precision and recall decreased, while SHD increased) as the sample size N increased in the 5-variable dense conditions, which was contrary to our expectations. In this appendix, we look into this issue in more detail by examining the 5-variable dense condition without a latent confounder scenario using the CCD algorithm. Note that the following explanation applies to the other two algorithms as well.

Figure F1(b) and Figure F1(c) present the PAGs generated by the CCD algorithm using a relatively small sample size of 500 and a large sample size of 3000, respectively. We can see that with the small sample size, the algorithm failed to detect the edge between X_2 and X_5, while with the large sample size, it did detect the edge. Upon the detection of an additional edge between X_2 and X_5, however, the resulting graph became very dense, with almost every vertex connected to every other vertex. As a consequence, the algorithm failed to orient any edges, as shown in Figure F1(c).

Figure F1. Graphs from the 5-variable dense condition without a latent confounder.

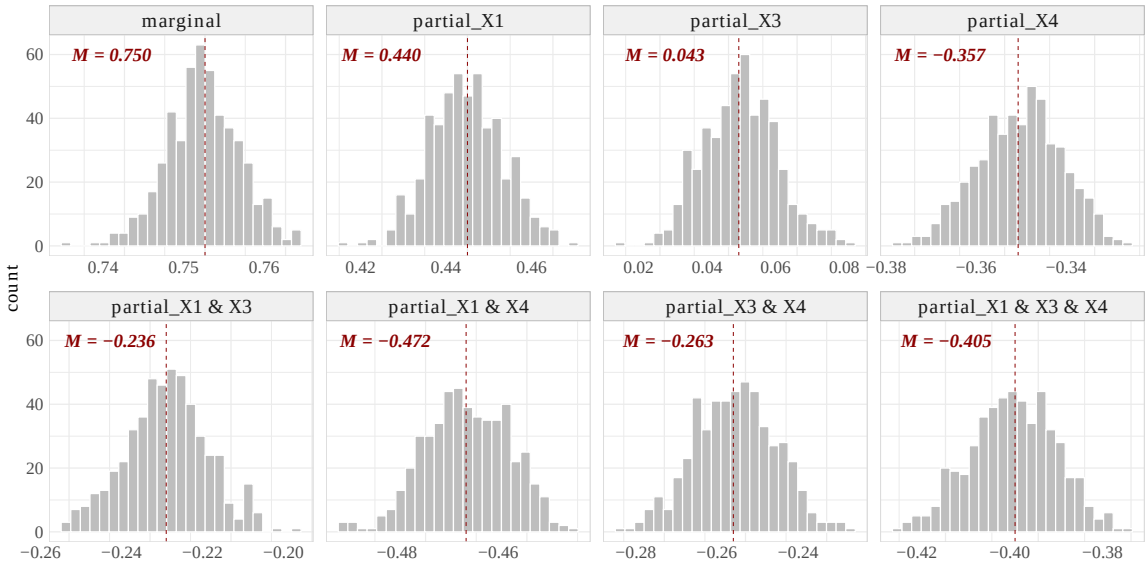


Note. Panel (a) shows the true ancestral graph for the 5-variable dense condition without a latent confounder. Panels (b) and (c) show the partial ancestral graphs (PAGs) estimated by CCD, using a relatively small sample size ($N = 500$) and a large sample size ($N = 3000$), respectively. The PAGs estimated by the FCI and CCI algorithms also exhibit a similar pattern.

Before the edge was detected, the algorithm was capable of performing orientation, and correctly oriented some edges by identifying the collider X_4, despite some wrongly oriented edges like the incoming arrows towards X_1. This explains the decrease in SHD and the slight increase in precision and recall with small sample sizes. Given the true graph shown in Figure F1(a), we can actually compute the SHD. The PAG with the small sample size in Figure F1(b) has an SHD value of 13, while the PAG with the large sample size in Figure F1(b) has an SHD value of 16.

To investigate why the algorithm failed to detect the edge with the small sample size, we examined all partial correlations between X_2 and X_5 across 500 simulations, using a large sample size of $N = 10000$. The resulting distributions of partial correlations are shown in Figure F2. Notably, the results reveal a very small partial correlation between X_2 and X_5 when conditioned on X_3 ($\rho_{X_2,X_5|X_3} = 0.04$). Also, we examined the results of conditional independence tests for all possible independence patterns between X_2 and X_5, which are summarized in Table F1. The null hypothesis of independence was not rejected in more than half of the cases until the sample size reached 2000, which led to the omission of the edge between X_2 and X_5. However, for sample sizes greater than 2000, the null hypothesis of independence was rejected, indicating that the algorithm was able to detect the edge between X_2 and X_5. Consequently, the resulting graph became too dense to orient any edges, leading to a fully undirected ancestral graph, as shown in Figure F1(c).

Figure F2. Distributions of marginal & partial correlations between X_2 and X_5.



Note. The figure displays the distributions of the estimated marginal and partial correlations between X_2 and X_5 across 500 simulations based on a sample size of $N = 10000$. The top left panel displays the distribution of marginal correlations between X_2 and X_5, which reflects their correlation without controlling for other variables. The remaining panels represent the distributions of partial correlations between X_2 and X_5 while controlling for one or more variables, as indicated by the labels. *M* denotes the mean value of the marginal/partial correlations.

Table F1. Conditional independence test results ($p > 0.01$).

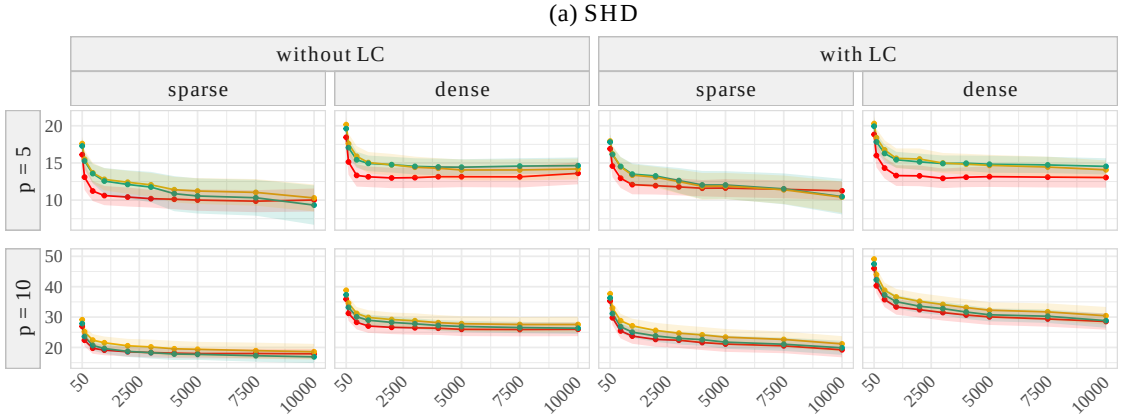
N	*50*	*150*	*500*	*1000*	*2000*	*3000*	*4000*	*5000*	*7500*	*10000*
Marginal	0	0	0	0	0	0	0	0	0	0
Cond. X_1	0.12	0	0	0	0	0	0	0	0	0
Cond. X_3	0.94	0.91	0.83	0.73	0.63	0.49	0.34	0.22	0.14	0.01
Cond. X_4	0.28	0	0	0	0	0	0	0	0	0
Cond. $\{X_1, X_4\}$	0.06	0	0	0	0	0	0	0	0	0
Cond. $\{X_1, X_3\}$	0.63	0.14	0	0	0	0	0	0	0	0
Cond. $\{X_3, X_4\}$	0.60	0.08	0	0	0	0	0	0	0	0
Cond. $\{X_1, X_3, X_4\}$	0.16	0	0	0	0	0	0	0	0	0

Note. The table shows all possible (conditional) independence patterns between X_2 and X_5 for different sample sizes (N) used in the simulation study. Each row corresponds to a specific independence relation, and each cell indicates the proportion of *p-values* that are greater than the α level of 0.01 ($p > 0.01$) out of 500 simulations, indicating that the null hypothesis of independence is not rejected. *Marginal* = not conditional on any variables; *Cond.* = conditional on.

Appendix G Sensitivity Analysis 1: Randomly Sampling Weights

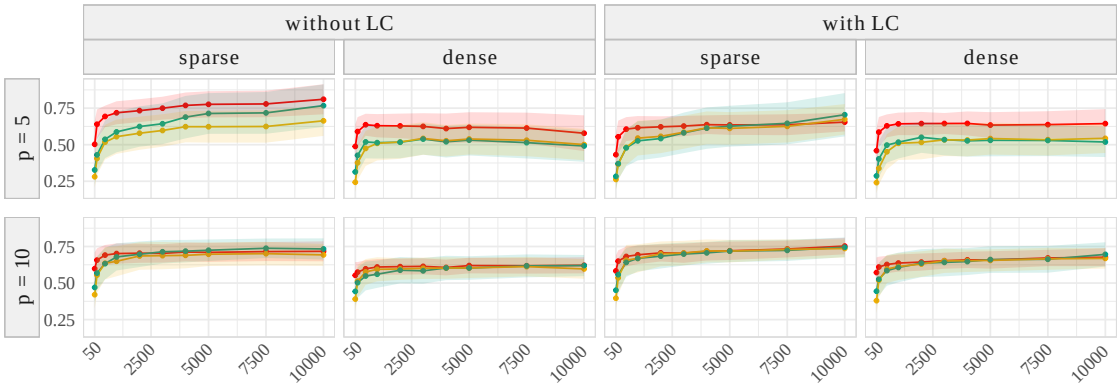
This appendix reports the complete results of the secondary simulation study where we randomly sampled parameters of **B** matrices. Figure G1 summarizes the results, which reveal several differences from the original study. First, CCD's performance appears to have improved across all conditions and shows comparable SHD values to CCI, even in sparse conditions (see Figure G1(a)). The outperformance of CCD is more evident when looking at precision (Figure G1(b)) and recall (Figure G1(c)). However, this is because CCD is more conservative and produces more circle endpoints than CCI, almost always recording higher uncertainty rates than CCI (see Figure G1(d)). Therefore, the overall conclusion remains the same as in the original simulation study; CCI is still the preferred choice when seeking to obtain the most information on causal directions with low uncertainty. Second, we observe that the FCI algorithm exhibits more stability and produces more circle endpoints, leading to similar levels of uncertainty as CCD across different conditions. However, its overall inferior performance compared to the other algorithms remains consistent with the findings of the original simulation study. Finally, we notice a significant change in the patterns in the *5-variable dense conditions*. Instead of the odd dips and spikes observed previously in the original simulation study, we now see a steady decrease in SHD and an increase in precision and recall values until they plateau. As explained in Section 4.4.2, the odd drops and spikes were a result of the specific weights chosen in those cases, and therefore, such patterns are expected to disappear when weight parameters are randomly sampled. Based on these findings, we can conclude that the results of the original simulation study are not substantially influenced by the particular choices of **B** matrices, except for the atypical *5-variable dense cases*.[17]

Figure G1. Performance analysis with randomly sampled weights for *B* matrices.

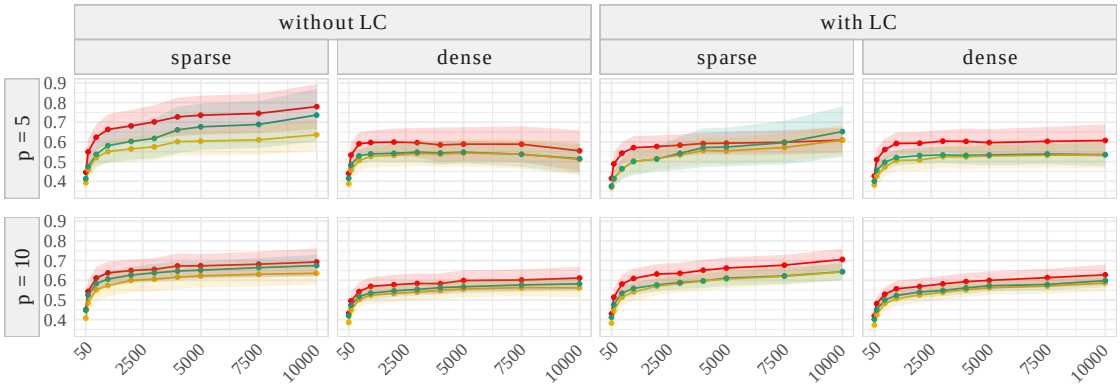


[17]Note that these results are based on sampling weights from both negative and positive values using a uniform distribution on $[-0.9, -0.1] \cup [0.1, 0.9]$. As an extra robustness check, we also performed the same analysis with only positive weights sampled from a Uniform($[0.1, 0.9]$). The outcomes were almost identical, with no discernible differences.

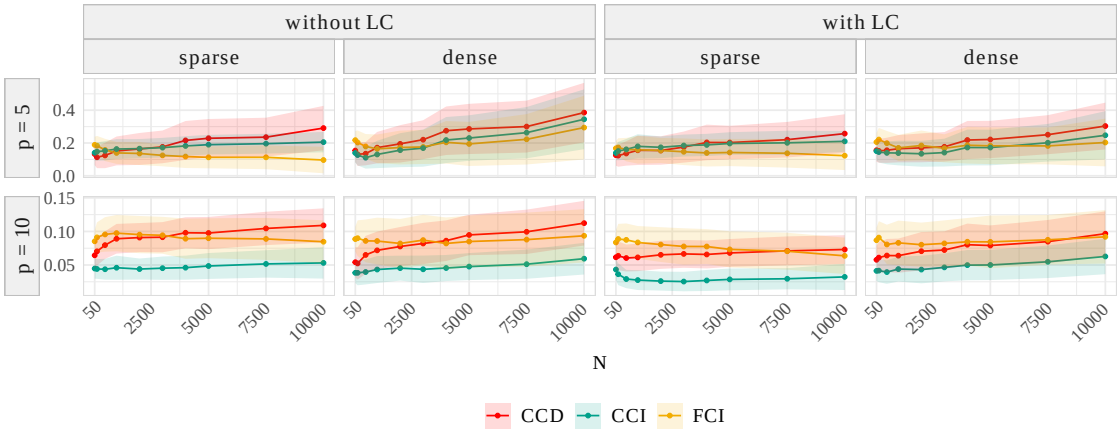
(b) Precision



(c) Recall



(d) Uncertainty

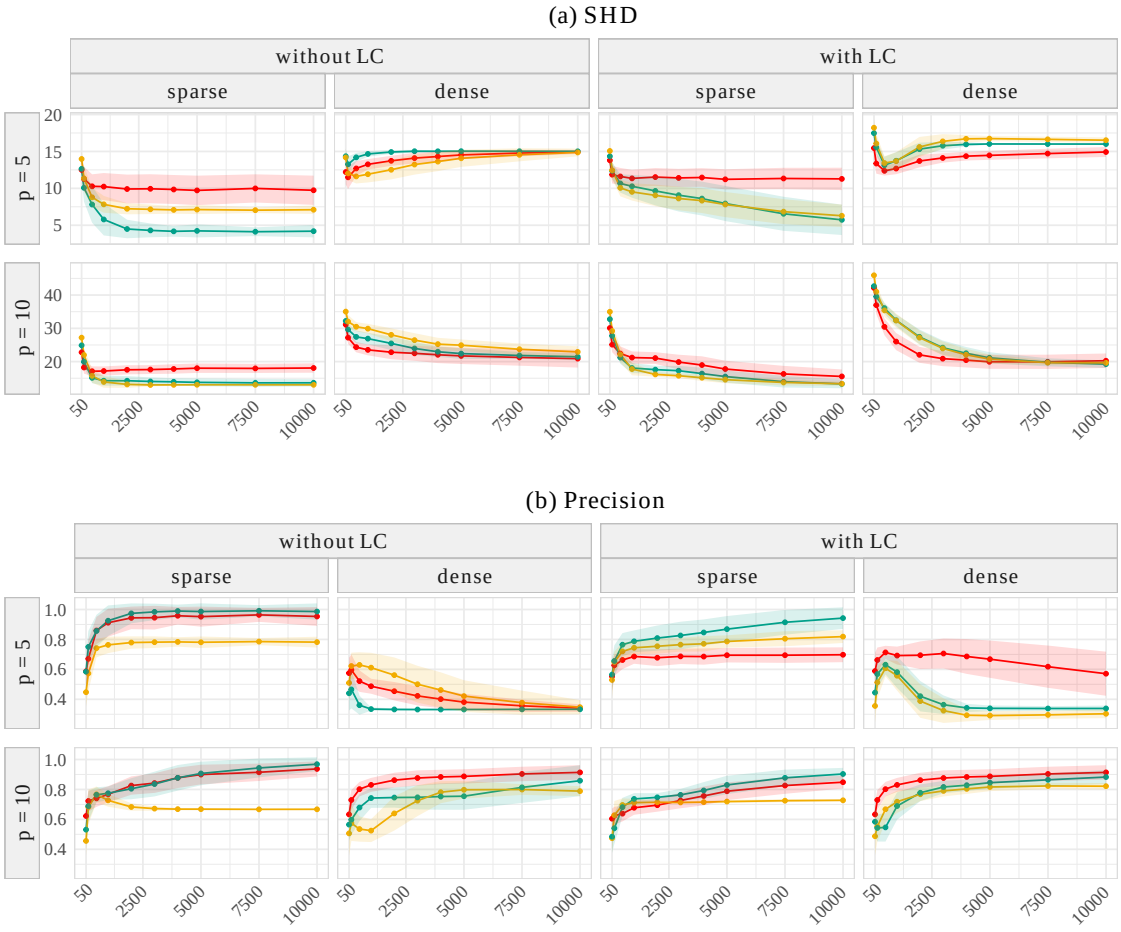


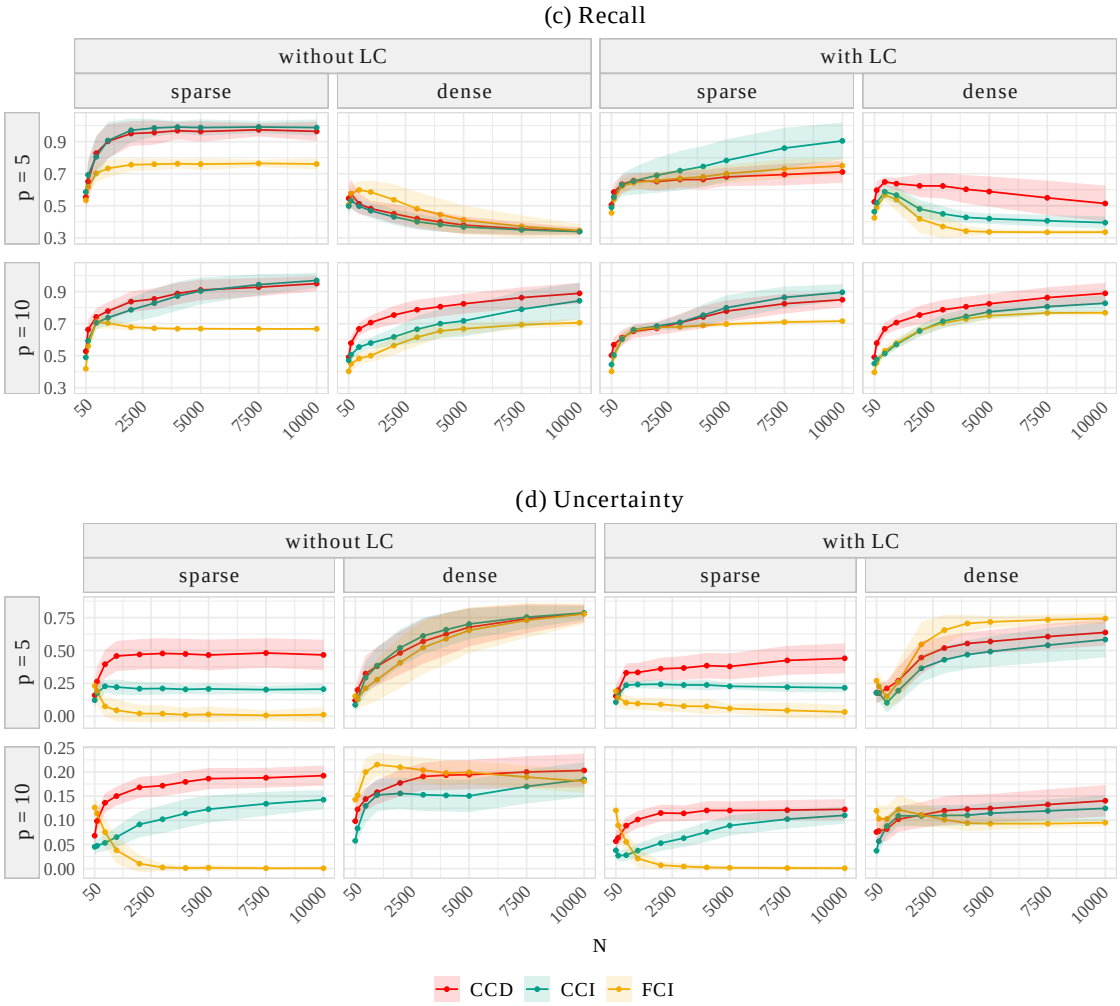
Note. Each point on the plot represents the average value of the corresponding metric across 500 iterations, while the shaded region represents the interquartile range (IQR) of the metric's distribution.

Appendix H Sensitivity Analysis 2: Varying α Levels

In this appendix, we present the findings of a secondary simulation study in which we varied the significance level (α) based on sample size. The α level plays a crucial role in constraint-based causal discovery (Spirtes et al., 2000), since it can be either too strict and indicate everything as independent or not strict enough and fail to find any independence relationships. To ensure consistent results from conditional independence tests, we lowered the α level as the sample size increased in this simulation study. The results can be found in Figure H1, which show that the observed patterns are almost identical to those from the original simulation study. Based on these findings, we can conclude that the results of our original simulation are robust to the fixed α level chosen in the study (i.e., $\alpha = 0.01$).

Figure H1. Performance analysis with varying α levels.





Note. Each point corresponds to the average value of a specified metric across 500 iterations, and the shaded area represents the interquartile range (IQR) of the metric's distribution.

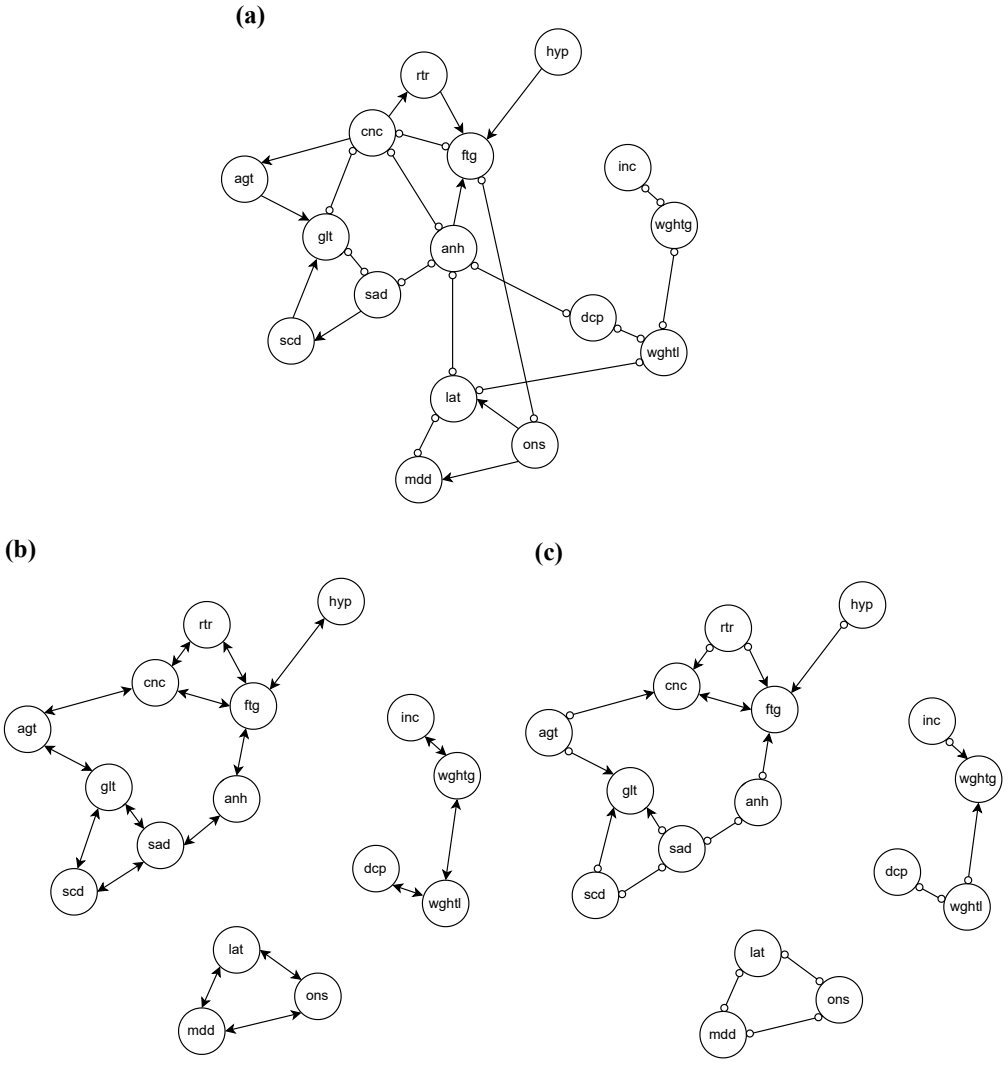
Appendix I Stability Analysis on Empirical Data

We performed a stability analysis on the empirical data in which we randomly sampled 90% of observations from the original dataset with $N = 408$, creating 1000 subsets with a sample size of 367 ($408 \times 0.9 = 367.2$). Subsequently, we applied all three algorithms (CCD, FCI, and CCI) to each of the 1000 subsets and retained only the edge-endpoints that were identified more than 70% of the time.[18] The PAGs resulting from each algorithm are presented in Figure I1. Overall, we observe that the causal structures found in the stability analysis are very similar to those obtained in the original analysis in Section 5.2. However, the resulting PAGs from the stability analysis are slightly more sparse, with a few missing edges. For example, the PAG estimated by CCD in Figure I1(a) no longer includes the edge between guilt (*glt*) and increased appetite (*inc*), which was present in the original analysis. Similarly, the PAGs estimated by CCI and FCI in Figure I1(b) and Figure I1(c),

[18]Note that the underlinings generated by the CCD algorithm are disregarded in this analysis.

respectively, do not show the edges between guilt (*glt*) and concentration (*cnc*) or anhedonia (*anh*) and decreased appetite (*dcp*). Additionally, there appears to be some loss of orientation in the PAGs, particularly in the one obtained from CCD. The CCD PAG in Figure I1(a) contains relatively more circle endpoints compared to the original result, which makes the cyclic structures observed in the original analysis less evident. Nevertheless, the results remained largely the same, except for only a couple of missing edges and a few arrow heads/tails replaced by circle marks. This suggests that our original results are stable and not significantly affected by minor variations in the data.

Figure I1. Estimated PAGs from stability analysis.

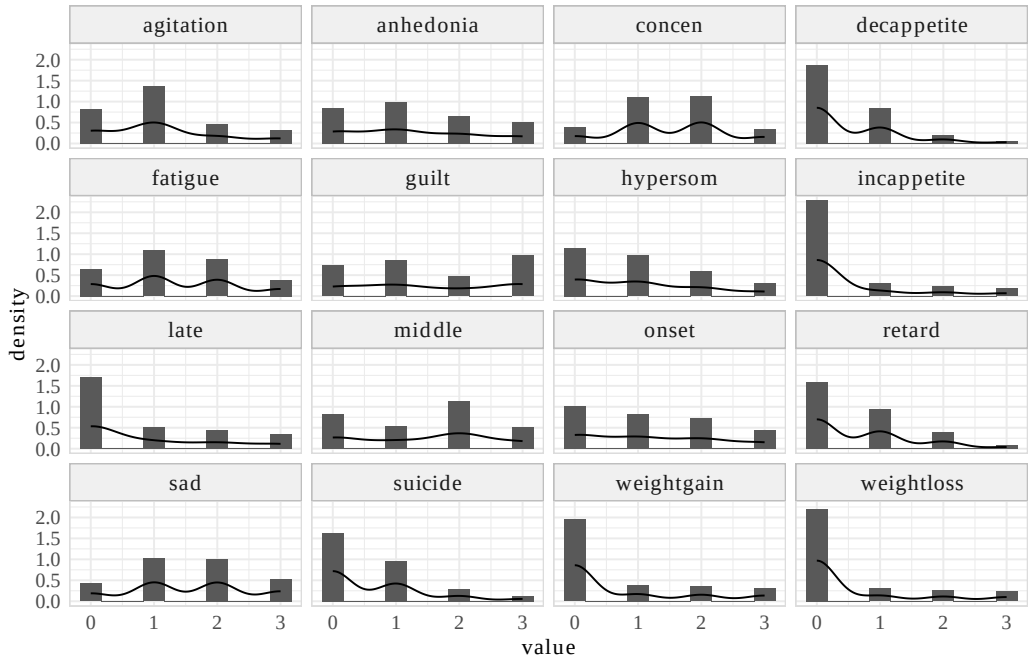


Note. Panels (a), (b), and (c) present the PAGs estimated by CCD, CCI, and FCI, respectively. Each PAG is obtained by selecting only the endpoints that were identified more than 70% of the times across 1000 subsets of the original data. *ons* = sleep onset insomnia; *mdd* = middle insomnia; *lat* = late (early morning awakening); *hyp* = hypersomnia; *sad* = sad; *dcp* = decreased appetite; *inc* = increased appetite; *wghtl* = weight loss; *wghtg* = weight gain; *cnc* = concentration impairment; *glt* = guilt; *scd* = suicidal thoughts; *anh* = anhedonia; *ftg* = fatigue; *rtr* = psychomotor retardation; *agt* = agitation.

Appendix J Distribution of Empirical Data

The distributions of depression symptoms are illustrated in Figure J1, which clearly indicates that most of them are skewed and do not follow a Gaussian distribution. Given that partial correlation is known to be a valid measure of conditional independence only in the case of multivariate Gaussian (Baba et al., 2004), it is possible that our analysis, which relied on partial correlations to test for conditional independence despite the non-Gaussian nature of the data, may have been misleading.

Figure J1. Distributions of depression symptoms.



Note. Each panel shows the distribution of individual depression symptom in the dataset from McNally et al. (2017). *concen* = concentration impairment; *decappetite* = decreased appetite; *hypersom* = hypersomnia; *incappetite* = increased appetite; *late* = late (early morning awakening); *middle* = middle insomnia; *onset* = sleep onset; *retard* = psychomotor retardation; *suicide* = suicidal thoughts.