Learning Cluster Causal Diagrams: An Information-Theoretic Approach

Xueyan Niu, Xiaoyun Li, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{niuxueyan, lixiaoyun996, pingli98}@gmail.com

Abstract

Many real-world phenomena arise from causal relationships among a set of variables. As a powerful tool, Bayesian Network (BN) has been successful in describing high-dimensional distributions. However, the faithfulness condition, enforced in most BN learning algorithms, is violated in the settings where multiple variables synergistically affect the outcome (i.e., with polyadic dependencies). Building upon recent development in *cluster causal diagrams* (C-DAGs), we initiate the formal study of learning C-DAGs from observational data to relax the faithfulness condition. We propose a new scoring function, the Clustering Information Criterion (CIC), based on information-theoretic measures that represent various complex interactions among variables. The CIC score also contains a penalization of the model complexity under the minimum description length principle. We further provide a searching strategy to learn structures of high scores. Experiments on both synthetic and real data support the effectiveness of the proposed method.

1 Introduction

There is a vast body of research on causal inference which aims at modeling the causal relations among random variables, e.g., [Saeed *et al.*, 2020; Chen *et al.*, 2021]. When there are many variables of interest, the probabilistic graphical models are popular in modern practice. For example, the Ising model, originated from statistical physics, has found applications in unsupervised learning and medical imaging [Morningstar and Melko, 2018]. Other undirected and directed graphical models include Markov random fields and Bayesian Networks (BNs), which are used actively in many machine learning problems [Liu *et al.*, 2017; Mokhtarian *et al.*, 2021]. Traditionally, the BN model represents causal relations using Directed Acyclic Graphs (DAGs) under the structural causal model framework [Pearl, 2000].

One notable limitation of the graphical models is that they capture pairwise interactions exclusively. Even though many studies have reported collective behavior among features in multivariate systems [Battiston *et al.*, 2020], investigating causalities emerged from collective interactions at the level of groups of variables is still a burgeoning research area. Recent developments include [Reing *et al.*, 2021], which studies how a group of variables causally influences the output of a neural network using sensitivity analysis. In particular, critiques have been made that graphical models, which comprise dyadic relations, fail to capture causal relations that emerge from synergistic interactions [James *et al.*, 2016]. The limitation is also related to the basic assumption in causality, known as the faithfulness condition, which is not well-justified in many situations [Uhler *et al.*, 2013]. Moreover, the rapid accumulation of data has challenged researchers to make meaningful deductions from high-dimensional data. For complex systems such as the human genome, connectome, and social networks, it is desirable that the whole shall be grouped into manageable subsystems to draw effective conclusions.

There have been works on capturing complex local dependencies. Qualitative Probabilistic Network (QPN) [Wellman, 1990; de Campos and Cozman, 2013] augments the graphical models by adding directed hyper-edges. [Tikka *et al.*, 2021] proposed transit cluster scheme which reduces the size of a DAG while preserving the identifiability properties of causal effects. Recently, [Anand *et al.*, 2022] introduced the *cluster causal diagrams* (C-DAGs) with extended d-separation rules and do-calculus. In C-DAG, causal relations are represented among clusters of variables but not among the variables within the clusters. Once the C-DAG is known, we are able to perform causal inference tasks, e.g., policy evaluation [Correa and Bareinboim, 2020], sponsored search advertising [Nabi *et al.*, 2020], and knowledge transferring [Pearl and Bareinboim, 2014], coarsely among groups of variables.

One question left open in [Anand *et al.*, 2022] is how C-DAGs can be effectively learned from data. In this paper, we address this question by making the following **contributions**:

- We adapt the *cluster causal diagram* (C-DAG) model, which groups random variables with synergistic effects into clusters, to relax the faithfulness assumption in structure learning.
- We propose the Clustering Information Criterion (CIC) score that represents the goodness of fit of the C-DAG model penalized with the model complexity.
- We develop a greedy search strategy to explore the space of C-DAGs to learn the structure that fits the data. Experimental results confirm its effectiveness.

2 Background

2.1 Notations

In the paper, we consider $n > 1$ random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, viewed as n vertices in G, and a partition of the n variables into clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, where $1 \leq m \leq n$, such that $\cup_{i=1}^{m} C_i = \boldsymbol{X}, C_i \subseteq \boldsymbol{X}, C_i \cap C_j = \emptyset, \forall i, j \in [m], i \neq j$. The set of all possible partitions of $\boldsymbol{X}$ is denoted by $\Pi(\boldsymbol{X})$. The number of elements in a cluster C_i is denoted by $|C_i|$. Suppose X_i has finite alphabet size s_i. $[n]$ denotes the set of numbers $\{1, 2, \ldots, n\}$.

A directed graph $G = (V, E)$ is called a Directed Acyclic Graph (DAG) if it contains no directed cycles. $\mathrm{Pa}(\cdot)$ denotes the set of parents: $\mathrm{Pa}(X_j) = \{X_i : (X_i, X_j) \in E\}$. X_j is said to be a descendant of X_i if there exists a directed path from X_i to X_j in G, and X_i is said to be an ancestor of X_j. The set of descendants and the set of ancestors of X_k are denoted as $\mathrm{De}(X_k)$ and $\mathrm{An}(X_k)$ respectively. A Structural Causal Model (SCM) [Pearl, 2000] is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{U}) \rangle$, where $\boldsymbol{U}, \boldsymbol{V}$ are exogenous (latent) and endogenous (observable) variables, and $\mathcal{F} = \{f_i\}_{i \in [n]}$ is a collection of functions that encode causal relations, i.e., $X_i = f_i(\mathrm{Pa}(X_i), U_i)$. So there is a directed edge $(\mathrm{Pa}(X_i) \rightarrow X_i)$ in G indicating the direct causation. A vertex X_j is called a collider if there is a v-structure $(X_i \rightarrow X_j \leftarrow X_k)$ in the graph. $P(\boldsymbol{U})$ is the joint distribution that encode the uncertainty of the system. The set of vertices in the graph correspond to the random variables, i.e., $\boldsymbol{V} = \{X_1, X_2, \ldots, X_n\}$. Further, $\langle G, P \rangle$ is called a Bayesian network if the following factorization holds: $P_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} P_{X_i \mid \mathrm{Pa}(X_i)}(x_i \mid \mathrm{Pa}(x_i))$. Let $C \in \mathcal{C}$ be a cluster. We use $\boldsymbol{X}_C$ to denote $(X_i \mid X_i \in C)$. and $P_{\boldsymbol{X}_C}$ to denote $P_{(X_i \mid X_i \in C)}$. We say that $\langle G, P \rangle$ satisfies the *Markov property* if for any vertex $X_i \in \boldsymbol{V}$, X_i is conditionally independent of its non-descendents given $\mathrm{Pa}(X_i)$.

2.2 Score-based DAG Learning

Given observed data $\mathcal{D}$, learning a BN is a structure learning procedure to find a DAG G and the parameters that best encode $\mathcal{D}$. In particular, the task of learning the DAG G is known as structure learning. One popular approach is through the information-theoretic scores, e.g., the Hill-Climbing algorithm [Heckerman *et al.*, 1995]. Score-based structure learning solves the problem

$$G^* = \arg\max_{G \in \mathcal{G}} \ \mathrm{Score}(G : \mathcal{D}),$$

where $\mathcal{G}$ is the set of admissible DAGs, and $\mathrm{Score}(\cdot)$ is a scoring function that evaluates how well the model fits the data. There are two types of scoring functions: Bayesian and information-theoretic. The Bayesian scoring functions, such as K2 and BDeu, aim at maximizing the posterior probability $P(G|\mathcal{D})$ under various assumptions given a prior distribution.

In this paper, we focus on the information-theoretic scores, which explore the Minimum Description Length (MDL) principle where the model with shorter description length is preferred. The description length includes two components, that required to describe the data and that required to describe the network. Using the Huffman code, the description length of the data is proportional to the log-likelihood, which can be expressed as

$$LL(G : \mathcal{D}) = -N \sum_{i=1}^{n} H_{\mathcal{D}}(X_i \mid \mathrm{Pa}(X_i)),$$

where N is the number of instances in $\mathcal{D}$, and $H_{\mathcal{D}}(\cdot|\cdot)$ is the conditional entropy using the empirical data distribution $P_{\mathcal{D}}(\cdot)$. The log-likelihood is penalized with the complexity of the model, $C(G)$. Therefore, in general, the information-theoretic score is in the form

$$\mathrm{Score}_{\mathrm{IT}}(G : \mathcal{D}) = LL(G : \mathcal{D}) - f(N)C(G),$$

where $f(N)$ is a proportionality factor that depends on the number of instances N. The popular Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores are taking $f(N)$ to be 1 and $\frac{1}{2} \log N$ respectively.

Using another information-theoretic measure, the mutual information (MI), [de Campos, 2006] proposed the MIT scoring function that achieves good performance. Specifically, they observed the equivalence between maximizing the log-likelihood and maximizing the sum of the MI:

$$\sum_{\mathrm{Pa}_G(X_i) \neq \emptyset} I_{\mathcal{D}}(X_i; \mathrm{Pa}_G(X_i)) = \frac{1}{N} LL(G : \mathcal{D}) + \sum_{i=1}^{n} H_{\mathcal{D}}(X_i), \tag{1}$$

where $H_{\mathcal{D}}(\cdot)$ and $I_{\mathcal{D}}(\cdot;\cdot)$ are the entropy and the MI based on the empirical distribution of $\mathcal{D}$. The score is similar to the one of [Chow and Liu, 1968] which aims at approximating the empirical distribution with tree structures.

3 C-DAG: Motivation, Definition and Setup

In this section, we introduce the definition of cluster causal diagram (C-DAG) and demonstrate its motivation through an example. Then we discuss the faithfulness of C-DAG.

3.1 Model and the Faithfulness Assumption

We formalise the C-DAG model, which encapsulates assumptions about how variables are related.

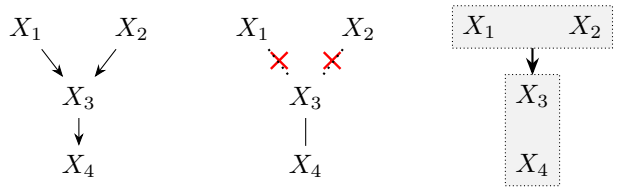
Assumption 1. *The causal mechanism is a SCM, where the exogenous variables are independent, and any two endogenous variables do not share a common exogenous parent.*

This assumption is standard in many empirical studies (see, for example, Section 9.2 of [Spirtes *et al.*, 2000]). Below, we adapt the definition of the C-DAG [Anand *et al.*, 2022] according to Assumption 1.

Definition 1 (Cluster Causal Diagram)**.** *Consider a SCM and the corresponding causal diagram $G = (\boldsymbol{V}, \boldsymbol{E})$. Given a partition $\mathcal{C} = \{C_1, \ldots, C_m\}$ of $\boldsymbol{V}$, construct a graph $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ as follows:*

$$\boldsymbol{E}_{\mathcal{C}} = \{(C_i, C_j) : \exists V_i, V_j \in C_j, s.t. V_i \in \mathrm{Pa}(V_j)\}. \tag{2}$$

If $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ contains no cycles, then the partition $\mathcal{C}$ is admissible, and $G_{\mathcal{C}}$ is called a C-DAG. We denote $\mathrm{Pa}_{G_{\mathcal{C}}}(C_j) := \{C_i : (C_i, C_j) \in \boldsymbol{E}_{\mathcal{C}}\}$. We shall omit the subscript of $\mathrm{Pa}(\cdot)$ when there is no ambiguity.



(a) True SCM. (b) Essential graph of the learned DAG. (c) C-DAG representation.

Figure 1: An example containing synergistic causal effect. (a) The distribution is not faithful to the true DAG. (b) Constraint-based algorithms are unable to detect synergistic effects. (c) C-DAG aims at learning the causal effects between groups of random variables.

Definition 1 is different from group DAGs [Parviainen and Kaski, 2017], which has been proposed to learn dependencies from *a prior* known groups. Similar to the BN, we assume the following factorization holds for C-DAG.

Assumption 2. *The joint distribution $P_{\boldsymbol{X}}$ factorizes over a C-DAG $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$, i.e., $P_{\boldsymbol{X}} = \prod_{C \in \mathcal{C}} P_{\boldsymbol{X}_C | \ \mathrm{Pa}(\boldsymbol{X}_C)}$.*

Faithfulness Assumption. Many structure learning algorithms, such as the PC algorithm [Kalisch and Bühlman, 2007] and the Max-Min Hill-Climbing (MMHC) algorithm [Tsamardinos *et al.*, 2006], rely on the faithfulness assumption [Uhler *et al.*, 2013; Pearl, 2000].

Definition 2 (DAG faithfulness)**.** *A distribution P is faithful to a DAG, G, if no conditional independence relations other than the ones entailed by the Markov property are present.*

In other words, A BN satisfies the faithfulness condition if the joint distribution $P(\boldsymbol{X})$ embodies only independencies that can be represented in the DAG [Spirtes *et al.*, 2000]. Despite the common acceptance, the faithfulness assumption has been shown to be questionable when learning DAGs [Uhler *et al.*, 2013]. In statistics, as discussed in [Robins *et al.*, 2003], the practice of selecting variables who have large regression coefficients also assumes the faithfulness condition, which may not hold in many situations in regression analysis. Next, we present a simple example where the absence of faithfulness leads to poor causal discovery performance.

3.2 A Motivating Example

Let $\boldsymbol{V} = \{X_1, X_2, X_3, X_4\}$ be the observed random variables, $\boldsymbol{U} = \{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$ be the exogenous binary random variables. Consider the causal mechanism (see Figure 1a):

$$X_1 = \epsilon_1,\ X_2 = \epsilon_2,\ X_3 = (X_1 \oplus X_2) + \epsilon_3,\ X_4 = X_3 + \epsilon_4,$$

where $X_1 \oplus X_2 := (X_1 + X_2) \mod 2$ is the binary addition operator. Specifically, $X_1 \oplus X_2 = 1$ if and only if $X_1 \neq X_2$. In biology, X_1, X_2 could indicate the existence or non-existence of two genes, and X_3 could indicate a phenotype that is observed if and only if one of X_1 and X_2 is present. X_4 could represent a measurement related to the phenotype.

Assume $\epsilon_1, \epsilon_2 \overset{i.i.d.}{\sim} \mathrm{Be}(\frac{1}{2})$. We use the classical Binary Symmetric Channel (BSC), to construct ϵ_3 and ϵ_4. A BSC between the input Y and output Z with crossover probability p is such that $P(Z = i|Y = j) = p$ if $i \neq j$ and $P(Z = i|Y = j) = 1-p$ if $i = j$ for $i, j \in \{0, 1\}$. Let ϵ_3 be such that $X_1 \oplus X_2$ and X_3 form a BSC with crossover probability δ, and let ϵ_4 be such that X_3 and X_4 form a BSC with crossover probability γ. The joint distribution $P_{X_1X_2X_3X_4}$ is factorized as $P_{X_1X_2X_3} = P_{X_1} \cdot P_{X_2} \cdot P_{X_3|X_1,X_2} P_{X_4|X_3}$, which implies three directed edges $(X_1 \to X_3)$, $(X_2 \to X_3)$, $(X_3 \to X_4)$.

We can compute the MI: $I(X_1; X_2) = I(X_1; X_3) = I(X_1; X_4) = I(X_2; X_3) = I(X_2; X_4) = 0$, and $I(X_3; X_4) = 1 - H(\gamma) \geq 0$. Note that zero MI implies statistical independency. Yet, these independence relations, e.g., $X_1 \perp X_3, X_2 \perp X_3$, are not entailed by the Markov property. Hence, the faithfulness condition is not satisfied for the DAG. If we pick variables using the coefficients from pairwise regression, then only one edge connecting X_3 and X_4 may be present, while the edges $(X_1 \to X_3)$ and $(X_2 \to X_3)$ will not be recovered, as illustrated in Figure 1b. That is, the causal relations between X_1, X_2 and X_3, X_4 are lost.

3.3 Faithfulness of C-DAG

Figure 1c illustrates the use of C-DAG to circumvent the issue raised by the example in Section 3.2 due to the violation of the faithfulness condition. By clustering random variables according to the synergistic measure of information, we can recover the causal relations at the clustering (group) level. Instead of requiring the Bayesian network $\langle G, P \rangle$ to be faithful, we assume faithfulness at a coarser scale, where the individual variables who violate the condition can be grouped together to preserve the faithfulness for C-DAGs.

Definition 3 (C-DAG faithfulness)**.** *A distribution P is faithful to a C-DAG $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ if no conditional independence relations other than the ones entailed by the Markov property, viewing clusters as individual vertices, are present.*

Theorem 1. *Consider a BN $\langle G, P \rangle$, where $G = (V, E)$ is a DAG. Given a clustering $\mathcal{C} = \{C_1, \ldots, C_m\}$ of V. Let $G_{\mathcal{C}}$ be the C-DAG induced by G and $\mathcal{C}$ according to Definition 1. If for any collider $V_i \in \mathbf{V}$, there exists $C_i \in \mathcal{C}$ such that either $C_i = \{V_i\} \cup \mathrm{De}(V_i)$ or $\{V_i\} \cup \mathrm{An}(V_i) \subseteq C_i$, then the faithfulness to G implies that to $G_{\mathcal{C}}$.*

The proof of the theorem follows from Theorem 4.6 of [Anand *et al.*, 2022]. Theorem 1 shows the faithfulness of a BN implies the faithfulness of the C-DAG representation. However, the converse does not hold (see the example in Section 3.2), i.e., the faithfulness of the C-DAG is a weaker condition therefore may have broader applications in real-world complex systems. This result differs from [Parviainen and Kaski, 2017] as the definitions for group DAG and C-DAG are different: C-DAG requires the edges to satisfy Eq. (2), whereas the clusterings in group DAGs are known *a prior*.

4 Information-Theoretic C-DAG Learning

Based on the discussion above, the C-DAG learning objective of this paper is formally stated as follows.

Learning objective: Given a set of N observations $\mathcal{D} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$, where $\boldsymbol{y}_i \in \mathbb{R}^n$, $\forall 1 \leq i \leq N$, the goal is to recover a C-DAG $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ that embodies the

causal relationships in the cluster causal diagram such that the learned C-DAG satisfies faithfulness (Definition 3).

Next, we will propose an information-theoretic scoring function, the Clustering Information Criterion (CIC), that captures how well the data fit the C-DAG model while encouraging low complexity. We will then introduce an algorithm to search for the C-DAG with high CIC score.

4.1 Basics of Information Measures

We first review some definitions in information theory. Let $H(\boldsymbol{X}) = -\int P_X \log P_X \, \mathrm{d}x$ be the *entropy* of a random variable X with density P_X, which could be vector-valued. The *mutual information* (MI) between X and Y with joint density $P_{X,Y}$, defined as

$$I(X,Y) = D_{KL}(P_{X,Y} \| P_X \cdot P_Y) = \int p_{X,Y} \log \frac{P_{X,Y}}{P_X P_Y} \, \mathrm{d}x,$$

measures the mutual dependence between X and Y. Particularly, zero MI implies statistical independence. The *conditional mutual information* is defined as

$$I(X,Y|Z) = D_{KL}(P_{X,Y|Z} \| P_{X|Z} \cdot P_{Y|Z}).$$

The *total correlation* extends MI to multi-variate case,

$$TC(X_1, \ldots, X_n) = D_{KL}(P_{X_1,\ldots,X_n} \| P_{X_1} \cdots P_{X_n}),$$

where $P_{X_1,\ldots,X_n}$ is the joint density of $(X_1, \ldots, X_n)$.

4.2 Information-Theoretic Evaluation of C-DAG

We will introduce our method in several steps. Given a dataset and candidate structures, we first evaluate the degree of fitness of the C-DAG to the data while endeavoring to group variables that may violate the faithfulness condition due to high inter-dependencies and synergistic effects.

Regarding the fitness of the data $\mathcal{D}$, we exploit the log-likelihood discussed in Section 2.2. According to Eq. (1), maximizing the log-likelihood of the DAG G is equivalent to maximizing the sum of the empirical MI

$$g_1(G : \mathcal{D}) = \sum_{\mathrm{Pa}_G(X_i) \neq \emptyset} I_{\mathcal{D}}(X_i; \mathrm{Pa}_G(X_i)),$$

noticing that the term $\sum_{i=1}^n H_{\mathcal{D}}(X_i)$ is independent of the model. To generalize the score to C-DAGs, we define the measure of log-likelihood w.r.t. a C-DAG $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ as

$$g_2(G_{\mathcal{C}} : \mathcal{D}) = \sum_{\mathrm{Pa}_{G_{\mathcal{C}}}(\boldsymbol{X}_C) \neq \emptyset} I_{\mathcal{D}}(\boldsymbol{X}_C; \mathrm{Pa}_{G_{\mathcal{C}}}(\boldsymbol{X}_C)) \quad (3)$$

The relationship between the two functions $g_1(G : \mathcal{D})$ and $g_2(G_{\mathcal{C}} : \mathcal{D})$ is non-trivial, because the MI is non-modular, i.e., the sum $\sum_{X_i \in \mathrm{Pa}(Y)} I(X_i; Y)$ may not equal to $I(\mathrm{Pa}(Y); Y)$. There is a line of research on how the MI can be decomposed into modular components including synergistic, redundant, and unique information [Lizier *et al.*, 2018; Niu and Quinn, 2019]. Roughly speaking, the MI between the clusters minus sum of pairwise MI indicates the presence of synergistic interaction [McGill, 1954], which will be used in our score function. In addition, to discover strongly correlated random variable, we use the total correlation which reflects the total amount of information present in a set of random variables. Adding these two elements together, we evaluate the C-DAG by:

$$\begin{aligned} g(G_{\mathcal{C}} : \mathcal{D}) = &\sum_{C \in \mathcal{C}} TC(\boldsymbol{X}_C) + g_2(G_{\mathcal{C}} : \mathcal{D}) \\ &- \sum_{C \in \mathcal{C}} \sum_{i \in C} \sum_{j \in \mathrm{Pa}_{G_{\mathcal{C}}}(\boldsymbol{X}_C)} I_{\mathcal{D}}(X_i; X_j). \end{aligned} \quad (4)$$

4.3 C-DAG Complexity Penalization

Highly complex model often incurs over-fitting. We present a regularization built upon the Minimum Description Length (MDL) principle. It is desirable that the derived clustering preserves as much information as possible while maintaining a short description length. From the perspective of MDL, a model can be seen as a prefix code, and it is preferred that the code compresses the data into fewer bits.

To encode the C-DAG, the following information is necessary and sufficient: (1) a partition of the n vertices; (2) a list of the parents of each cluster, and (3) the set of conditional probabilities associated with each vertice needed to parametrize the network.

Suppose the n random variables are partitioned into m clusters $\mathcal{C} = \{C_1, \ldots, C_m\}$, and that each cluster C_i has k_i parent clusters. To encode the partitioning, we consider the number of ways to partition a set of n elements into $m \leq n$ non-empty subsets. This number is called the Stirling number of the second kind,

$$S(n,m) = \frac{1}{m!} \sum_{i=0}^{m} (-1)^i \binom{m}{i} (m-i)^n.$$

Therefore, the number of bits to encode a particular partition $\mathcal{C}$ is $\log(S(n,m))$. In practice, the number can be efficiently computed using recursions. Moreover, to encode the parent clusters, we need $k_i \log(m)$ bits. To parametrize the network, we need the conditional probabilities $P_{C_i|\mathrm{Pa}(C_i)}$, which requires $(\prod_{X_j \in C_i} s_j - 1) \prod_{X_k \in \mathrm{Pa}(C_i)} s_k$ real numbers. To summarize, we regularize the model complexity based on the minimal description length principle with a penalty:

$$\begin{aligned} C(G_{\mathcal{C}}) = &\log(S(n,m)) + \sum_{i=1}^{m} k_i \log(m) \\ &+ \sum_{i=1}^{m} \Big(\prod_{X_j \in C_i} s_j - 1 \Big) \prod_{X_k \in \mathrm{Pa}(C_i)} s_k \end{aligned} \quad (5)$$

4.4 The Clustering Information Criterion Score

With the scoring function (4) and the penalization function (5), we are now ready to introduce our Clustering Information Criterion (CIC). Throughout the previous discussion, we have omitted the sample size N. In general, the score should be normalized with factors that depend on N, i.e., $score(G : \mathcal{D}) = f_1(N) g(G : \mathcal{D}) - f_2(N) C(G)$. To choose proper factors, we use $f_1(N) = 2N$ in the Mutual Information Test (MIT) score [de Campos, 2006] and $f_2(N) = \frac{1}{2} \log N$ in the BIC score [Schwarz, 1978]. The final expression of the proposed CIC scoring function is:

$$\mathrm{Score}_{\mathrm{CIC}}(G_{\mathcal{C}} : \mathcal{D}) = 2N \cdot g(G_{\mathcal{C}} : \mathcal{D}) - \frac{\log N}{2} \cdot C(G_{\mathcal{C}}). \quad (6)$$

4.5 Search Strategy

In order to efficiently find the structures with high scores, in this section, we develop a search strategy over the space of C-DAGs. We take advantage of the decomposibility of the score given a clustering, i.e., we only need to re-calculate the score based on that local changes including adding/deleting/reversing an edge. We follow the standard greedy search strategy. Our Algorithm 1 implements Z steps of greedy search, Algorithm 2. After obtaining the clusters and edges, we run conditional independency test on each separator to make sure the $v-$structures are corrected recovered.

Greedy Search. We apply a local greedy search operation to search for the local optimal structure. To explore the structures, we define the neighborhood of a C-DAG $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ by applying one of the following possible moves that does not result in a cycle: adding, deleting, and reversing an edge in $\boldsymbol{E}_{\mathcal{C}}$. We define the neighborhood of a C-DAG $G_{\mathcal{C}}$ as

$$\mathcal{N}(G_{\mathcal{C}}) := \{G'_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}'_{\mathcal{C}}) \text{ admissible } : \boldsymbol{E}'_{\mathcal{C}} \text{ derived by adding/deleting/reversing an edge from } \boldsymbol{E}_{\mathcal{C}}\},$$

and the greedy search is to compare scores over the space of neighboring C-DAGs.

Random sampling. Algorithm 1 involves randomly sampling a clustering $\mathcal{C}$ and edges $\boldsymbol{E}_{\mathcal{C}}$. To get a partition of $[n]$, we first sample the number of clusters m and then assign each random variable to a randomly selected cluster. For each vertex, the probability of being assigned to a particular cluster is uniform. We can control the clustering sparsity by specifying the distribution of m. To sample the edges $\boldsymbol{E}_{\mathcal{C}}$, we apply the standard Erdős-Rényi procedure. We first generate undirected edges $\boldsymbol{E}_{\text{und}}$ such that $e_{ij} \overset{i.i.d.}{\sim} \text{Be}(\rho), \ \forall\, 1 \le i < j \le m$, where e_{ij} indicates whether there is an edge between the two clusters C_i and C_j. Next, for any edge $\{C_i, C_j\} \in \boldsymbol{E}_{\text{und}}$, we set the order (C_i, C_j) if $i < j$. The last step is to randomly permutate the cluster labels. The resulting graph $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$ is guaranteed to be a C-DAG.

Algorithm 1 Learning C-DAG representation

Input: Data $\mathcal{D}$, Number of steps Z, Sampling parameter ρ
Output: A learned structure G^*

Randomly sample a clustering $\mathcal{C}$ and edges $\boldsymbol{E}_{\mathcal{C}}$.
$G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$, $G^* = G_{\mathcal{C}}$, $score = \text{Score}_{\text{CIC}}(G_{\mathcal{C}} : \mathcal{D})$.
for $i = 1$ **to** Z **do**
  Randomly sample a clustering $\mathcal{C}$ and edges $\boldsymbol{E}_{\mathcal{C}}$.
  Set $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$.
  $G'_{\mathcal{C}} = \text{GreedySearch}(G_{\mathcal{C}})$
  if $\text{Score}_{\text{CIC}}(G'_{\mathcal{C}} : \mathcal{D}) > score$ **then**
    $G^* = G'_{\mathcal{C}}$, $score = \text{Score}_{\text{CIC}}(G'_{\mathcal{C}} : \mathcal{D})$.
  end if
end for
for each $C_i - C_j - C_k$ appears in G^* **do**
  if $\boldsymbol{X}_{C_i} \perp \boldsymbol{X}_{C_k} | \boldsymbol{X}_{C_j}$ is False **then**
    $E^*_{\mathcal{C}} = (E^*_{\mathcal{C}} \setminus \{(C_j, C_i), (C_j, C_k)\}) \cup \{(C_i, C_j), (C_k, C_j)\}$.
  end if
end for
return G^*

Algorithm 2 Greedy Search

Input: Data $\mathcal{D} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$, Structure $G_{\mathcal{C}} = (\mathcal{C}, \boldsymbol{E}_{\mathcal{C}})$
Output: A local optimal $G'_{\mathcal{C}}$

$G'_{\mathcal{C}} = G_{\mathcal{C}}$, $score = \text{Score}_{\text{CIC}}(G'_{\mathcal{C}} : \mathcal{D})$.
for $G''_{\mathcal{C}} \in \mathcal{N}(G'_{\mathcal{C}})$ **do**
  if $\text{Score}_{\text{CIC}}(G''_{\mathcal{C}} : \mathcal{D}) > score$ **then**
    $G'_{\mathcal{C}} = G''_{\mathcal{C}}$
    $score = \text{Score}(G'_{\mathcal{C}} : \mathcal{D})$
  end if
end for
return $G'_{\mathcal{C}}$

5 Experiments

In this section, we provide numerical results to demonstrate the practical benefits of our work, that the proposed Clustering Information Criterion (CIC) score and C-DAG learning algorithm coalesce into a principled framework for learning cluster causal diagrams from observational data.

5.1 Synthetic Datasets

We begin our experimental investigation by considering two Boolean functions exhibiting different synergistic behaviors that are only available through the interaction among a set of variables. The functions of interests are the *parity* and the *majority*. The output of the *parity* is true if and only if the number of ones in the input is odd. The output of the *majority* is true if and only if at least half of the inputs are true. For these functions, it is impossible to determine the output using an individual input. In particular, the *parity* is purely synergistic in terms of the MI: let $X_i \sim \text{Be}(p_i), i \in [n]$, and $Y = Parity(X_1, \ldots, X_n)$, then $I(\boldsymbol{X} \setminus \{X_i\}; Y) = 0$, but $I(\boldsymbol{X}; Y) > 0$, i.e., every strict subset of $\boldsymbol{X}$ is independent of the output, whereas the output is deterministic of the inputs.

We use the two functions of triple-wise interactions as building blocks of the clusters to construct the C-DAGs. Below, $\neg$ and $\wedge$ denote the logic negation and logic AND operations. We use the Binary Symmetric Channel (BSC) to simulate noises. $X = \text{BSC}_p(Y)$ denotes $P(X = i | Y = i) = 1 - p$ and $P(X = \neg i | Y = i) = p$ for $i = 0, 1$. We consider an underlying SCM using different combinations of the *parity* and the *majority*, and present the results of three representative scenarios. Let $X_1, X_2, X_3, X_4, X_5, X_6 \overset{i.i.d.}{\sim} \text{Be}(\frac{1}{2})$, and we generate X_7, X_8 by

Scenario 1.

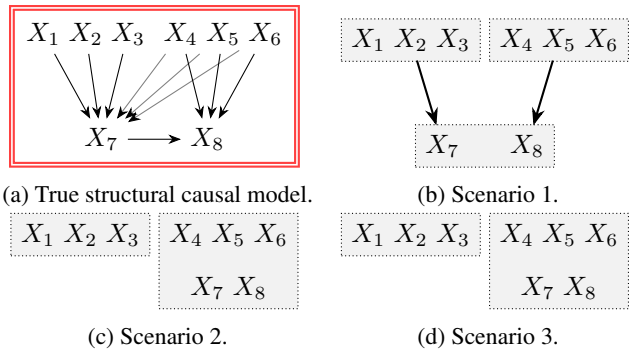
$$\text{BSC}_{p_1}(X_7) = Parity\,(X_1, X_2, X_3) \wedge Parity\,(X_4, X_5, X_6),$$
$$\text{BSC}_{p_2}(X_8) = X_7 \wedge Parity\,(X_4, X_5, X_6).$$

Scenario 2.

$$\text{BSC}_{p_1}(X_7) = Maj\,(X_1, X_2, X_3) \wedge Maj\,(X_4, X_5, X_6),$$
$$\text{BSC}_{p_2}(X_8) = X_7 \wedge Maj\,(X_4, X_5, X_6).$$

Scenario 3.

$$\text{BSC}_{p_1}(X_7) = Parity\,(X_1, X_2, X_3) \wedge Maj\,(X_4, X_5, X_6),$$
$$\text{BSC}_{p_2}(X_8) = X_7 \wedge Parity\,(X_4, X_5, X_6).$$

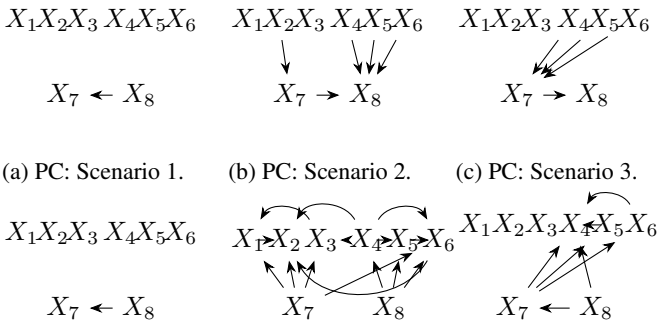


(a) True structural causal model. (b) Scenario 1.

(c) Scenario 2. (d) Scenario 3.

Figure 2: Learned C-DAGs using the proposed method.

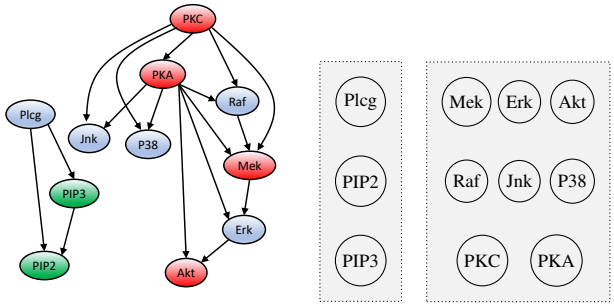
In each scenario, we simulate $N = 1000$ data points and run Algorithm 1 with $Z = 500$ iterations. We set the parameter of the Erdős-Rényi graph $\rho = 0.5$ for random sampling and the channel noise parameters $p_1 = p_2 = 0.1$. We visualize the true DAG (Figure 2a, for all three cases) and the learned C-DAGs in Figure 2. We see that for scenario 1, our method successfully recovers the synergistic effects caused by the *parity* function (Figure 2b). In the other two scenarios, the *majority* function exhibit less synergistic effects than the *parity* function. Still, the learned C-DAG well identifies the clustering structure of $C_1 = \{X_1, X_2, X_3\}$. The cluster $C_2 = \{X_4, X_5, X_6, X_7, X_8\}$ in Figure 2c and 2d is due to the high inter-dependencies. For scenario 2 and 3, although there is a missing edge between the two clusters according to (2), the correlation between the two clusters can be detected from the data, and the two directions $C_1 \to C_2$ and $C_2 \to C_1$ are not statistically distinguishable. As a comparison, in Figure 3, we report the results of the PC and the Hill-Climbing (HC) algorithms using the open-source `pgmpy` package in Python. The presented result is with significance level $s = 0.01$. We also tested $s = 0.05, 0.1$, which leads to minor changes in the edges but does not change the main conclusion. In all cases, these classical learning algorithms hardly find any informative causal relation in the data.



(a) PC: Scenario 1. (b) PC: Scenario 2. (c) PC: Scenario 3.

(d) HC: Scenario 1. (e) HC: Scenario 2. (f) HC: Scenario 3.

Figure 3: Recovered DAGs using the PC and the HC algorithms.



(a) Ground truth SCM from [Sachs *et al.*, 2005]. (b) Recovered C-DAG using the proposed method.

Figure 4: Results on human immune system protein-signaling networks. The recovered C-DAG in (b) complies with the true DAG in (a). The group containing Plcg, PIP3, and PIP2 is well discovered.

5.2 Protein-Signaling Networks

As an example of real-world application, we apply our method to the protein signaling dataset [Sachs *et al.*, 2005], which contains the expression levels of $n = 11$ proteins and phospholipids in human immune system cells, with $N = 7466$ observations. The cellular protein measurements are continuous random variables. To efficiently estimate the MI, we apply the traditional kNN-based non-parametric estimator [Kraskov *et al.*, 2004] with $k = 5$. We run our algorithms with $\rho = 0.5$ and $Z = 500$. The resulting C-DAG, shown in Figure 4b, complies with definition (2). In particular, the algorithm successfully discovered the two groups of closely related molecules, {Plcg, PIP3, PIP2} and {PKC, PKA, Jnk, Raf, P38, Mek, Erk, Akt}, in the biological process, as expected from the ground truth.

6 Conclusion

To represent causal relations among clusters of random variables, the *cluster causal diagrams* (C-DAGs) extend the approach of the Bayesian networks to causal inference. In this paper, we show that the faithfulness assumption, enforced in most BN learning algorithms, can be relaxed using C-DAGs. We propose an information-theoretic scoring function, the Clustering Information Criterion (CIC), that represents how well the C-DAG structure fits the data. To effectively discover C-DAGs of high CIC scores, we develop a greedy searching strategy. We implement our method to recover C-DAGs using both synthetic and real data that involves complex interactions among groups of variables. We believe this work is a positive first step towards a tractable method for learning cluster causal diagrams from observational data.

References

[Anand *et al.*, 2022] Tara V Anand, Adele H Ribeiro, Jin Tian, and Elias Bareinboim. Effect identification in cluster causal diagrams. *arXiv preprint arXiv:2202.12263*, 2022.

[Battiston *et al.*, 2020] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks

beyond pairwise interactions: structure and dynamics. *Physics Reports*, 874:1–92, 2020.

[Chen *et al.*, 2021] Xinshi Chen, Haoran Sun, Caleb Ellington, Eric Xing, and Le Song. Multi-task learning of order-consistent causal graphs. *Advances in Neural Information Processing Systems*, 34, 2021.

[Chow and Liu, 1968] CK Chow and CN Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[Correa and Bareinboim, 2020] Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of AAAI*, pages 10093–10100, 2020.

[de Campos and Cozman, 2013] Cassio de Campos and Fabio Cozman. Complexity of inferences in polytree-shaped semi-qualitative probabilistic networks. In *Proceedings of AAAI*, pages 217–223, 2013.

[de Campos, 2006] Luis M de Campos. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(10), 2006.

[Heckerman *et al.*, 1995] David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[James *et al.*, 2016] Ryan G James, Nix Barnett, and James P Crutchfield. Information flows? a critique of transfer entropies. *Physical Review Letters*, 116(23):238701, 2016.

[Kalisch and Bühlman, 2007] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[Kraskov *et al.*, 2004] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6), 2004.

[Liu *et al.*, 2017] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen Change Loy, and Xiaoou Tang. Deep learning markov random field for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1814–1828, 2017.

[Lizier *et al.*, 2018] Joseph T Lizier, Nils Bertschinger, Jürgen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work, 2018.

[McGill, 1954] William McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.

[Mokhtarian *et al.*, 2021] Ehsan Mokhtarian, Sina Akbari, Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Learning Bayesian networks in the presence of structural side information. *arXiv preprint arXiv:2112.10884*, 2021.

[Morningstar and Melko, 2018] Alan Morningstar and Roger G Melko. Deep learning the ising model near criticality. *Journal of Machine Learning Research*, 18:1–17, 2018.

[Nabi *et al.*, 2020] Razieh Nabi, Joel Pfeiffer, Murat Ali Bayir, Denis Charles, and Emre Kıcıman. Causal inference in the presence of interference in sponsored search advertising. *arXiv preprint arXiv:2010.07458*, 2020.

[Niu and Quinn, 2019] Xueyan Niu and Christopher J Quinn. A measure of synergy, redundancy, and unique information using information geometry. In *Proceedings of ISIT*, pages 3127–3131, 2019.

[Parviainen and Kaski, 2017] Pekka Parviainen and Samuel Kaski. Learning structures of bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88:110–127, 2017.

[Pearl and Bareinboim, 2014] Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4), 2014.

[Pearl, 2000] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[Reing *et al.*, 2021] Kyle Reing, Greg Ver Steeg, and Aram Galstyan. Influence decompositions for neural network attribution. In *Proceedings of AISTATS*, pages 2710–2718, 2021.

[Robins *et al.*, 2003] James M Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

[Sachs *et al.*, 2005] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[Saeed *et al.*, 2020] Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. Causal structure discovery from distributions arising from mixtures of DAGs. In *Proceedings of ICML*, pages 8336–8345, 2020.

[Schwarz, 1978] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, Prediction, and Search*. MIT Press, 2000.

[Tikka *et al.*, 2021] Santtu Tikka, Jouni Helske, and Juha Karvanen. Clustering and structural robustness in causal diagrams. *arXiv preprint arXiv:2111.04513*, 2021.

[Tsamardinos *et al.*, 2006] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[Uhler *et al.*, 2013] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

[Wellman, 1990] Michael P Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990.