Causal Structure Learning when Data is not Independent and Identically Distributed (non-IID)

by

Shishir Adhikari

A written critique submitted as partial fulfillment of the requirements for

Ph.D. Qualifier Examination

Department of Computer Science
College of Engineering

University of Illinois at Chicago
Chicago, Illinois

February 2020

Abstract

Many disciplines of sciences are concerned with the task of understanding the causal data generating mechanism and making prediction or inference when naturally occurring mechanisms are subjected to external interventions. Causal structure learning concerns with developing a structural representation of underlying causal interactions and inferring the causal relations from observational data, optionally using domain knowledge or experimental data. The conventional causal Bayesian network and the structure learning algorithms built on top of it assume the independence and identical distribution (IID) of observations. However, real-world data have intrinsic semantic, temporal, and causal couplings as well as heterogeneity in the distribution. These real-world data complexities may lead to unreliable or misleading conclusions when applying the models with the IID assumption. We explore the existing work on causal structure learning from non-IID data. First, we introduce the relational model that captures the entities, relationships, and attributes with their causal dependencies. Also, the relational causal discovery (RCD) algorithm for learning the causal relationship between the attributes in the model is discussed. Next, we will present causal structure learning for heterogeneous or non-stationary environments where the data collection conditions vary or data distribution changes over time. Lastly, we investigate the causal structure learning in the multivariate time series settings where a variable is auto-correlated and influenced by past or contemporary values of other variables.

Contents

Abstract ii

List of Tables v

List of Figures vi

1 Introduction 1
- 1.1 Causal Structure Learning . . . 1
- 1.2 Real-world Data Complexities . . . 1
- 1.3 Research Motivation . . . 2
- 1.4 Overview and Organization . . . 2

2 Background 4
- 2.1 Notations and Terminologies . . . 4
- 2.2 Structural Causal Models . . . 4
- 2.3 Conditional Independence . . . 5
  - 2.3.1 Local Markov Condition . . . 5
  - 2.3.2 d-separation . . . 6
  - 2.3.3 Markov Equivalence Class . . . 6
  - 2.3.4 Statistical Independence Tests . . . 7
- 2.4 Causal Structure Learning Problem . . . 7
  - 2.4.1 Faithfulness Assumption . . . 7
  - 2.4.2 Causal Sufficiency Assumption . . . 8
  - 2.4.3 Problem Definition . . . 8
- 2.5 Causal Structure Learning Algorithms . . . 8
  - 2.5.1 Constraint Based . . . 8
  - 2.5.2 Score Based . . . 8
  - 2.5.3 Hybrid . . . 9
  - 2.5.4 Functional Causal Model Based . . . 9

3 Causal Structure Learning from Relational Data 10
- 3.1 Representation . . . 10
  - 3.1.1 Relational Model . . . 10
  - 3.1.2 Abstract Ground Graphs . . . 12
- 3.2 Methodology . . . 12
  - 3.2.1 Skeleton Estimation . . . 13
  - 3.2.2 Causal Orientation Determination . . . 13
- 3.3 Critique . . . 15
  - 3.3.1 Causal Structure Learning Approach . . . 15

3.3.2 Contributions and Limitations . . . 15

4 Causal Structure Learning from Nonstationary/Heterogeneous Data 16
4.1 Representation . . . 16
4.1.1 Misleading Conclusion with Traditional Approach . . . 16
4.1.2 Assumptions . . . 17
4.1.3 SEM for Non-identical Distribution . . . 17
4.2 Methodology . . . 17
4.2.1 Skeleton Estimation . . . 18
4.2.2 Causal Orientation Determination . . . 18
4.3 Critique . . . 19
4.3.1 Causal Structure Learning Approach . . . 19
4.3.2 Contributions and Limitations . . . 19

5 Causal Structure Learning from Time series Data 21
5.1 Representation . . . 21
5.1.1 SVAR and Dynamic DAGs . . . 21
5.1.2 Dynamic DAGs with Latent Variables . . . 22
5.2 Methodology . . . 23
5.2.1 Skeleton Estimation . . . 23
5.2.2 Causal Orientation Determination . . . 24
5.3 Critique . . . 24
5.3.1 Causal Structure Learning Approach . . . 24
5.3.2 Contributions and Limitations . . . 24

6 Discussion and Further Research Direction 25
6.1 Discussion . . . 25
6.2 Further Research Direction . . . 27

Bibliography 28

List of Tables

2.1 Common terminologies and their alternatives used interchangeably within this paper 4

3.1 Synopsis of RCD algorithm in the context of causal structure learning 15

4.1 Synopsis of CD-NOD algorithm in the context of causal structure learning 20

5.1 Synopsis of SVAR-FCI and SVAR-GFCI algorithms in the context of causal structure learning 24

6.1 Comparison of main concepts in Maier et al. (2013a), Zhang et al. (2017), and Malinsky and Spirtes (2018) 26

List of Figures

2.1 Three typical sub-graphs of a DAG capturing conditional independence 6
2.2 Equivalence class of DAG satisfying $A \perp\!\!\!\perp C \mid B$ 7

3.1 Example Relational Model for Movie Industry Domain 10
3.2 Relational skeleton showing instantiation of the example Relational Model 11
3.3 A ground graph for the example relational model and skeleton presented in Figures 3.1 and 3.2 respectively 12
3.4 An abstract ground graph from (a) ACTOR perspective and (b) MOVIE perspective for the example relational model and ground graph presented in Figures 3.1 and 3.3 respectively 12
3.5 Edge orientation rules used in RCD algorithm 14

4.1 An illustration of misleading conclusion with traditional approach. (a) True causal model with confounding effect of non-stationary/heterogeneous data (b) The estimated skeleton graph in asymptotic case 16
4.2 Two possible situations when $V1 \to V2$ and both $V1$ and $V2$ are found to be changing modules (adjacent to C)(a) Independent changing parameters (b) A confounder in addition to changing parameters 19

5.1 Dynamic DAGs with latent confounder interactions (a) L_1 is called an "auto-lag" confounder and L_2 may be called "contemporaneous confounder" (b) L_3 may be called "cross-lag confounder" 22

5.2 Dynamic MAGs for Marginalized dynamic DAG in Figure 5.1 (a) and (b) 23

1. Introduction

1.1 Causal Structure Learning

Recently the performance of data science and artificial intelligence systems have improved in several tasks such as classification, regression, object detection, language understanding, and strategic game playing. The success of these systems is mainly due to the availability of large annotated datasets, the emergence of powerful non-linear discriminative models, speed-up gains obtained with hardware acceleration, and flexibility of software frameworks. However, many disciplines of sciences are concerned with the task of understanding the causal data generating mechanism and making prediction or inference when naturally occurring mechanisms are subjected to external interventions. The past three decades have seen a number of conceptual developments and some partial solutions toward the task of learning the underlying causal data generating mechanism (Spirtes, 2010). Still, the practical application of these developments is limited by additional challenges such as domain-specific mechanisms, unavailability of ground truth, the presence of unmeasured causes, a large number of variables, and heterogeneity and bias present in big data. A traditional approach for determining the cause-effect relationship is a randomized experiment where the value of a treatment variable is manipulated randomizing over other variables to record the average effect in the outcome variable. A well-known example is a medical drug trial, where the decision of giving a drug or placebo to the participants is determined by random for testing the effectiveness of the drug. However, it is often infeasible to conduct such experiments due to the limited resources (time and cost), ethical issues, and a large number of variables. Therefore, the task of inferring the causal relations from observational data, optionally using domain knowledge or experimental data, is known as Causal Discovery. The casual relations are effectively represented with a directed acyclic graph (DAG) $G(\mathbf{V}, \mathbf{E})$ where each vertex denotes a random variable and a directed edge $A \rightarrow B$ denotes A is the direct cause of B. A causal discovery problem can be framed as a problem of learning a $N \times N$ adjacency matrix where $(i, j)^{th}$ element signifies whether i^{th} variable is the direct cause of j^{th} variable, and N is the number of variables. Thus causal discovery is interchangeably termed as Causal Structure Learning (CSL) or Causal Structure Discovery (CSD).

1.2 Real-world Data Complexities

Most of the learning problems in statistics and machine learning assume that features or random variables are independent and identically distributed (IID or i.i.d.). The independence assumption considers the samples are not dependent on one another. Similarly, the identical distribution assumption posits the underlying distribution of each random variable or feature is the same for all samples. The IID assumption is used in most of the statistical learning algorithms, including algorithms for causal structure learning (Glymour et al., 2019), for the theoretical guarantee of correctness and completeness. Moreover, the IID assumption is the foundation of model selection

techniques such as cross-validation and bootstrapping in statistics and machine learning. However, the IID assumption is rarely satisfied in real-world scenarios. There are natural coupling mechanisms where an observation is linked to other observations. Similarly, the distribution of data may change with time as well as domain. Some examples of couplings (Cao, 2014) found in data are:

- **Serial or Temporal Coupling**: The observations are related by order of occurrence (in time) where one event takes place only after another event. For example, a retweet can be done only after a tweet is posted.
- **Semantic and Syntactic coupling**: The observations have semantic and/or syntactic relationship. For example, comments in a Facebook post are related to the post by design.
- **Causal coupling**: The observations have a cause-effect relationship. For example, a piece of breaking news may cause a trend of related tweets.

Although the coupling examples presented above are in the context of social network data, these coupling mechanisms can be found in data sources from other domains as well. In addition to the coupling between the data samples, there may also be a distribution shift of the features or random variables. The data may follow a seasonal trend (for example, clothes sales data). Also, the data collection conditions and experimental settings may be different across domains and datasets. The data where the independence and identically distributed assumption does not hold is known as non-IID data. The complex couplings and heterogeneity in data distribution may lead to unreliable or misleading conclusions when traditional structure learning algorithms with IID assumptions are applied to non-IID data.

1.3 Research Motivation

The problem of maximally learning the causal structure from data is both NP-hard (Chickering et al., 1994) and NP-complete (Chickering, 1996). However, the adjacency in real-world causal graphs is generally assumed to be sparse. Also, background domain knowledge is available in some cases that can be incorporated into the learning problem. The past three decades of research have produced theoretical foundations, conceptual developments, and computational methods for causal structure learning (Spirtes et al., 2000; Pearl, 2009). Several compelling real-world applications (Borboudakis and Tsamardinos, 2016; Oktay et al., 2010; Lagani et al., 2016) of causal structure learning in the field of social sciences, natural sciences, and engineering have emerged. On the other hand, several data complexities such as non-IIDness, selection bias, missing data, and noisy data create challenges for learning reliable and robust causal structures from observational data. Thus, the area of causal structure learning has both challenges and opportunities that require non-trivial solutions.

1.4 Overview and Organization

In this work, I examine causal structure learning from non-IID data specifically relational (Maier et al., 2013a), non-stationary/heterogeneous (Zhang et al., 2017), and time-series (Malinsky and Spirtes, 2018) data. I gather the evidence from the sources to show conventional approaches with IID assumption produce inconsistent results when applied to non-IID data. Also, most importantly, the sources reveal the non-IID nature of data provides additional knowledge for learning causal structure.

I present background on conceptual developments and assumptions for causal structure learning from observational data in Chapter 2. I briefly discuss the representation of causal structure with

a causal Bayesian network and structural equation model (SEM), conditional independence tests, common assumptions for learning causal structure from data, and classification of causal structure learning algorithms. Chapter 3 reviews an approach to causal structure learning from relational data with an entity-relationship-attribute schema. Chapter 4 summarizes the representation of non-stationary or heterogeneous data for CSL as well as the proposed algorithms to learn the adjacency (skeleton) graph and orientation of edges. Chapter 5 reports the CSL methods for multivariate time series data as well as the representation of unmeasured common causes and contemporaneous causal relationships in time series data. Moreover, a critique highlighting the strengths and limitations of each CSL approach is presented after the summarization in Chapters 3 to 5. Finally, I discuss the collective contributions of each paper toward learning causal structure from non-IID data. Moreover, I consider the limitations and challenges in learning causal structure from non-IID data and develop further research direction and potential ideas for practical applications in the real world.

2. Background

2.1 Notations and Terminologies

This section provides standard notation used throughout the paper unless otherwise specified as well as defines common terminologies with alternatives. The symbols with boldface signify sets. The random variables are denoted by uppercase letters and indexed with a subscript. For example, X is a random variable, V_i is a random variable from a set $\mathbf{V}$ and $X_{,t}$ and $V_{i,t}$ are time-indexed random variables. The instances of random variables are represented by lower case letters and follow similar indexing. $G(\mathbf{V},\mathbf{E})$ is a graph with a set of vertices $\mathbf{V}$ and edges $\mathbf{E}$. $A \rightarrow B$ denotes an edge E from A to B in a graph G. The node A is the parent of the node B. The symbol $pa(V_i)$ indicates a set of nodes that are parents of node V_i. Similarly, B is a child of A. A directed path is a sequence of nodes obtained following the direction of the edges. The nodes preceding the tail node of the directed path are the ancestors of the tail node. Similarly, the nodes following the head node of the directed path are descendants of the head node. The symbols $anc(V_i)$ and $des(V_i)$ indicate the sets of ancestors and descendants of V_i respectively. A directed graph with no paths starting and ending on the same node is known as a directed acyclic graph (DAG).

The Table 2.1 below summarizes common terminologies with their alternatives and description. The terminology and its alternatives are used interchangeably.

Table 2.1: Common terminologies and their alternatives used interchangeably within this paper

Terminology	Alternatives	Description
Causal Structure Learning	Causal Discovery, Causal Structure Discovery	The problem of learning causal relationships using observational data and optionally domain knowledge or experimental data
Causal Graph	Causal Bayesian Network, Causal DAG, Causal Graphical Model, Graphical Model	A graph that represents random variables and captures causal relationships as well as conditional independence
Vertex	Node	A vertex V of a Graph $G(\mathbf{V},\mathbf{E})$
Latent Confounder	Hidden Confounder, Unmeasured Confounder	Hidden common cause of two or more variables
Structural Equation Model	Functional Causal Model	A model that defines a variable as a function of its direct causes and error term

2.2 Structural Causal Models

A structural causal model (SCM) (Pearl et al., 2016) is used to model the causal assumptions by representing the relevant features and their interaction. An SCM models how nature assigns values to the features of interest using a set of variables $\mathbf{U},\mathbf{V}$, and a set of functions $\mathbf{f}$ that assign values to each variable in $\mathbf{V}$ using other variables. The variables $\mathbf{U}$ termed as exogenous variables are external to the causal model and often considered errors or disturbances. The causal relations of the endogenous variables V are explained by the functions of at least one exogenous variable and optionally other endogenous variables. The SCMs are either represented by Structural Equation Models (SEM) or Causal Graphical Models.

Structural Equation Model The causal effect on a variable can be explained by a function of its known direct causes and unknown disturbances i.e.

$$V_c = f_c(\mathbf{V}^{(\mathbf{c})}, \mathbf{U}^{(\mathbf{c})}) \tag{2.1}$$

, where $\mathbf{V}^{(\mathbf{c})} \subset \mathbf{V}$ is a set of direct causes of variable of interest V_c and $\mathbf{U}^{(\mathbf{c})} \subset \mathbf{U}$ is a set of unmeasured disturbances. The function f_c can be any linear or non-linear function. The function f_c can also be unknown for a non-parametric model.

Causal Graphical Model The causal relationship among the random variables can also be represented using a DAG where a variable has incoming edges from its direct causes and unmeasured disturbances. Unless otherwise specified, the unmeasured disturbances of each variable are assumed to be independent with other disturbances and not shown explicitly in the graph. For the SEM presented in Equation 2.1, V_c has edges from $\mathbf{V}^{(\mathbf{c})}$ and $\mathbf{U}^{(\mathbf{c})}$ (implicitly). The DAG is, in fact, a causal Bayesian Network with nodes representing the random variables and edges representing causal direction. In addition to the intuitive representation of causality, causal Bayesian Networks capture the joint probability of all the random variables in the network. The joint probability of Bayesian Network factorizes to the product of the conditional probability of each random variable given its parents. Formally,

$$P(V_1, V_2, ..., V_n) = \prod_{i=1}^{n} P(V_i | pa(V_i)) \tag{2.2}$$

This factorization saves huge space needed for the joint probability table which can be replaced by much smaller conditional probability tables (CPT) for each variable. Moreover, the Bayesian Network captures independence relationship between random variables which is discussed in next section.

2.3 Conditional Independence

2.3.1 Local Markov Condition

The factorization in equation 2.2 follows from the chain rule of probability theory and conditional independence (CI) relations. Both the structural equation and graphical models are constructed such that a random variable is the function of its direct causes and disturbances. Thus, a random variable is independent of all other variables expect its parents and its descendants. The influence of a random variable can flow to its child and then to the descendants. However, given the parents, the ancestor variables are independent of a random variable. This conditional independence property is formally known as Local Markov Conditions as defined in equation 2.3.

$$V_i \perp\!\!\!\perp \mathbf{V} \setminus \{V_i, pa(V_i), des(V_i)\} | pa(V_i) \tag{2.3}$$

These conditions are, however, not all the conditional independence relationships captured by the model. The next section describes the rules for dependency separation (d-separation) that capture all the conditional independencies encoded in a graphical model.

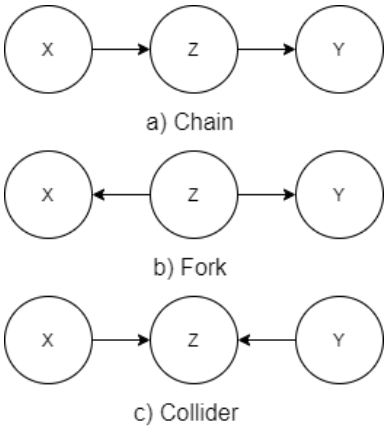


Figure 2.1: Three typical sub-graphs of a DAG capturing conditional independence

2.3.2 d-separation

Two random variables are likely dependent if there is an edge between those variables. This implies those random variables may also be independent under some data distribution. The dependency separation (d-separation) finds all the conditional independencies that hold for any data distribution that is generated by the mechanism described by a graphical model. The conditional independence relationships for the following three graphical structures help to formally define the rules of d-separation.

Chain Figure 2.1(a) shows a chain structure that is a subgraph of a DAG with a unidirectional path. The random variables X and Y are independent conditioned on Z i.e. $X \perp\!\!\!\perp Y \mid Z$.

Fork The fork structure is shown in Figure 2.1(b) where random variable Z is a common cause of variables X and Y. The variables X and Y are independent when Z is observed i.e. $X \perp\!\!\!\perp Y|Z$.

Collider Figure 2.1(c) depicts a collider structure where variable X and Y have common effect on variable Z. The independence relation for a collider is a bit different compared to chains and forks. X and Y are marginally independent when Z is not observed. However, when either Z or its descendants are observed, X and Y are likely dependent.

An undirected path is said to be **blocked** by a node Z with a conditioning set $\mathbf{S}$ of observed variables if one of the two conditions hold: (i) $Z \in \mathbf{S}$ and Z is not a collider or (ii) Z is a collider and neither Z nor its descendants belong to conditioning set $\mathbf{S}$. Two nodes are said to be d-separated by a conditioning set $\mathbf{S}$ if and only if all the paths between the nodes are blocked by $\mathbf{S}$ (Guo et al., 2018). The d-separated nodes are independent of one another conditioned on set $\mathbf{S}$. Given a causal DAG, all the conditional independencies encoded by the model are given by the d-separation rule. These sets of conditional independencies can be compared against the independencies in the data for accessing the fitness of the model.

2.3.3 Markov Equivalence Class

A single conditional independence relation can be satisfied by multiple causal graphical models. Figure 2.2 demonstrates three causal graphs that satisfy the independence relationship $A \perp\!\!\!\perp C \mid B$. These DAGs satisfying the same set of conditional independencies are referred to as an equivalence

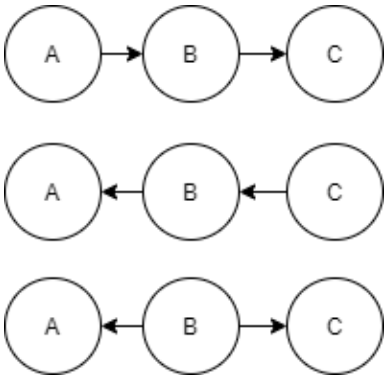


Figure 2.2: Equivalence class of DAG satisfying $A \perp\!\!\!\perp C \mid B$

class or Markov equivalence class. Thus, conditional independencies from data enable us to learn the equivalence class of a DAG.

2.3.4 Statistical Independence Tests

The translation of conditional independence to the structure of graphical models makes statistical independence tests the backbone for learning causal relationships. Typically statistical independence tests answer, with a p-value, whether two random variables are independent conditioned on a possibly empty set **S** using the observations. The null hypothesis of the statistical test is the variables are independent conditioned on **S**. The choice of statistical tests depends on the nature of random variables such as continuous, discrete or mixed data distribution, linear/non-linear SEM assumption, computational complexity, data dimensionality, and so on (Strobl et al., 2019).

2.4 Causal Structure Learning Problem

In addition to causal relationships, an SCM can quantify the actual causal effects on a random variable with the conditional probability tables (CPT) in graphical models or the function parameters in SEM. The task of causal structure learning, however, centers on only learning the skeleton graph and the orientation of the edges. In other words, a causal structure learning concerns with learning a non-parametric model. However, learning causal structure from the data is possible only with some assumptions about the data.

2.4.1 Faithfulness Assumption

The skeleton estimation step concerns with determining whether the random variables are connected by an edge. If there is an edge between two variables in a graphical model, we can say these variables are **likely dependent**. However, their dependence cannot be guaranteed as there may exist some data distribution where these variables are independent. Hence, an assumption called the faithfulness assumption has to be made to construct the skeleton graph from the data. The faithfulness assumption presumes all the conditional independencies observed from the data are entailed by the d-separation conditions of a causal graph.

2.4.2 Causal Sufficiency Assumption

The exogenous variables $\mathbf{U}$ of an SCM are associated with each endogenous variables $\mathbf{V}$. The exogenous variables are not explicitly included in causal graphs as all the variables in $\mathbf{U}$ are assumed to be independent of one another. In other words, it is assumed that there are no unmeasured common causes of variables in $\mathbf{V}$. This assumption is known as the causal sufficiency assumption.

2.4.3 Problem Definition

With the faithfulness and causal sufficiency assumption, we can say V_i and V_j in a DAG $G(\mathbf{V}, \mathbf{E})$ are adjacent if and only if V_i and V_j are dependent conditional on every subset of $\mathbf{V} \setminus \{V_i, V_j\}$. The causal sufficiency can be a strong assumption as the chances of having one or more unmeasured confounders are relatively high in many domains. Some causal structure learning methods (Spirtes et al., 2000, p. 123) can handle the unmeasured confounders by introducing a bidirectional edge. The general problem formulation for learning a causal DAG under faithfulness and causal sufficiency assumption is simply learning a $N \times N$ adjacency matrix where $N(i, j) = 1$, if V_i is the direct cause of V_j, and $N = |\mathbf{V}|$. The formulation above can be expanded to the structures with unmeasured confounders by allowing the matrix element to have multiple values reflecting if an edge is due to direct cause or confounding effect (Glymour et al., 2019).

2.5 Causal Structure Learning Algorithms

Several causal structure learning algorithms have been developed in the past few decades. These algorithms can be summarized by categorizing them according to their approach.

2.5.1 Constraint Based

The first class of algorithms called Constraint-Based algorithms use the statistical tests to check the independencies in the data and estimate the skeleton and orientation of the causal structure using these independencies as constraints. These algorithms generally start with a complete undirected graph and construct a skeleton adjacency graph and then orient the edges using conditional independencies and faithfulness assumption. These algorithms have flexibility to include domain knowledge as constraints as well. Peter-Clark (PC) and Fast Causal Inference (FCI) (Spirtes et al., 2000) are commonly used constraint based algorithms.

2.5.2 Score Based

Score based algorithms consider the causal discovery problem as fitting a causal graph to the data. A relevant score function is defined that relates how well the graph captures the conditional independencies in the data. These algorithms relax the faithfulness assumption but the causal sufficiency assumption is made. Greedy Equivalence Search (GES) (Chickering, 2002) is a popular score based causal structure learning algorithm that starts with an empty graph and keeps adding the edges that increase the goodness of fit score. The algorithm then removes the edges to return the equivalence class of DAGs with maximum score.

2.5.3 Hybrid

This class of algorithms combines both constraint-based and score-based approaches to utilize the conditional independencies tests as well as scoring functions. GFCI (Ogarrio et al., 2016), a hybrid algorithm, uses GES like scoring for skeleton determination and FCI for pruning the skeleton and orienting the edges.

2.5.4 Functional Causal Model Based

These approaches use non-linear SEM and are typically used to determine the direction of edges in case of two variables (Glymour et al., 2019). The general principle is that the regression in causal direction makes error terms independent with direct causes. However, regressing in anti-causal direction makes error terms correlated.

3. Causal Structure Learning from Relational Data

Causal Bayesian networks provide an intuitive representation of causal relationships and encode the conditional independencies between random variables. Moreover, Bayesian networks can compactly represent the joint probability distribution over the set of variables as conditional probability tables due to the factorization property of equation 2.2. However, the Bayesian networks assume the data instances are independent and identically distributed (IID). The real-world data-generating mechanisms mostly involve various heterogeneous entities with complex relationships. The Bayesian network is not rich enough to encode the relational data involving diverse entities and their relationships. For example, social network data consists of individuals, contents and their interactions. All the random variables in a Bayesian network are implicitly assumed to belong to a single entity or unit. The research in the probabilistic graphical models has developed more expressive representations such as probabilistic relational model (PRM) (Getoor et al., 2007) which removes the assumption of IID instances. The development of theory on representation and conditional independencies in the relational models (Maier et al., 2013b) has enabled constraint-based causal structure learning from relational data (Maier et al., 2013a). In this chapter, I summarize the work by Maier et al. (2013a) on the representation of relational models and the approach for constraint-based causal discovery using relational d-separation.

3.1 Representation

3.1.1 Relational Model

Figure 3.1: Example Relational Model for Movie Industry Domain

Relational models are designed to capture the real-world entities, their attributes as well as relationships with other entities. Let us take a running example by Maier et al. (2013a) that considers the movie industry domain with two entities ACTOR and MOVIE. Figure 3.1 depicts the relational model for the example domain. Each entity has one or more attributes and the entities are connected by a relationship with cardinality constraint. In the example above, "Popularity" and "Success" are the only attributes for the entities ACTOR and MOVIE respectively. The two entities are connected by a relationship "STARS-IN" with many-to-many cardinality. This means an actor can star in one or more movies and a movie can cast many actors. Moreover, the the directed edge from ACTOR.Popularity to MOVIE.Success signify the underlying causal mechanism where an actor's popularity causes success of the movie she starts in.

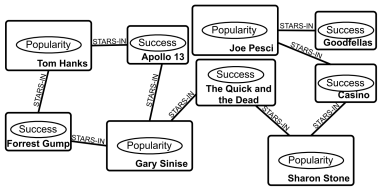


Figure 3.2: Relational skeleton showing instantiation of the example Relational Model

Relational Schema: The first component of a relational model is a relational schema $\mathcal{S} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}\}$ that describes the entities, relationship, and attribute classes as well as contains the cardinality constraints. It is typically represented as an entity-relationship (ER) diagram similar to figure 3.1 except for the causal edges.

Relational Skeleton: A relational skeleton σ is an instantiation of the relational schema. The instances of entities interact according to relationship and cardinality constraints. Figure 3.2 shows the relational skeleton for the example model.

Relational Path: A relational path is an alternating sequence of entities and relationship classes according to the schema subjected to cardinality constraints. Some relational paths for the model in figure 3.1 are $[ACTOR], [ACTOR, STARS-IN, MOVIE], [ACTOR, STARS-IN, MOVIE, STARS-IN, ACTOR]$ signifying a single actor, an actor starring in a movie, and co-actors of a movie respectively

Relational Variables: A relational variable is defined by a pair of a relational path and an attribute of the class ending on the relational path. For example: $[ACTOR].Popularity, [ACTOR, STARS-IN, MOVIE].Success, [MOVIE, STARS-IN, ACTOR].Popularity$ and so on are the relational variables.

Relational Dependency: A relational dependency is a pair of relational variables with a common first item termed as *perspective*. For example, the pair of relational variables $[MOVIE, STARS-IN, ACTOR].Popularity \rightarrow [MOVIE].Success$ suggests the success of a movie is dependent on the popularity of actors starred in the movie. A dependency is said to be a canonical dependency if the effect variable has only one entity in the relational path.

Relational Model: Formally a relational model $\mathcal{M} = (\mathcal{S}, \mathcal{D})$ is a collection of relational dependencies $\mathcal{D}$ in a canonical form defined over a schema $\mathcal{S}$. Like Bayesian networks, relational models are parameterized by a set of conditional probability distribution one for each relational variable $\mathcal{A}(\mathcal{I})$, where $\mathcal{I} \in \mathcal{E} \cup \mathcal{R}$.

Ground Graph: A ground graph $GG_{\mathcal{M}\sigma}$ is a model instantiation produced when the relational model M is paired to the relational skeleton σ. The ground graph is a directed graph with a node for all the attributes of each instance and an edge between the instances of relational variables participating in relational dependencies. Figure 3.3 depicts a ground graph for the example relational model and skeleton presented in Figures 3.1 and 3.2 respectively. The ground graph follows similar factorization as in Bayesian networks and given by Equation 3.1

$$P(GG_{\mathcal{M}\sigma}) = \prod_{I \in \{\mathcal{E}, \mathcal{R}\}} \prod_{X \in \mathcal{A}(I)} \prod_{i \in \sigma(I)} P(i.X | pa(i.X)) \tag{3.1}$$

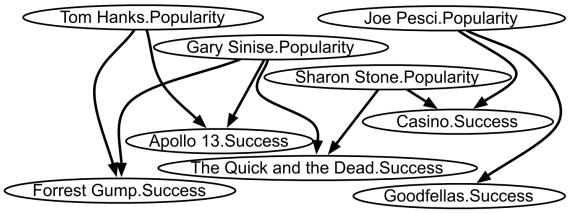


Figure 3.3: A ground graph for the example relational model and skeleton presented in Figures 3.1 and 3.2 respectively

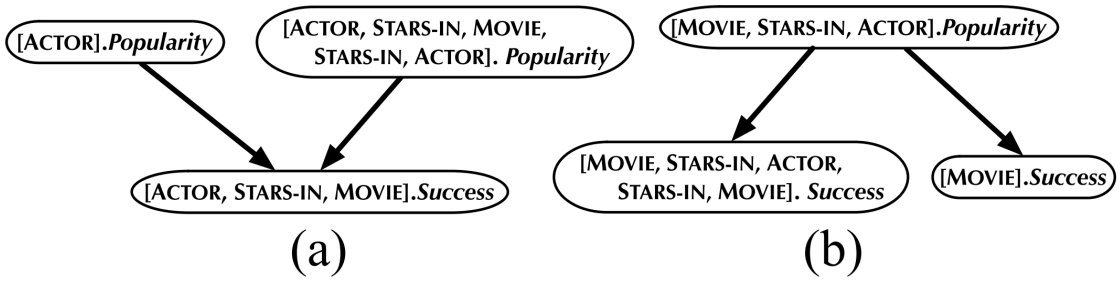


Figure 3.4: An abstract ground graph from (a) ACTOR perspective and (b) MOVIE perspective for the example relational model and ground graph presented in Figures 3.1 and 3.3 respectively

3.1.2 Abstract Ground Graphs

The d-separation property of Bayesian networks does not work accurately when applied directly to relational models (Maier et al., 2013a). The ground graphs are dependent on all the instances of a relational model which may lead to a large number of nodes. Maier et al. (2013b) developed an abstract representation of ground graphs and showed the soundness and completeness of using d-separation to reason about conditional independencies encoded in those representations. An abstract ground graph $AGG_{\mathcal{M}Bh}$ is a directed graph defined for a relational model $\mathcal{M}$ with perspective B and a scalar hop threshold h that captures the dependency among relational variables for any possible ground graph. The number of nodes in an $AGG_{\mathcal{M}Bh}$ depends on the number of relational variables starting with perspective B and having a maximum of h path length. An edge is present between the relational variables in an abstract ground graph if there exists any ground graph with relational dependency between those variables. Figure 3.4 shows the abstract ground graphs from the perspective of ACTOR and MOVIE with a hop limit of 4. As seen from the figure, a single relational model can have multiple abstract ground graphs for different perspectives. Also, a single dependency in a relational model can be translated to multiple edges in an abstract ground graph.

3.2 Methodology

Maier et al. (2013a) proposed Relational Causal Discovery (RCD) algorithm for learning causal structure from relational data using abstract ground graphs (AGG) representation. RCD is a constraint-based algorithm that reasons about the conditional independencies in AGG using rules of relational d-separation. Similar to other constraint-based algorithms, RCD involves two phases: (i) skeleton estimation and (ii) edges orientation determination. Under causal faithfulness and

sufficiency assumption, the RCD is algorithm is shown to be sound and complete. Soundness and completeness of the algorithm provide a theoretical guarantee under the satisfaction of the assumptions and correct conditional independencies test, no other algorithm can find more orientations from the observational data. The pseudocode for RCD algorithm is outlined in Algorithm 1 and described in the following sections.

ALGORITHM 1: RCD($schema$, $depth$, $hopThreshold$, P)

```
PDs ← getPotentialDeps(schema, hopThreshold)
N ← initializeNeighbors(schema, hopThreshold)
S ← {}
// Phase I
for d ← 0 to depth do
    for X → Y ∈ PDs do
        foreach condSet ∈ powerset(N[Y] \ {X}) do
            if |condSet| = d then
                if X ⫫ Y | condSet in P then
                    PDs ← PDs \ {X → Y, Y → X}
                    S[X, Y] ← condSet
                    break
// Phase II
AGGs ← buildAbstractGroundGraph(PDs)
AGGs, S ← ColliderDetection(AGGs, S)
AGGs, S ← BivariateOrientation(AGGs, S)
while changed do
    AGGs ← KnownNonColliders(AGGs, S)
    AGGs ← CycleAvoidance(AGGs, S)
    AGGs ← MeekRule3(AGGs, S)
return getCanonicalDependencies(AGGs)
```

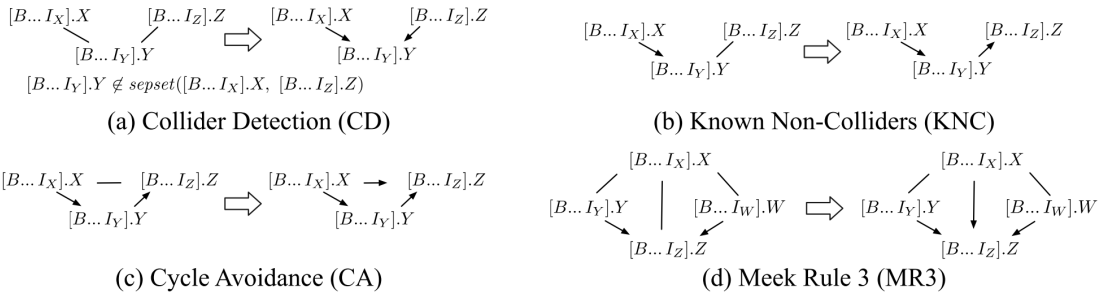
Algorithm 1: Relational Causal Discovery (RCD) Algorithm (Maier et al., 2013a)

3.2.1 Skeleton Estimation

The skeleton estimation for the RCD algorithm is similar to one employed by the constraint-based PC algorithm in terms of conditional independence tests performed. The PC algorithm starts with a complete graph and prunes edges if a pair of variables are found to be independent given a conditioning set. Unlike the PC algorithm, the RCD algorithm first estimates the potential dependencies (**PDs**) from all perspectives in a canonical form using the relational schema and hop threshold. The potential dependencies are then pruned using the conditional independence tests and a set **S** containing separating sets for each pruned dependency is maintained. The skeleton estimation phase gives a set of undirected dependencies from all perspectives (pruned **PDs**) and a set of conditioning variables **S**.

3.2.2 Causal Orientation Determination

The causal direction of edges is determined using orientation rules on the abstract ground graphs constructed with potential dependencies. The orientation rules reason about d-separation and condi-



(a) Collider Detection (CD)

(b) Known Non-Colliders (KNC)

(c) Cycle Avoidance (CA)

(d) Meek Rule 3 (MR3)

Figure 3.5: Edge orientation rules used in RCD algorithm

tional independence tests. First, all the colliders are determined and then orientation propagation rules are used iteratively until no edges can be oriented further. The orientation rules used by the RCD algorithm are described below. Let B be the prespective of AGG and I_W an entity or relationship with attribute W. So, $[B...I_W].W$ is a relational variable in the AGG.

Collider Detection The unshielded colliders can be detected using the d-separation property of colliders described in section 2.3.2. For simplicity let us consider a skeleton graph of three random variables $V_i - V_j - V_k$. The edges can be oriented as $V_i \rightarrow V_j \leftarrow V_k$ if V_i and V_k are independent when V_j is not in a conditioning set but dependent when V_j is included in the conditioning set. This can be extended to the relational variables of an AGG as depicted in Figure 3.5(a).

Bivariate Edge Orientation Bivariate edge orientation, a novel contribution by Maier et al. (2013a), is realized by autocorrelation between the entities involved in ONE-MANY or MANY-MANY relationship. Figure 3.4 shows the relational bivariate dependency where two relational variables of the same entity interact with another entity with different path lengths. Intuitively, Figure 3.4 (a) shows the dependency between co-actors' popularity and the movie's success. Let, $[I_X].X - [I_X...I_Y].Y - [I_X...I_Y...I_X].X$ be the skeleton graph. The edges are oriented as a collider if $[I_X...I_Y].Y$ is not in the separating set of $[I_X].X$ and $[I_X...I_Y...I_X].X$. Otherwise, the edges are oriented such that $[I_X...I_Y].Y$ is the common cause.

Known Non-collider Edge Orientation This rule is applied after the collider detection rules and works on the assumption that the orientation should not produce any new colliders. Again, for simplicity let us consider a skeleton graph of three random variables $V_i \rightarrow V_j - V_k$. The undirected edge should be oriented as $V_i \rightarrow V_j \rightarrow V_k$ to avoid a new collider. The extension to relational variables is depicted in Figure 3.5(b).

Cycle Avoidance This rule orients the edges such that directed cycles are avoided. Let, $V_i \rightarrow V_j \rightarrow V_k$ be the directed edges and $V_i - V_k$ be the undirected edge. The edge must be oriented as $V_i \rightarrow V_k$ to avoid the cycle. Figure 3.5(c) shows the extension of this rule to relational variables.

Meek Rule 3 This rule was introduced by Meek (1995) and is referred to as Meek Rule 3 by the authors. It reasons about the orientation of edge $V_x - V_z$ in presence of two partially oriented structures $V_x - V_w \rightarrow V_z$ and $V_x - V_y \rightarrow V_z$. In order to avoid a new collider and directed cycle, the edge must be oriented as $V_x \rightarrow V_z$. Figure 3.5(d) shows the extension for relational data.

3.3 Critique

In this section, I summarize the work by Maier et al. (2013a) in a broader context of causal structure learning and discuss the contributions and limitations of the approach.

3.3.1 Causal Structure Learning Approach

The Table 3.1 below provides the synopsis of the approach for causal structure learning from relational data.

Table 3.1: Synopsis of RCD algorithm in the context of causal structure learning

Concept	Description
Input	Relational Schema and Data
Representation Type	✔Graphical Models ✖SEM
Representation	Relational Model and Abstract Ground Graphs
Algorithm Type	✔Constraint-based ✖Score-based ✖Hybrid
Causal Assumption	✔Faithfulness ✔Sufficiency
Contribution by non-IIDness	Relational Bivariate Orientation
Output	Equivalence classes of AGG for each perspective

3.3.2 Contributions and Limitations

Maier et al. (2013a) provided a significant contribution in the area of relational causal discovery. RCD algorithm not only provided an approach to handle relational data for causal structure learning but also came with the theoretical guarantee of soundness and completeness making it equivalent to the PC algorithm for IID data. Moreover, the RCD algorithm was able to leverage the autocorrelation found between the entities involving in ONE-MANY or MANY-MANY relationship. This enabled a novel bivariate orientation rule for determining the direction of edges. The experimental results with the synthetic data showed this rule had a significant role in orienting the edges for relational data with multiple entities. The results also showed RCD performed better than the algorithms that used a modification of the PC algorithm for relational data. For the case with only one entity, the performance of the PC and RCD algorithm was similar. This justified the need for richer representations for relational data. Moreover, the RCD algorithm was applied to real-world data from the movie industry to discover interesting causal insights.

The RCD algorithm has some limitations due to assumptions made and the practicality of implementation. The algorithm takes account of relationships only with certain path length. Since casual structure learning is unsupervised, the knowledge of the hop limit may not be known apriori. Moreover, the RCD algorithm assumes the there are no latent confounders present in the data. However, this is a strong assumption and can be violated. The RCD algorithm uses a linear method for testing conditional independencies which may produce incorrect results for complex data distributions. The RCD algorithm takes relational schema as input to discover the naturally occurring causal mechanisms. However, the design of schema may be different across different domains for the same underlying mechanisms. This algorithm does not consider temporal data where there can be feedback creating cycles. Moreover, data measurement errors, selection bias, and missing data challenge the faithfulness assumption which is still an open problem for all causal discovery approaches.

4. Causal Structure Learning from Nonstationary/Heterogeneous Data

The availability of big data has created new challenges for causal discovery. Traditionally causal discovery algorithms were applied to a relatively small dataset with an identical distribution of random variables. However, the large dataset available in present days is integrated from multiple sources. Similarly, the data is retained over a comparatively longer period with fine granularity. Thus, it is more common to encounter heterogeneous and nonstationary data with underlying distribution changing across data sources and time. The heterogeneous and nonstationary data violate the IID assumption of Bayesian networks and most of the algorithms built on top of it. Moreover, the causal mechanism learned by a model is assumed to remain constant over time. In this chapter, I summarize the approach by Zhang et al. (2017) for representing and learning causal structure from heterogeneous and nonstationary data.

4.1 Representation

4.1.1 Misleading Conclusion with Traditional Approach

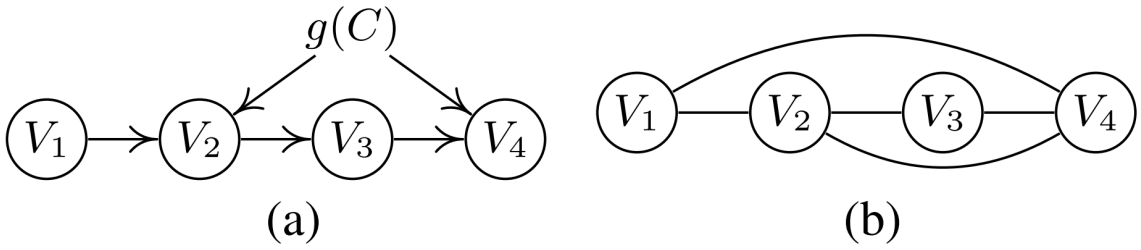


Figure 4.1: An illustration of misleading conclusion with traditional approach. (a) True causal model with confounding effect of non-stationary/heterogeneous data (b) The estimated skeleton graph in asymptotic case

Causal Bayesian networks are preferable models for representing the causal mechanisms as they encode the joint probability and conditional independencies of the data. However, Bayesian networks are assumed to have identically distributed random variables and a fixed probability distribution. These assumptions are violated for nonstationary and heterogeneous data. The distribution shift across time and domains may change the conditional independence found in the data and hence the causal models. The variation in the causal model may be due to change in the underlying data generating process, change in causal strengths, or level of disturbances. If a DAG $G(\mathbf{V}, \mathbf{E})$ represents the underlying causal model for each time point or domain, the joint

probability distribution factorizes to $P(\mathbf{V}) = \prod_i P(V_i \mid pa(V_i))$. For nonstationary or heterogeneous data at least some $P(V_i \mid pa(V_i))$ are changed with time and domain. These modules are called changing causal modules. Zhang et al. (2017) assume the quantities that change over time or domain can be represented as a function of time or domain index represented as C. The values of index C are available from the nonstationary dataset. Thus, the function $g(C)$ can be considered as a confounder affecting one or more modules simultaneously and responsible for non-identical distribution. Figure 4.1(a) depicts such assumption where true causal model is a chain from V_1 to V_4. The nodes V_2 and V_4 are simultaneously being affected by some unknown function of time or domain index. The Figure 4.1(b) shows skeleton graph where the confounding effect is responsible for false identification of $V_1 - V_4$ and $V_2 - V_4$ as adjacent nodes. Some workarounds proposed for handling such issues include using a sliding window of time or handling data of each domain separately. However, such approaches suffer due to the scarcity of samples and a high number of conditional independence tests. Zhang et al. (2017) formulate the problem to detect the changing causal modules all at once using complete data.

4.1.2 Assumptions

The representation of causal models for nonstationary and heterogeneous data should be developed such that the underlying causal model is allowed to change with time or domain. Zhang et al. (2017) allow unmeasured confounders in the model by dropping the causal sufficiency assumption. However, it is assumed that each confounder can be written as a smooth function of time or domain index. This assumption implies the confounders are fixed are a given time or domain. Such an assumption is termed as pseudo causal sufficiency assumption. Moreover, Zhang et al. (2017) assume the data is independent but not identically distributed. This means only instantaneous or contemporaneous causal relations are considered. Similar to other constraint-based algorithms, the data distribution is assumed to be faithful to the underlying causal model.

4.1.3 SEM for Non-identical Distribution

Let $\mathbf{g} = \{g_l(C)\}_{l=1}^{L}$ be a set of confounders represented as smooth function of domain index or time. Then, each random variable V_i is represented by the following structural equation model:

$$V_i = f_i(pa(V_i), \mathbf{g}^i(C), \theta_i(C), \varepsilon_i) \tag{4.1}$$

, where $\mathbf{g}^i(C) \subseteq \mathbf{g}$ are the confounders affecting V_i, $\theta_i(C)$ is time or domain dependent function only affecting V_i, and ε_i is the disturbance term independent of C. This equation can be converted to equivalent DAG by assuming C as a random variable and the joint probability distribution to be over $\mathbf{V} \cup \mathbf{g} \cup \{\theta_m(C)\}_{m=1}^{n}$. The graph formed after representing the SEM in equation 4.1 as a DAG is referred as augmented graph G^{aug}.

4.2 Methodology

The equation 4.1 and equivalently G^{aug} provide representation to capture the data with non-identical distribution. However, the joint probability distribution consists of hidden variables that cannot be measured. Thus, the time or domain index C is used as a surrogate variable for all the unobserved variables. This enables the application of conditional independence tests to $\mathbf{V} \cup \{C\}$ for recovering changing causal modules and skeleton graph. Moreover, the nonstationarity helps to provide

additional information regarding causal orientations. The skeleton estimation and edge orientation procedure of Constraint-based Causal Discovery from Nonstationary/heterogeneous Data (CD-NOD) (Zhang et al., 2017) algorithm is presented below.

4.2.1 Skeleton Estimation

The skeleton estimation phase uses conditional independence tests to reason about the adjacency of random variables. For the CD-NOD algorithm, skeleton estimation starts with a complete graph and prunes edges in two steps:(i) Detection of changing modules and (ii) Recovery of causal adjacency.

Detection of changing modules: In this step, the marginal and conditional independence between surrogate variable C and each random variable V_i is tested. If the variables are found to be independent given any conditioning set $\mathbf{S}$, the edge between the variables is removed. All the variables adjacent to node C after this procedure are the changing modules that are affected by some unobserved factors across time or domain.

Recovery of adjacency: The edge between two variables V_i and V_j is removed if equation 4.2 holds.

$$V_i \perp\!\!\!\perp V_j \mid \mathbf{S_k} \cup C \tag{4.2}$$

, where $\mathbf{S_k} \subseteq \{\mathbf{V} \setminus \{V_i, V_j\}\}$. The asymptotic correctness of the rule above follows from the assumption of distribution faithful to DAG and the assumption that all the unobserved variables $\mathbf{g} \cup \{\theta_m(C)\}_{m=1}^{n}$ are a deterministic function of the index C. Since the deterministic function of index C may to be a complex non-linear function Zhang et al. (2017) recommend kernel-based non-parametric tests (Zhang et al., 2011) for the correctness of conditional independence test.

4.2.2 Causal Orientation Determination

In addition to the Collider Detection(CD), Known Non-Collider (KNC) and Cycle Avoidance (CA) rules used in SGS algorithm (Spirtes et al., 2000), Zhang et al. (2017) propose additional orientation rules for determining the causal direction. These additional rules are realized due to the changing causal modules for nonstationary or heterogeneous data.

Case I: Unshielded Triples The first case considers unshielded triples including the domain or time index variable. Let $C - V_k - V_l$ such a triple. Since the changing modules are function of the index, we can assign $C \to V_k$. Now, the conditional independence test $C \perp\!\!\!\perp V_l \mid \mathbf{S}$ gives the orientation of $V_k - V_l$. If $V_k \in \mathbf{S}$ then the triples must be a chain $C \to V_k \to V_l$. Otherwise, the triples will be a collider $C \to V_k \leftarrow V_l$.

Case II: Shielded Triples The second case of orienting a shielded triple $C - V_k - V_l$ and $C - V_l$ is a bit complicated. The edges can be paritally oriented as $C \to V_k - V_l$ and $C \to V_l$. However, $V_k - V_l$ can be either confounded by $\mathbf{g}$ or $\{\theta_k(C), \theta_l(C)\}$ can independently affect V_k and V_l respectively. Zhang et al. (2017) propose an approach based on principle of independent changes that states $P(cause)$ and $P(effect|cause)$ are independent given cause and effect change independently and confounder(function of C) is not present. Let $V_k = V_1$ and $V_l = V_2$ and without loss of generality $V_1 \to V_2$ be the true causal direction. Figure 4.2 (a) depicts the case of independent changes and (b) shows the case where a confounder is present. Zhang et al. (2017) show the case for Figure 4.2(a) can be solved using the property that independent changes $\theta_i(C)$ are independent only in causal direction. This can be estimated from data by calculating Kullback-Leibler divergence score

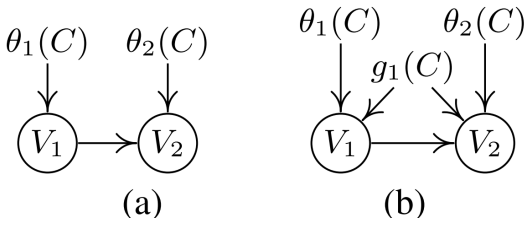


Figure 4.2: Two possible situations when $V1 \rightarrow V2$ and both $V1$ and $V2$ are found to be changing modules (adjacent to C)(a) Independent changing parameters (b) A confounder in addition to changing parameters

(Δ, given by equation 4.3) for both causal orientations and choosing the orientation that provides minimum score.

$$\Delta_{V_2 \rightarrow V_1} = < log \frac{\bar{P}(V_1 \mid V_2)}{< \hat{P}(V_1 \mid V_2) >} > \tag{4.3}$$

, where <.> denotes sample average, $\bar{P}(V_1 \mid V_2)$ is calculated from whole data, and $< \hat{P}(V_1 \mid V_2) >$ is estimated for each domain or time window.

For the case with confounder as shown in Figure 4.2(b), Zhang et al. (2017) conjecture the approach still work assuming the influence of the confounder $g_1(C)$ is not very strong. Zhang et al. (2017) reason the Δ captures influence of confounder for true causal direction whereas the wrong causal direction has additional disturbances along with confounder influence.

4.3 Critique

4.3.1 Causal Structure Learning Approach

CD-NOD algorithm (Zhang et al., 2017) determines skeleton and edges orientation for non-identical distribution changing with domain or time. The SEM representation is extended to include time/domain dependent unobserved influences as well as confounders that affect random variables. The CD-NOD approach uses both causal graphical model and structural equation model for the representation of causal assumptions and non-identical data generating process. The algorithm assumes faithfulness of the distribution but relaxes the causal sufficiency assumption such that there can be hidden common causes as a function of time or domain index. This modification is termed as pseudo causal sufficiency assumption. CD-NOD adds an orientation determination approach similar to the functional causal model in addition to regular constraint-based rules. The table 4.1 summarizes the causal structure learning approach.

4.3.2 Contributions and Limitations

The main contributions of Zhang et al. (2017) were to develop a causal representation that allowed random variables with non-identical distribution, estimate the changing modules along with a skeleton graph, and orient the edges using the heterogeneity or nonstationarity properties of data (Section 4.2.2). The experimental results on the simulated dataset showed that in comparison to the traditional SGS algorithm, the CD-NOD (enhanced) algorithm reduced the false positives significantly. This result follows from the example shown in Figure 4.1 where traditional algorithms are affected by time or domain confounding. Moreover, the experiments on real word brain imaging and breast tumor dataset showed the algorithm's applicability in the real-world for reduced false

Table 4.1: Synopsis of CD-NOD algorithm in the context of causal structure learning

Concept	Description
Input	Nonstationary/Heterogeneous Data with time/domain index
Representation Type	✔Graphical Models ✔SEM
Representation	Functional Causal Model and Augmented Graphs (With time/domain index)
Algorithm Type	✔Constraint-based ✖Score-based ✖Hybrid
Causal Assumption	✔Faithfulness ✔Sufficiency (Pseudo)
Contribution by non-IIDness	Changing modules detection and Kullback-Leibler divergence score based edges orientation
Output	Equivalence classes DAG with variables and domain/time index

positives. The integration of the principle of independent changes allowed edge orientation with the KL divergence score and added a new perspective to the constraint-based algorithm. Also, Zhang et al. (2017) explicitly specified the type of conditional independence tests to be performed for better results. Although the conditional independence test is the backbone for most of the constraint-based algorithms, the details are typically left obscure.

The experimental results of the CD-NOD algorithm were compared to the SGS algorithm which is a weak baseline. The PC algorithm and its order-independent variant (Colombo and Maathuis, 2014) with asymptotic completeness guarantee would be a much stronger baseline. Moreover, both PC and SGS algorithms have assumptions of causal sufficiency which is violated by the confounding effect of time or domain index. FCI, the constraint-based algorithm capable of recognizing confounders could be used as a baseline to see out of the box performance. The determination of causal orientation for the case of Figure 4.2(b) was not justified theoretically and empirically. The assumptions of independent data, confounders only being a deterministic function of time or domain index could be challenged. The authors highlighted further improvements such as theoretical justification for the case of Figure 4.2(b), handling case when causal direction reverse, temporal feedbacks and so on.

5. Causal Structure Learning from Time series Data

Time series data is one of the most prominent non-IID data occurring in the real world. Time series data presents different challenges in learning causal structures from data. Most of the causal structure learning approaches are based on directed acyclic graphs and assume there is no temporal feedback. Temporal feedback occurs when a random variable's value is dependent on its past values. Such models are also known as vector autoregressive models (VARs) and extensively studied in econometrics. Both RCD (Maier et al., 2013a) and CD-NOD (Zhang et al., 2017) algorithms discussed in the previous chapters explicitly assume the causal relationships are instantaneous or contemporaneous and there is no temporal feedback. In this chapter, I summarize the work on causal structure learning from multivariate time series data in the settings with unmeasured confounding (Malinsky and Spirtes, 2018). Malinsky and Spirtes (2018) propose constraint-based and hybrid algorithms to handle multivariate time series by relaxing the causal sufficiency assumption. Moreover, Malinsky and Spirtes (2018) allow contemporaneous causal relations in time series data. There is a philosophical debate regarding contemporaneous effect in time series data (Malinsky and Spirtes, 2018; Granger, 1988). The Granger's causality and so-called VARs models assume there is no true contemporaneous causal effect. The causal effects believed to be propagated from one variable in subsequent time step and the observed contemporaneous relations are accounted for by the unmeasured confounding. The Structural Vector Autoregressive (SVAR) model used by (Malinsky and Spirtes, 2018), on the other hand, allow contemporaneous causal relationships. These contemporaneous causal relationships are possible when the true causal frequency is undersampled or aggregated over a period during the measurement. In the following section, I discuss the representation of the SVAR model with contemporaneous causal relationships using so-called dynamic DAG as well as the representation of unmeasured confounders in time series data.

5.1 Representation

5.1.1 SVAR and Dynamic DAGs

Structural Vector Autoregressive (SVAR) models are typically used in time series analysis to represent the a variable's lag dependencies (with past values) as well as contemporaneous dependencies (with present values). A k-dimensional order-p SVAR process can be written as equation 5.1 (Malinsky and Spirtes, 2018).

$$X_{i,t} = f_i(\mathbf{X}_t^{-i}, \mathbf{X}_{t-1}, .., \mathbf{X}_{t-p}, \varepsilon_{i,t}) \tag{5.1}$$

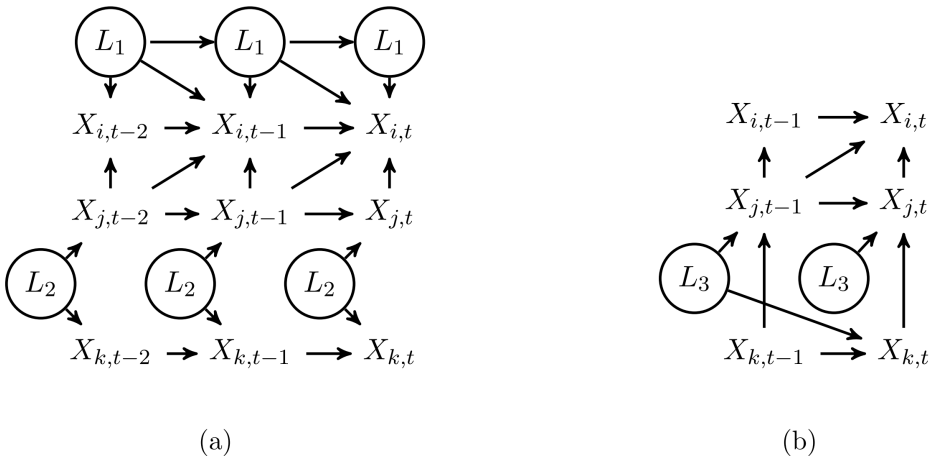


Figure 5.1: Dynamic DAGs with latent confounder interactions (a) L_1 is called an "auto-lag" confounder and L_2 may be called "contemporaneous confounder" (b) L_3 may be called "cross-lag confounder"

, where $X_{i,t}$ denotes a random variable $X_i \in \{X_1, X_2 ..., X_k\}$ at time t, $\mathbf{X}_t^{-i} = \mathbf{X_t} \setminus X_{i,t}$ signifies contemporaneous effects, $\mathbf{X}_{t-p}$ the lag variables of past p time step, and $\varepsilon_{i,t}$ is independent noise. The SVAR process is a special SEM. Thus, using the equivalence of SEM and Bayesian network, we can construct a joint probability distribution and corresponding DAG, known as dynamic DAG. The joint distribution is given by equation 5.2.

$$P(\mathbf{X_t}, .., \mathbf{X}_{t-p}) = \prod_{i \in \{1,...,k\}, s \in \{t,...,t-p\}} P(X_{i,s} | pa(X_{i,s}) \tag{5.2}$$

As seen from equation 5.1, self-feedback is not allowed on a random variable although contemporaneous effects are allowed.Malinsky and Spirtes (2018) also assume there is no selection bias and the data generating process is stationary with time. These assumptions allow the representation of causal structure in time series data with a Bayesian network known as a dynamic DAG.

5.1.2 Dynamic DAGs with Latent Variables

The equations 5.1 and 5.2 can be modified to include the unmeasured variables as well. Figure 5.1 depicts two dynamic DAGs showing the latent confounders and their types. Although the time series is (semi-)infinite, only a finite segment of the graph can be represented since the causal relations are repeating. The variables at $t - p$ have exact incoming edges as time t although it is not shown explicitly in the graph. There can be different types of confounders.

- **contemporaneous confounder**: Common causes of variables in the same time step.
- **cross-lag confounder**: Common causes of different variables across different time steps.
- **auto-lag confounder**: Common causes of a variable and its own past lags.

The latent confounders can not be measured from the data. Hence, we need a representation that marginalizes over the unobserved variables and at the same time maintain causal information. Maximal Ancestral Graph (MAG) is used to represent the marginalized distribution. MAG allows directed edges (→) as well as bidirected (↔) edges. $A \to B$ denote A is an ancestor of B. A bidirectional edge means both nodes are not the ancestors of one another and the dependency is due to a confounder. The conditional independencies entailed by an underlying DAG are preserved by

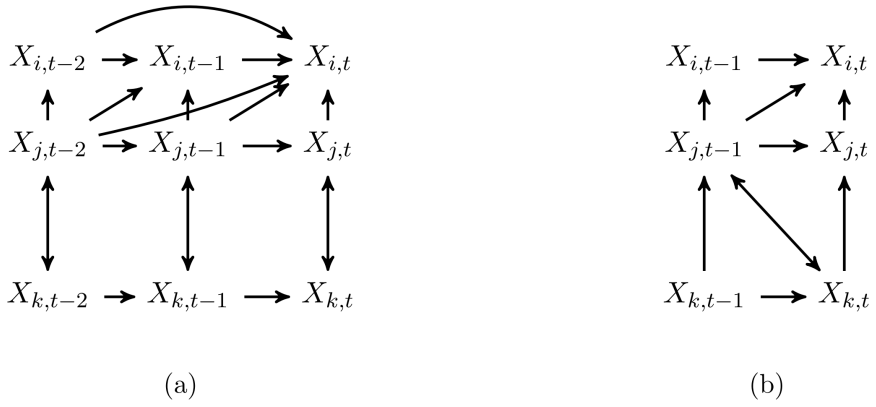


Figure 5.2: Dynamic MAGs for Marginalized dynamic DAG in Figure 5.1 (a) and (b)

MAG and m-separation, equivalent to d-separation, can be used to reason about independence. Two variables V_i and V_j are adjacent if the nodes are not d-separated in the underlying graph condition on the observed variables. The equivalence classes of a MAG graph is represented by so-called Partial Ancestral Graph (PAG) with a possibility of an edge with a circle in either one or both ends. $Ao \rightarrow B$ indicates A has an arrowhead at some MAG in equivalence class whereas a tail in others. Figure 5.2 depicts the marginalized dynamic DAG as a dynamic MAG.

5.2 Methodology

Malinsky and Spirtes (2018) modify the constraint-based FCI (Spirtes et al., 2000) and hybrid GFCI (Ogarrio et al., 2016) to develop SVAR-FCI and SVAR-GFCI respectively for handling multivariate time series data. These modified algorithms use temporal ordering and repeating structures in underlying DAG to estimate the skeleton as well as orient the edges in addition to prior approaches. The temporal ordering ensures the causal direction can never be toward past direction. The repeating structures are defined as **homologous pairs**.

A pair of vertices $(X_{i,s}, X_{j,t})$ is called a homologous pair with $(X_{m,a}, X_{n,b})$ if $i = m$, $j = n$ and $s - t = a - b$. Let, $hom(X_{i,s}, X_{j,t})$ denote all the homologous pairs to $(X_{i,s}, X_{j,t})$.

In figure 5.2 (a) $(X_{k,t-2}, X_{j,t-2})$ are homologous with $(X_{k,t-1}, X_{j,t-1})$ and $(X_{k,t}, X_{j,t})$. Similarly, $(X_{k,t-2}, X_{j,t-1})$ is homologous pair to $(X_{k,t-1}, X_{j,t})$. The adjacency and causal orientations in homologous pairs are the same. The followings sections outline the contribution by timeseries data and its representation for skeleton estimation and causal edge orientation.

5.2.1 Skeleton Estimation

The skeleton estimation starts with a complete PAG. The adjacency of edges are checked by following the similar procedure as PC and RCD algorithm. If a pair of nodes $(X_{i,s}, X_{j,t})$ are d-separated by some conditioning set, then the edge between $X_{i,s}$ and $X_{j,t}$ as well as all the edges between its homologous pairs $hom(X_{i,s}, X_{j,t})$ are removed.

5.2.2 Causal Orientation Determination

The causal edge is oriented to adjacent vertices from $X_{i,s}$ to $X_{j,t}$ if $s < t$ using the property of temporal order. After orientation obtained after the application of generic constraint based rules such as Collider Detection (CD), Known Non-collider (KNC), and Cycle Avoidance (CD) is copied to all the homologous pairs.

5.3 Critique

5.3.1 Causal Structure Learning Approach

Malinsky and Spirtes (2018) develop representation of time series data with contemporaneous causal effect using SVAR models. The equivalent dynamic DAG model allowed presence of latent confounders. Since Malinsky and Spirtes (2018) implement two algorithms for causal structure learning from time series data, the synopsis table 5.1 below includes the features of both algorithms. The main assumptions made by both algorithms are causal faithfulness, stationary distribution with time, and no selection bias.

Table 5.1: Synopsis of SVAR-FCI and SVAR-GFCI algorithms in the context of causal structure learning

Concept	Description
Input	Multivariate time series data (Stationary with no self-contemporaneous feedback)
Representation Type	✔Graphical Models ✔SEM
Representation	SVAR model and Maximal Ancestral Graphs (MAG)
Algorithm Type	✔Constraint-based(SVAR-FCI) ✖Score-based ✔Hybrid(SVAR-GFCI)
Causal Assumption	✔Faithfulness ✖Sufficiency
Contribution by non-IIDness	Temporal ordering for edge orientation, repeating homologous structures for skeleton estimation and edges orientation
Output	Equivalence classes of MAG called Partial Ancestral Graph (PAG)

5.3.2 Contributions and Limitations

The novel contribution by Malinsky and Spirtes (2018) is that the authors worked on contemporaneous causal effects in time series with latent variables settings for the first time. Malinsky and Spirtes (2018) developed an SVAR based dynamic DAG representation to encode the temporal ordering, causal dependency, and conditional independence relations. The representation allowed additional insights for skeleton estimation and edges orientation determination. The authors provided preliminary results that showed high precision but low recall for skeleton estimation and edges orientation in simulation experiments with a low sample size. Also, the experimental setup was limited to linear models with Gaussian noise. The assumption of stationary data allowed skeleton estimation and edge orientation using homologous pairs property. However, such an assumption may be a strong one in many domains. Although the model assumed contemporaneous causal effects, it ignored self-feedback. The authors pointed out it could be one of the future work.

6. Discussion and Further Research Direction

6.1 Discussion

Chapter 3 to 5 presented approaches of learning causal structure from relational, nonstationary/heterogeneous and time series dataset respectively. These data types are the most prominent non-IID data sources with either coupled dependencies between observations, or non-identical distribution of features or both. The common themes observed from the analysis of all three approaches are as follows.

1. Traditional algorithms fail or produce misleading results when directly applied to non-IID data.
2. More expressive representation is needed to capture the underlying mechanism in non-IID data.
3. Non-IID nature of data and its representation provides additional knowledge about causal dependency and orientation.

The theoretical guarantees and correctness in the causal structure learning literature come with many assumptions about the underlying data-generating mechanisms. The IID samples assumption made by Bayesian networks, the underlying representation, is the most challenging one. The application of approaches with IID assumption for non-IID data produce false skeleton or orientation that propagates over time in the most constraint-based algorithms. Maier et al. (2013a) and Zhang et al. (2017) showed traditional PC and SGS algorithms produced misleading results. Malinsky and Spirtes (2018), on the other hand, built on top of existing work on causal discovery from time series data to relax the causal sufficiency and no contemporaneous causal effects assumptions.

The representations in causal structure learning have to capture the underlying data generating mechanism as well as the conditional independencies found in data. Maier et al. (2013a) designed a relational model and abstract ground graph representation to capture the interactions in relational data. Zhang et al. (2017) allowed non-stationary/heterogeneous data generating processes with the modification of SEM and equivalent DAGs with a surrogate variable. Malinsky and Spirtes (2018) introduced graphical representation termed as dynamic MAG for so-called SVAR processes that handled unmeasured confounders.

Maier et al. (2013a) leveraged the auto-correlation between an entity attribute with itself through different relationship paths to develop relational bivariate edge orientation rule. Zhang et al. (2017) used heterogeneity and non-stationarity to allow modules changing with time/domain and detect these changing modules. Malinsky and Spirtes (2018) utilized the temporal ordering and repeating structures in time series data. The table 6.1 summarizes the main concepts in all three papers.

Table 6.1: Comparison of main concepts in Maier et al. (2013a), Zhang et al. (2017), and Malinsky and Spirtes (2018)

Concept	**Maier et al. (2013a)**	**Zhang et al. (2017)**	**Malinsky and Spirtes (2018)**
Data for causal structure learning	Relational	Nonstationary/ heterogeneous	Multivariate time series
Causal Discovery Approach	Constraint-based	Constraint-based (and hint of Functional causal model)	Constraint-based and Hybrid
Algorithms	RCD	CD-NOD	SVAR-FCI and SVAR-GFCI
Non-IIDness in data	Both not independent or non-identical but with out temporal or self feedback	Only non-identical distribution but independent	Not Independent but stationary with no contemporaneous self-feedback
Causal Representation	Relational Model and Abstract Ground Graphs (AGG)	SEM and Augmented Bayesian Network	SEM and Dynamic Maximal Ancestral Graphs (MAGs)
Causal Assumptions	Faithfulness and Causal Sufficiency	Faithfulness and Pseudo Causal Sufficiency	Faithfulness
Other Assumptions	No temporal feedback, Only contemporaneous causal relations	Independent distribution, Confounders smooth function of time or domain index, only contemporaneous causal relations	stationary distribution, no selection bias
Domain Knowledge or Parameter tuning	Relational Schema and hop-limit	Time / domain index	Time-order p
Conditional independence test used	Linear	Kernel-based non-parametric	Linear with Gaussain noise
Contribution by non-IIDness	Relational Bivariate Orientation	Changing modules detection and and Kullback-Leibler divergence score based edges orientation	Temporal ordering and repeating homologous structures
Output	Equivalence classes of AGG from each perspective	Equivalence classes of DAG with changing modules adjacent to time/domain index	Equivalence classes of MAGs (called PAG) showing marginalized causal dependencies

6.2 Further Research Direction

Although the sources reviewed (on Table 6.1) try to solve some aspects of non-IIDness of data separately, there is no generic representation (like Bayesian network for IID data) to capture all the non-IID nature of data. Real-world data is mostly relational, non-stationary and temporal at the same time. A long term research direction is to come up with relational models for non-stationary/heterogeneous time series data. This would make all other cases of non-IID data discussed above special cases.

Most of the causal structure learning algorithms come with explicit assumptions for correctness. The real-world complexities like unmeasured confounders, missing data, selection bias, etc make it difficult to fulfill all the assumptions. So, a research direction is making these models robust to these real-world data complexities.

Causal structure learning algorithms reason about conditional independencies in the data to come up with causal relationships. However, these models (mostly constraint-based) return only the equivalence classes. Thus, causal structure learning naturally fits with an interactive or active learning paradigm where domain knowledge and experimental results can be incorporated.

A challenge for causal discovery methods is the evaluation of the performances. Due to the lack of ground truth, most of the models are evaluated using synthetic data. However, the use of synthetic data is non-standard and it can possibly be influenced by the assumptions of the authors (Gentzel et al., 2019). A research direction can be to design a more robust evaluation framework of causal structure learning algorithms. The availability of empirical benchmark data in non-IID settings will help enhance the state-of-the-art in the field.

Conditional independence tests are the backbone of most causal structure learning algorithms. The mistakes in conditional independence tests are propagated to all other phases. However, most of the conditional independence tests come with their own assumptions of data distribution and parameters. Although the causal models are mostly non-parametric, the evaluation of most causal models is performed with linear SEM model with Gaussian noise. Thus, a research direction is to find methods for fast non-parametric conditional independence tests.

In conclusion, some theoretical foundations have been set up for representation of different non-IID data separately. With some assumptions and restrictions, partial solutions have been developed to reason about causal dependency in realtional, nonstationary/heterogeneous and time series data. Although there are challanges to handle the non-IID nature of data, these models actually utilize the underlying data structure in non-IID data to provide addition insights. There are still open challanges which require non-trivial solutions in the field of causal structure learning. Causal structure learning benefits most when the domain knowledge and experimental results are utilized interactively.

Bibliography

Borboudakis, G. and Tsamardinos, I. (2016). Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1435–1444.

Cao, L. (2014). Non-iidness learning in behavioral and social data. *The Computer Journal*, 57(9):1358–1370.

Chickering, D. (1996). Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer.

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

Chickering, D., Geiger, D., and Heckerman, D. (1994). Learning bayesian networks is np-hard. Technical report, Citeseer.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782.

Gentzel, A., Garant, D., and Jensen, D. (2019). The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, pages 11717–11727.

Getoor, L., Friedman, N., Koller, D., Pfeffer, A., and Taskar, B. (2007). Probabilistic relational models. *Introduction to statistical relational learning*, 8.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10.

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

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2018). A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*.

Lagani, V., Triantafillou, S., Ball, G., Tegner, J., and Tsamardinos, I. (2016). Probabilistic computational causal discovery for systems biology. In *Uncertainty in biology*, pages 33–73. Springer.

Maier, M., Marazopoulou, K., Arbour, D., and Jensen, D. (2013a). A sound and complete algorithm for learning causal models from relational data. In *In Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence.*, volume 2017, pages 371–380. AUAI Press.

Maier, M., Marazopoulou, K., and Jensen, D. (2013b). Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*.

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

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410.

Ogarrio, J. M., Spirtes, P., and Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. In *Conference on Probabilistic Graphical Models*, pages 368–379.

Oktay, H., Taylor, B. J., and Jensen, D. D. (2010). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9.

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

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662.

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

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

Zhang, K., Huang, B., Zhang, J., Glymour, C., and Schölkopf, B. (2017). Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813.