Estimating Causal Effects Considering Unmeasured Common Causes

by

José Sebastián Bejos Mendoza

Dissertation submitted in partial fulfilment of the requirements
for the degree of Master in Sciences

Instituto Nacional de Astrofísica, Óptica y Electrónica
Tonantzintla, Puebla, Mexico
August 2020

Supervisors:

Dr. Luis Enrique Sucar Succar
Dr. Eduardo Morales Manzanares
Department of Computer Sciences
INAOE

Abstract

Causal Bayesian networks (CBNs) are now widely used as causal models and they are the starting point for much of the research on automated causal discovery and causal inference in the artificial intelligence literature. The causal effect between a pair of variables (X, Y), that belongs to a set of variables of a joint distribution, it is a measure of how much the variable Y is modified when manipulating the variable X. The causal effect can be estimated from the post-intervention distribution over a CBN when the causal structure that generated the data is known *a priori*. Nevertheless, given only observational data, constraint-based causal discovery methods are solely able to find a finite collection of possible causal structures, i.e., a Markov Equivalence Class (MEC), but they can not identify the causal structure that best represents the data inside the MEC.

In this dissertation, we propose the LV-IDA+ algorithm to estimate bounds on causal effects, between pairs of variables (X, Y) in a system, considering unmeasured common direct causes. This algorithm is based on the IDA ([MKB09]) and the LV-IDA ([MS17]) algorithms. As in IDA and LV-IDA, we consider the case where we only know the MEC of the causal structure and the system $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is jointly Gaussian, and as in LV-IDA, we use Maximal Ancestral Graphs (MAGs) and Partial Ancestral Graphs (PAGs), to represent the causal structure and the MEC of the system, respectively. The LV-IDA algorithm cannot always identify the causal effect for some pairs of variables and then returns missing values as output. This is due to the fact that in special instances there is no adjustment set for some pairs of variables, in some, and occasionally all MAGs in the PAG. Our main contribution proposes a way to approximate the causal effect when these undetermined cases arise on the LV-IDA algorithm.

The LV-IDA+ algorithm uses the covariate adjustment method over the canonical Directed Acyclic Graphs (DAGs) associated with the MAGs in the PAG to approximate the causal effects in these unresolved cases. To our knowledge, no other way has been proposed to give at least an approximate answer instead of an inconclusive one to calculate the causal effect for such cases, and the proposed approximation closes this gap. In the experimental evaluation that was carried out over synthetic canonical DAGs, higher accuracy is observed in the estimation of the bounds on causal effects using this approximation approach. With the bonus that instead of having missing values we have at least an approximation of the bounds on causal effects for the mentioned special cases.

Resumen

Las Redes Bayesianas Causales (CBN) ahora se utilizan ampliamente como modelos causales y son el punto de partida para gran parte de la investigación sobre el descubrimiento causal automatizado y la inferencia causal en la literatura sobre inteligencia artificial. El efecto causal entre un par de variables (X, Y), que pertenece a un conjunto de variables en una distribución conjunta, es una medida de cuánto se modifica la variable Y al manipular la variable X. El efecto causal se puede estimar a partir de la distribución posterior a la intervención sobre una CBN cuando se conoce *a priori* la estructura causal que generó los datos. Sin embargo, dados solo datos observacionales, los métodos de descubrimiento causal basados en restricciones solo pueden encontrar una colección finita de posibles estructuras causales, i.e., una Clase de Equivalencia de Markov (MEC), pero no pueden identificar la estructura causal que mejor representa los datos dentro de esta MEC.

En esta disertación, proponemos el algoritmo LV-IDA+ para estimar cotas de los efectos causales, entre todos los pares de variables (X, Y) en un sistema, considerando causas directas comunes no medidas. Este algoritmo se basa en los algoritmos IDA ([MKB09]) y LV-IDA ([MS17]). Como en IDA y LV-IDA, consideramos el caso en el que conocemos únicamente la MEC de la estructura causal y el sistema $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ es conjuntamente Gaussiano y como en LV-IDA, usamos Grafos Ancestrales Maximales (MAGs) y Grafos Ancestrales Parciales (PAGs), para representar la estructura causal y el MEC del sistema, respectivamente. El algoritmo LV-IDA no siempre puede identificar el efecto causal para algunos pares de variables y devuelve valores faltantes como salida. Esto se debe al hecho de que, en casos especiales, no hay un conjunto de ajustes para algunos pares de variables, en algunos y ocasionalmente todos los MAG en el PAG. Nuestra principal contribución propone una forma de aproximar el efecto causal cuando estos casos indeterminados surgen en el algoritmo LV-IDA.

El algoritmo LV-IDA+ utiliza el método de ajuste por covariantes sobre los Grafos Dirigidos Acíclicos (DAG) canónicos asociados con los MAG en el PAG para aproximar los efectos causales en estos casos no resueltos. Hasta donde sabemos, no se ha propuesto ninguna otra forma de dar al menos una respuesta aproximada en lugar de una no concluyente para calcular el efecto causal para tales casos y la aproximación propuesta

cierra esta brecha. En la evaluación experimental que se llevó a cabo sobre DAG canónicos sintéticos, se observa una mayor precisión en la estimación de las cotas de los efectos causales utilizando este enfoque de aproximación. Con la ventaja extra de que en lugar de tener valores faltantes tenemos al menos una aproximación de las cotas de los efectos causales para los casos especiales mencionados.

This work is dedicated to my family and all my teachers.

Acknowledgments

It is impossible not to understate my gratitude to Dr. Enrique Sucar and Dr. Eduardo Morales, who advised this research and supported me during the whole process, who welcomed me with open arms into this now beloved institute and from whom I always received very cordial treatment. I particularly thank them for their trust and patience throughout this extreme adventure together in the area of causal inference.

I would also like to thank all the professors with whom I interacted during my graduate studies, especially Dr. Francisco Martínez Trinidad and Dr. Ariel Carrasco, with whom I wrote my first research paper in the area of machine learning and from whom I learned many things in this area. I would also like to make a special mention to Dr. Manuel Montes with whom it was a privilege to share ideas for several semesters and who introduced me to the area of natural language processing. I could not fail to thank Dr. Angélica Muñoz also from INAOE and Professor Manuel Valadez from UNAM, whom I considered great friends and whose lives were not long enough to reach the time of the conclusion of this work, but who were very important people in the path that led me to do these graduate studies.

Infinite gratitude to my family for all their understanding and patience towards my madness and love of mathematics. I apologize for all my absences and for having chosen a path so different from that of everyone in our beautiful family. Thanks to my parents, Rafael and Mónica, for their very special love. Thank you to my grandmother and grandfather, whom I adore with all my heart. I thank my brothers and sisters, Eduardo, Jose Maria, Blake, Benjamin, David, Erick, Daniela, and María, who were close to me during this adventure in Puebla and who made my life very special. To my great friend, who used to be my student, Ivanovich, for teaching me so much. To Zenchi, my great companion, for his unconditional love and company. Finally, to Chigüina Natalia, who saved me from sadness and gave me the opportunity to walk by her side even during this challenging time for me.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The notion of causality has been studied extensively, in science and philosophy, for many centuries. However, in this research, we focus solely on the probabilistic context of causal inference, where data is used to inform decisions about actions. In particular, the problem of causal inference in which we are interested is the one that concerns about the analysis of observational data on the variables of a system, to infer and quantify the causal relationships between the variables in the causal system. Though there are different formalism for study this statistical causal inference problem, a framework based on the so-called Causal Graphical Models is used here (see [Pea09; SGS00]). The formalism of Causal Graphical Models is based on Probabilistic Graphical Models (see [KF09; Suc15]) but extended for causal reasoning.

One of the main differences between a probabilistic model and a causal model is the special capacity of the latter to intervene in the system being modeled. Intervening can be understood as actively doing something to the system, instead of passively observing it. In this sense, a causal model is interested in measuring how much the manipulation of a certain set of variables $\boldsymbol{X}$ affects another set of variables $\boldsymbol{Y}$ in the system. As well as representing these types of relationships in a compact way when they are present in the system of study. Under the framework of Causal Graphical Models, the measure of the effect on a set of variable on another after intervening other in a system is called the total causal effect, and the representation of this type of relationship is employing different types of graphs: Directed Acyclical Graph (DAGs), Maximal Ancestral Graphs (MAGs), for example, as will be shown later.

The causal effect between a pair of variables (X, Y) it is a measure of how much the variable Y is modified when manipulating the variable X. The prevailing mean for estimating such effects is using randomized controlled experiments, in which the treatment variable (the cause) is randomized while the outcome variable (the effect) is passively observed. Fisher defines randomized controlled experiments as the procedure to physically

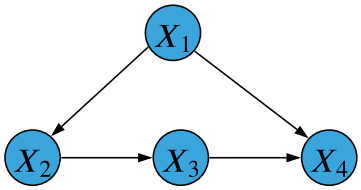
manipulate reality such that an outcome variable can be evaluated under different conditions ([Fis74]). This method is indeed one of the most pervasive techniques used throughout the sciences, and it is often deemed the gold standard for causal inference. For instance, the process of drug's approval is conducted following Fisher's method, for example, to estimate the effect of a drug (X) on a person (Y).

Nevertheless, experimental data are not always available given that randomized controlled experiments can be unethical, infeasible, time consuming, or expensive. On the other hand, observational data, i.e., data associated with processes that cannot be reproduced and are therefore not appropriate for conducting controlled experiments, are often abundant. In this dissertation, we consider the problem of estimating the causal effects between pairs of variables given only observational data and contribute to the state of the art methods to solve this problem. Causal effect estimation is notoriously hard to calculate from this kind of data, but recently there has been lot of interest in this problem (see for instance [MOS18; Per+18; ZLT19; MS17; MKB09; Maa+10] and [JZB19]). Most works consider the assumption of causal sufficiency. This assumption stipulates that no variables which are common direct causes of at least two measured variables are unmeasured. In this research we relax this assumption, which is often not realistic in applied contexts. Since it assumes that all possible causes interacting in a causal system are known and measurable.

1.1 Causal Graphical Models

In general, the Causal Graphical Models (CGMs) are described as a pair $(\mathcal{G}, \Phi_{\mathcal{G}})$ where $\mathcal{G}$ is a directed graph where the set of vertices represents a set $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ of jointly distributed random variables, named the causal structure of the CGM (see Figure 1.1 (b)) and $\Phi_{\mathcal{G}}$ is the set of parameters of the model. The set of parameters is given as a set of functional relationships among the variables which are known as Structural Equations (SE) (see Figure 1.1 (a)). That is, in the parameters set $\Phi_{\mathcal{G}}$ a function $X_i = f_i(\boldsymbol{pa}(X_i), U_i)$ is assigned to each $X_i \in \boldsymbol{V}$ and a probability distribution $P(U_i)$ to each U_i, where $\boldsymbol{pa}(X_i)$ are the parents of X_i in $\mathcal{G}$ and where U_i is a random disturbance (the CGMs will be explained in more detail later in Chapter 2). In this research, we focus exclusively on a linear Gaussian parameterization of the causal structure $\mathcal{G}$, that is, the set of parameters $\theta_{\mathcal{G}}$ are linear SE with Gaussian errors (see Chapter 2 Section 2.2 for details).

The most widely used graphs as causal structure are directed acyclic graphs (DAGs). Directed acyclic graphs are a great tool in statistical modelling in virtue of their probabilis-



(a)

$$X_1 = \epsilon_1$$
$$X_2 = 2X_1 + \epsilon_2$$
$$X_3 = 3X_2 + \epsilon_3$$
$$X_4 = X_1 + 2X_3 + \epsilon_4$$
$$\epsilon_1, \ldots, \epsilon_4 \sim \mathcal{N}(\mu, \sigma)$$

(b)

Figure 1.1 An example of a CGM. (a) The causal directed graph $\mathcal{G}$ of the CGM. (b) The set $\Phi_{\mathcal{G}}$ of parameters of the model given as a linear Gaussian SE.

tic semantics: a DAG over a set of random variables encodes a set of conditional independence constraints on the joint probability distribution by its local Markov property. The local Markov property specifies that every variable in the DAG is independent of its non-descendants conditional on its parents. These local conditional independence constraints imply others that are called global conditional independence, which can be read in the DAG by a graphical criterion known as *d*-separation, which will be discussed in the following chapters.

More importantly for our purpose, DAGs have a natural causal semantics: vertices represent variables, and arrows represent direct causal relationship between pairs of variables. By direct causal relationship we meant that there is no variable that mediates the relationship, or intuitively, that there exists a manipulated change in the variable cause that will be followed by a change in the variable effect, while holding all other variables fixed (see 1.2).

Even though most of the work in causal modeling with CGMs assumes causal sufficiency, i.e., it is assumed that no variables which are common direct causes of at least two measured variables are unmeasured. A major worry among statisticians towards inferring causation from correlation between random variables, is that the existence of unobserved or latent variables (variables of which we can not measure the values) that contribute to the observed correlation pattern among observed variables (variables of which we can and do measure the values) is very natural.

Maximum ancestral graphs (MAG) are an alternative representation for the causal structure of a CGM which allows modeling insufficient systems (see [Zha08a; RS02; Zha08b]). This ancestral graphs that generalize DAGs can represent conditional independence infor-

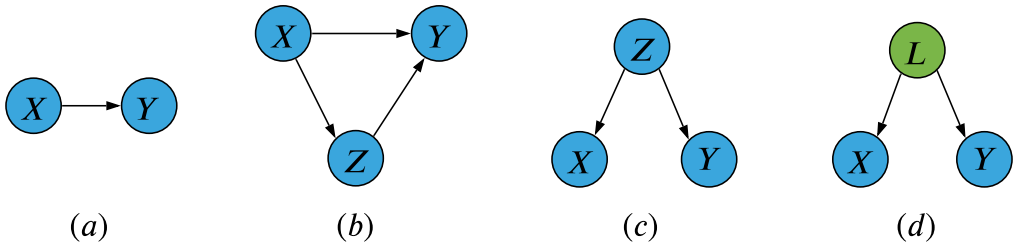


Figure 1.2 Different causal relationships within a DAG. (a) The simplest relationship between a pair of variables (X, Y), the direct causal relationship. (b) The variable Z is a mediator in the existing causality between X and Y. (c) The variable Z is a common cause between X and Y. (d) The variable L is an unmeasured (or latent) common cause between X and Y.

mation and causal relationships that include unmeasured variables (also called hidden or latent variables). Since in this work we are interested in the calculation of causal effects considering the possibility of non-measurable common causes, we work mostly with this type of causal structure for our modeling. We leave details of the semantics of ancestral graphs for later chapters (see Chapter 3).

1.2 Causal Effects Computations

In the context of Causal Graphical Models three different techniques have been used to calculate the causal effect between a pair of random variables (X, Y): Covariate Adjustment, Inverse Probability Weighting (IPW), and Instrumental Variables. In all these methods, the causal structure plays an important role, because it tells us which variables can be used for covariate adjustment, which variables can be used as instruments variables, and which weights should be used in inverse probability weighting.

In this dissertation, we deal with the causal effect computation between pairs of variables in a system using the general approach of the covariate adjustment method in the context of causal graphical models. This method can be divided in three parts: (i) First, a causal structure represented by a graph is estimated, where the vertices are the variables in the system and the edges indicate direct causal relationships between them. (ii) Then, the structure of the causal graph is used to find a set of variables, called the adjustment set, which is a sufficient set, along with the treatment variable, to compute the needed intervention. (iii) Lastly, using an adjustment set, the post-intervention distribution required to

estimate total causal effect can be calculated (see Chapter 2, Section 2.3 and Chapter 4 for details).

An adjustments set can be seen as the set of variables that influence the flow of causality over the causal structure among the variables of interest and these sets are regularly not unique for a pair of variables sets, but may be several. For the case of a DAG causal structure, the set $\boldsymbol{pa}(\boldsymbol{X})$ of the parent of the variable X always satisfies the backdoor criteria ([MC15]), i.e, is a adjustment set, and may be use for the causal effect computation by covariate adjustment. Besides, it always exists an adjustment set for DAG causal model. The search for adjustment sets when the causal structure is given by a MAG is more sophisticated. Moreover, it may not exist one for some pairs of variables (see [ZLT19; Per+18; MC15; JZB19]). This peculiarity about the insufficient causal systems represented by MAGs is a central motivation for the contribution to the state of the art in this research, so we will inquire into this in Chapters 4 and 5.

Our contribution deals with the second and third parts of the covariant adjustment method. Concerning the first part, we consider that the causal structure is estimated by a constraint-based algorithm. These algorithms consider conditional dependencies on the observational distribution, to infer the causal directed acyclic graph (DAG) that generated the data. Unfortunately, multiple DAGs can encode the same set of conditional independence relationships so these methods can exclusively find an equivalence class, called the Markov Equivalence Class (MEC), of the underlying causal structure given a set of observational data (see [GZS19]).

If the assumption of causal sufficiency is relaxed constraint-based algorithms, such as the Fast Causal Inference algorithm (FCI), will estimate a Partial Ancestral Graph (PAG), which encodes a MEC of MAGs (see [Zha08b; OSR16]).

Regarding the search for adjustment sets, in [MC15; Per+18; JZB19] and [ZLT19] are defined graphical criterion over DAGs and MAGs for finding sets of variables that can be used in order to estimate the causal effect using observational data by covariate adjustment. In particular, in [Per+18] and [ZLT19], sound and complete algorithms are shown for constructing sets that satisfy their criterion for MAGs. These results play a fundamental role for the causal effects estimations in the algorithm we propose.

Since by mean of a constraint-based algorithm it is not possible to specify a single causal structure $\mathcal{G}$ but a summary structure $\mathcal{P}$ that represents the MEC to which $\mathcal{G}$ belongs. It is possible to unfold $\mathcal{P}$ to list all the structures represented by $\mathcal{P}$ and calculate the causal

effect among a variables pair (X, Y) in each of the causal structures in $\mathcal{P}$, performing the second and third part of the covariante adjustment method.

Under this framework, we can bound the real causal effect between a pair of variables (X, Y) as the minimum and maximum of the effects calculated in each of the models in $\mathcal{P}$. This framework to bound causal effects in causal sufficient system, i.e., considering that $\mathcal{G}$ is given as a DAG and $\mathcal{P}$ as a Completed Partially Directed Acyclic Graph, was proposed in [MKB09] as the IDA ("Intervention when the DAG is Absent") algorithm. The IDA algorithm was extended for insufficient system in [MS17], where the LV-IDA (Latent Variables IDA) algorithm was proposed. This algorithm works with MAGs as the underlying causal structures $\mathcal{G}$ and PAGs as their summary structure $\mathcal{P}$, allowing the possibility of latent confounders (equivalent to unmeasured common causes) in the modeling.

1.3 Problem Statement

The fundamental question that we focus to answer is this dissertation is the following.

Is there any plausible approximation for the value of the causal effect between a pair of variables (X, Y) when it is not possible to find an adjustments set in an insufficient causal system $\boldsymbol{V} = \{X_1, \ldots X_p\}$ modeled with a Maximal Ancestral Graph?

1.3.1 Motivation

In the covariate adjustment method for causal graphical models, an adjustment set represents a subset of variables on which the original probability distribution (before the intervention) can be marginalized to identify the post-intervention probability distribution.

We refer to the original probability distribution also as the pre-intervention distribution. It is this distribution the one that represents the purely observational data-set. On the other hand, the post-intervention distribution is the one representing the data after performing an intervention on some of the variables in the system, as its name indicates.

Identifying the post-intervention probability distribution means finding an expression for its formulation in terms of conditional probabilities between the system variables. Thus, the post-intervention distribution can be computed through marginalization and conditioning operations over the pre-intervention distribution, i.e., using purely observational data. To refer to the post-intervention distribution for a pair of variables (X, Y) in a system, i.e., the marginal distribution of Y after the intervention of variable X, we use the notation $P(Y \mid do(X))$. With the identification of $P(Y \mid do(X))$ for the pair of variables (X, Y), it

is possible estimation of the causal effect of X on Y, i.e., to measure the variation of the variable Y under the manipulation on the variable X.

Causal systems are modeled using MAGs when the assumption of causal sufficiency is relaxed. On this type of models it is not always possible to find an adjustment set for some pairs of variables (X, Y). Therefore, it is not possible to identify the post-intervention distribution between some pairs of variables (X, Y). In principle, this does not allow the calculation of the causal effect between some pairs of variables (X, Y) in the system by mean of the covariate adjustment method.

1.3.2 Research Questions

The following questions guide this research work:

1. Given that the causal effect for any pair of variables (X, Y) is always identifiable on a causal DAG, and knowing that a causal MAG is very general model that represents an infinity of possible causal DAGs with different number of no measured variables: Is there a representative candidate DAG* on the set of DAGs encoded by a MAG, such that the estimation of the causal effect by covariate adjustment over the DAG* is a good approximation for the cases where the causal effect is not identifiable in the MAG?

2. Considering a representative directed acyclic graph DAG*, like the one mentioned in the previous question, this will be a DAG with latent variables for which we will not have data, nor will know to what extent this latent variables interact with the observed variables of the system. In this sense, the following question arises: How could a complete parameterization for this DAG* with latent variables be obtained?

3. As we can only get a MEC when trying to learn the causal MAG from observational data using a constraint-based algorithm, i.e., the PAG that represents the MEC of the MAG. Can we use this representative DAG* to compute at least bounds on the causal effects given only such PAG?

4. How can we evaluate the accuracy of such bounds on causal effects given solely a PAG that represents the MAG of the underlying causal MAG and given that for some MAGs in a PAG, part of the causal effects are not identifiable by covariate adjustment?

1.3.3 Justification

There is substantial research on finding adjustment sets in the causal inference community (see for example [MC15; Per+18; ZLT19; JZB19]). All these recent research works have focused on characterizing under which circumstances there are adjustment sets on MAGs and other graphs, establishing different graphical criteria. Furthermore, the most recent ones have shown that their criteria are sound and complete (see [Per+18; ZLT19; JZB19]).

On the other hand, recently Malinsky and Spirtes (see [MS17]) proposed a way to estimate upper and lower bounds on causal effect over pairs of variables in insufficient systems. Their LV-IDA algorithm receives as input, in addition to the observational data or the set of dependencies between them, a PAG that could have been learned by a constraint-based causal discovery method. Its method returns missing values for the causal effect of some pair of variables when it was not identifiable in some or all of the MAGs of the input PAG. This implies that the computed bounds of the causal effects by their algorithm are not trustworthy when these missing values values (NA values) are returned on some of the MAG in the PAG. Moreover, when it is the case for all MAGs in the PAG and not just some, their algorithm cannot return the bounds on causal effects.

As far as we know, no one has presented any related research work on how to give an approximation of the causal effect for cases where we cannot find adjustment sets for the pair of variables in question. All works are limited to answering that it is not possible to identify the causal effect directly for such cases. In this sense, it seems important to propose solutions to the question of how such a causal effect can be approximated when it cannot be calculated directly, i.e., when it is not possible to identify the post-intervention distribution $P(Y \mid do(X))$ for some pairs of variables (X, Y) over a MAG. In principle, this could improve the results in the work of Malinsky and Spirtes ([MS17]) because these approximations would help to get better bound on causal effects given a PAG.

1.3.4 Hypothesis

In consequence of the research questions we establish the general hypothesis of this research as: Estimating the causal effect for a pair (X, Y) by covariant adjustment on the canonical DAG $\mathcal{D}^*$ associated with a MAG may be a good approximation for cases where the causal effect over (X, Y) is not identifiable on a MAG. To obtain a complete parameterization of this canonical DAG $\mathcal{D}^*$ with latent variables, we can use expectation-maximization techniques.

1.3.5 Objectives

The following objectives are oriented to afford answers to the research questions presented before for confirming or refuting the hypothesis.

The general objective of this research is to develop and validate a way to approximate the causal effect between a pair of variables when it cannot be directly identified in a MAG to estimate bounds on causal effects in a PAG. We have the following specific objectives:

1. To implement a way to compute a canonical DAG from a MAG and obtain a complete parameterization of it using expectation-maximization techniques to be able to estimate causal effects between pairs of variables in this canonical DAG.

2. To extend the LV-IDA algorithm ([MS17]) to estimate bounds on causal effects given a PAG by calculating the causal effects over the canonical DAGs associated with the MAGs in the PAG, as approximations for cases when they can not be calculated directly on the MAGs.

3. To present a data generation process in which a random PAG is first constructed, then a MAG within the MEC represented by this PAG is random selected, and finally the canonical DAG associated with this MAG is computed to generate the data according to this DAG.

4. To evaluate our proposal to approximate the causal effects for cases when they can not be calculated directly on the MAGs by comparing the quality of the estimated bounds on causal effects with our extension to the LV-IDA algorithm that implements our hypothesis, with the original LV-IDA algorithm.

1.4 Research Contribution

In this dissertation, we propose an algorithm based on the IDA and LV-IDA algorithms to calculate the bounds of the causal effects between any pair of variables (X, Y) in a insufficient system, i.e., considering non-measured direct causes. As in IDA and LV-IDA, we consider the case that the system $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is jointly Gaussian, and as in LV-IDA, we use MAGs and PAGs, to represent the causal structure and the MEC, respectively, of the system.

The main difference of the LV-IDA algorithm with respect to the proposed algorithm LV-IDA+, is that LV-IDA+ always guarantees the calculation of the causal effect between any pair of variables in the system. Whereas LV-IDA cannot always calculate this effect for some pairs of variables and then occasionally returns missing values as output. This is due to the fact that in these cases there is no adjustment set for the pair of variables (X, Y) in some, and occasionally all, MAGs in the PAG. Our main contribution proposes a way to approximate the causal effect when these degenerate cases are presented.

The LV-IDA+ algorithm uses the adjustment sets of the canonical Directed Acyclic Graphs (DAGs) associated with the MAGs in the PAG to approximate the causal effects in these unresolved cases. To our knowledge, no other way has been proposed to give at least an approximate answer instead of an inconclusive one to calculate the causal effect for such cases, and the proposed approximation closes this gap.

In the experimental evaluation that was carried out over synthetic canonical DAGs, a higher accuracy is observed in the estimation of the bounds on causal effects using this approximation approach. With the bonus that instead of having missing values for the mentioned especial cases we have at least an approximation of the bounds on causal effects for every pair of variables in the system.

1.5 Structure of the Dissertation

This dissertation has been organized in three parts and seven chapters. In the first part, the theoretical bases on which this work is based are set out. This part, which we called Background, is composed of Chapters 2 and 3. In chapter two, the foundations on the Causal Bayesian Networks are exposed. Much of what is presented is for the particular case when these have a linear Gaussian parameterization. In Chapter 3 the necessary definitions for the Ancestral Graphs Markov Models are presented, which are a type of graphs specialized for modeling causal structures over insufficient causal systems. In Related Work, the second part of this document, which consists only of Chapter 4, the ideas of the IDA framework and its extension to insufficient systems, the LV-IDA algorithm, are presented. In the third and last part of this document, which we named Contribution to the Field, we present in Chapter 5 the LV-IDA+ algorithm, in Chapter 6 the experimental evaluation of this algorithm, and in Chapter 7 we show the conclusions of the work.

Part I

Background

Chapter 2

Causal Bayesian Networks

In this chapter we introduce the fundamental concepts about Causal Bayesian Networks (CBNs) when the causal structure is given by a directed acyclic graph. In particular, we developed the theory regarding the case where the CBNs are parameterized with conditional linear Gaussian distributions, and show how interventions are calculated for this kind of CBNs. At the end of the chapter we talk about the use of Expectation Maximization techniques for learning the parameters of these models when hidden variables are present.

2.1 Causal Bayesian Nets Basics

Throughout this dissertation we denote sets in bold (for example $\boldsymbol{X}$), graphs in calligraphic font (for example $\mathcal{G}$) and vertices in a graph or variables in uppercase letters (for example X).

2.1.1 Directed Acyclic Graphs

. A **graph** $\mathcal{G}$ is an ordered pair of two sets $(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of objects called vertices and $\boldsymbol{E}$ a set of two elements subsets $\{V_i, V_j\}$ of $\boldsymbol{V}$, called edges. If the graph is directed, i.e., a **directed graph**, then the set of edges $\boldsymbol{E}$ is a set of ordered pairs of distinct vertices (V_i, V_j). Two vertices are **adjacent** in a graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ if there is an edge associating them, i.e., for vertices $V_i, V_j \in \boldsymbol{V}$, $\{V_i, V_i\} \in \boldsymbol{E}$ denoted as $V_i - V_j$ for an undirected graph, and $(V_i, V_i) \in \boldsymbol{E}$ denoted as $V_i \to V_j$ for a directed graph. In this work we consider simple directed graphs, meaning that there is at most one edge between any pair of vertices, for which we refer as directed graphs for brevity. The undirected graph resulting from substituting the directed edges for undirected ones in a directed graph $\mathcal{D}$ is called the **skeleton** of $\mathcal{D}$. Within a directed graph $\mathcal{D}$, a **walk** is a sequence of vertices $(V_1, \ldots, V_k)$, with initial vertex V_1 and terminal vertex V_k, such that V_i and V_{i+1} are adjacent for all i with $i = 1, \ldots, k-1$. A walk in which no vertex is repeated is a (directed) **path**. A walk in which only the initial and terminal vertices are the same is a (directed) **cycle**.

Definition 2.1.1. A directed graph $\mathcal{D}$ in which there are no directed cycles is called a **directed acyclic graph (DAG)**.

Definition 2.1.2. In a DAG $\mathcal{D}$ with $X, Y \in \boldsymbol{V}$,

$$\text{If } \left\{ \begin{array}{l} X \to Y \\ X \leftarrow Y \end{array} \right\} \text{then } X \text{ is a } \left\{ \begin{array}{l} \textbf{parent} \\ \textbf{child} \end{array} \right\} \text{of } Y \text{ and } \left\{ \begin{array}{l} X \in \boldsymbol{pa}(Y) \\ X \in \boldsymbol{ch}(Y) \end{array} \right\}$$

Definition 2.1.3. In a DAG $\mathcal{D}$ with $X, Y \in \boldsymbol{V}$, a vertex X is said to be an **ancestor** of a vertex Y, denoted as $X \in \boldsymbol{an}(Y)$, if there is a directed path $X \to \cdots \to Y$ from X to Y. Symmetrically, a vertex X is said to be a **descendant** of a vertex Y, denoted as $X \in \boldsymbol{de}(Y)$, if there is a directed path $Y \to \cdots \to X$ from Y to X.

A vertex X on a path in a DAG is said to be a **collider** if two directed edges meet at X (i.e., $\to X \leftarrow$). On the other hand, a vertex X is said to be a **common cause** if two directed edges diverge at X (i.e., $\leftarrow X \to$) on a path in a DAG.

2.1.2 Bayesian Networks

. A **probabilistic graphical model (PGM)** is a compact representation of a joint probability distribution, from which we can obtain marginal and conditional probabilities. There are two major tasks when using and exploiting PGMs: learning and inference. Learning

involves estimating the structure $\mathcal{G}$ and parameters ϕ of the model given an observational dataset over a set of variables $\boldsymbol{V}$. Inference deals with answering probabilistic queries by obtaining the conditional or marginal probability distribution of a subset of variables in $\boldsymbol{V}$. Bayesian Networks are a especial type of PGMs in which their structure is formed by DAGs (see [Suc15]), formally:

Definition 2.1.4. A **PGM** over a set of variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is a pair $(\mathcal{G}, \phi)$, where $\mathcal{G}$ is a graph that represents the structure of the model, and $\phi = \{f(\boldsymbol{y})\}$, with $\boldsymbol{Y} \subset \boldsymbol{V}$, is a set of local functions that defines the parameters of the model, such that, the joint probability is obtained by the product of the local functions:

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} f(\boldsymbol{y})$$

Definition 2.1.5. A **Bayesian Network (BN)** over a set of variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is a PGM $(\mathcal{G}, \phi)$ where $\mathcal{G}$ is a DAG $\mathcal{D}$ and ϕ is given as a set of **conditional probability distributions (CPDs)** of the form $P(x_i \mid \boldsymbol{pa}(x_i))$, i.e., the joint distribution for $\boldsymbol{V}$ can be expressed as the product

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(x_i \mid \boldsymbol{pa}(x_i))$$

.

In this case, we say that the distribution $P(X_1, \ldots, X_p)$ is factored according to the DAG $\mathcal{D}$. So the structure for a Bayesian Network is a DAG whose vertices represents random variables $\{X_1, \ldots, X_p\}$ and encodes the set of conditional independencies $(X_i \perp \boldsymbol{nde}(X_i) \mid \boldsymbol{pa}(X_i))$, for all X_i with $i = 1 \ldots p$, where $\boldsymbol{nde}(X_i)$ denote the set of non descendants of X_i. The expression $(X_i \perp \boldsymbol{nde}(X_i) \mid \boldsymbol{pa}(X_i))$ is read as: X_i is conditional independent of the variables in $\boldsymbol{nde}(X_i)$ given the set of variables in $\boldsymbol{pa}(X_i)$. The conditional independencies in this set are called **local independencies** and denoted by $\boldsymbol{I}_l(\mathcal{D})$ (see [KF09]). All the conditional independencies that hold for a structure given as a DAG, including the local independencies, can be described in term of the d-separation graphical criteria, defined as follows.

Definition 2.1.6. Let X, Y be two vertices and $\boldsymbol{W}$ a set of vertices with $X, Y \notin \boldsymbol{W}$ in a DAG $\mathcal{D}$. The vertices X and Y are d-**separated** given $\boldsymbol{W}$ in $\mathcal{D}$ if and only if, there exists no undirected path π between X and Y in the skeleton of $\mathcal{D}$, such that:

i Every collider on π has a descendent in W and,

ii no other vertex on π is in $\boldsymbol{W}$.

Definition 2.1.7. If $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{W}$ are three disjoint set of vertices in a DAG $\mathcal{D}$ with $\boldsymbol{X}, \boldsymbol{Y} \neq \emptyset$ then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated given $\boldsymbol{W}$ if an only if every pair $(X, Y) \in \boldsymbol{X} \times \boldsymbol{Y}$ is d-separated given $\boldsymbol{W}$ (see [SGS00]).

We use $\boldsymbol{I}_d(\mathcal{D})$ to denote the set of independencies that correspond to d-separation.

2.1.3 Markov Equivalence Classes

.

Definition 2.1.8. Let P be a joint distribution over a set of random variables $\{X_1, \ldots, X_p\}$ and let $\boldsymbol{I}(P)$ be the set of conditional dependencies of the form $(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z})$ that hold in P. We say that a structure $\mathcal{G}$ is an **independence map** (**$\boldsymbol{I}$-map**) for a set of independencies $\boldsymbol{I}(P)$ if $\boldsymbol{I}(\mathcal{G}) \subseteq \boldsymbol{I}(P)$.

Definition 2.1.9. A structure $\mathcal{G}$ is a **minimal $\boldsymbol{I}$-map** for a set of independencies $\boldsymbol{I}(P)$ if it is an $\boldsymbol{I}$-map, and the removal of even a single edge from $\mathcal{G}$ made it not an $\boldsymbol{I}$-map, and it is a **perfect $\boldsymbol{I}$-map** if we have that $\boldsymbol{I}(\mathcal{G}) = \boldsymbol{I}(P)$.

Several DAGs can encode the same conditional independencies via d-separation. Such DAGs form a **Markov Equivalence Class (MEC)** which can be described uniquely by a **Completed Partially Directed Acyclic Graph (CPDAG)**. A CPDAG C has the same adjacencies as any DAG in the MEC encoded by C. A directed edge $X \to Y$ in a CPDAG C corresponds to a directed edge $X \to Y$ in every DAG in the MEC described by C. For any non-directed edge $X - Y$ in a CPDAG C, the Markov equivalence class described by C contains a DAG with $X \to Y$ and a DAG with $X \leftarrow Y$. Thus, CPDAGs contain directed and non-directed edges (see Figure 2.1).

2.1.4 Causal Graphical Models

.

A **causal graphical model (CGM)** is a pair $(\mathcal{G}, \Phi_{\mathcal{G}})$ where $\mathcal{G}$ is a graph that is called the causal structure and $\Phi_{\mathcal{G}}$ is the set of parameters of the model. It is common to represent the causal structure of the system, i.e. the set of random variables $\{X_1, \ldots, X_p\}$, by a DAG $\mathcal{D}$

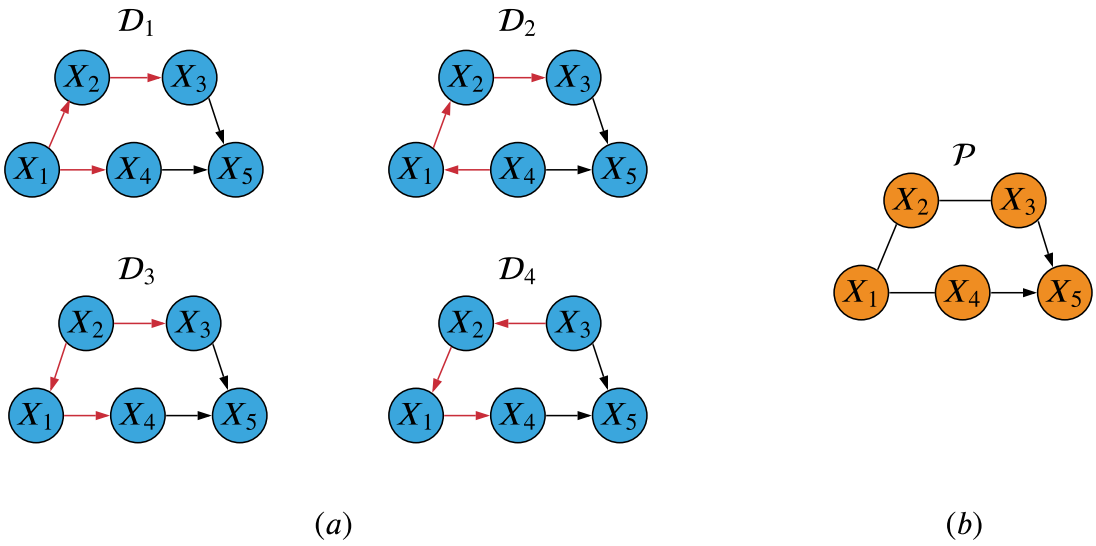


Figure 2.1 A set of DAGs that forms a MEC and its correponding CPDAG. (a) The set of four DAG in the MEG. (b) The CPDAG that represents this MEC of DAGs

where the vertices in $\boldsymbol{V}$ corresponds to the set of variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ and each edge in $\boldsymbol{E}$ represents a direct functional relationship among the corresponding variables, which is expressed by saying that X_i is the direct cause of X_j for an edge $(X_i, X_j) \in \boldsymbol{E}$. The set of parameter $\boldsymbol{\Phi}$ is given as set of functional relationship among the variables which are known as Structural Equations (SE). This set of SEs $\boldsymbol{\Phi}_{\mathcal{G}}$ assigns a function $X_i = f_i(\boldsymbol{pa}(X_i), u_i)$ to each $X_i \in \boldsymbol{V}$ and a probability distribution $P(u_i)$ to each u_i, where u_i is a random disturbance distributed according to $P(u_i)$, independently of all other u_j with $i \neq j$ (see [Pea09]). A special case, important for this work, is when the f_i functions are linear and the errors u_i are Gaussian distributed random variables.

Definition 2.1.10. A causal graphical model over a set of variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ can be represented as a **Causal Bayesian Network (CBN)**, i.e., as a pair $(\mathcal{G}, f)$, where the causal structure $\mathcal{G}$ is given as a DAG $\mathcal{D}$ and f is the joint distribution for $\boldsymbol{V}$ that factorizes as

$$f(\boldsymbol{V}) = \prod_{i=1}^{p} f(X_i \mid \boldsymbol{pa}(X_i)),$$

where these factors act as the set of parameters $\Phi_{\mathcal{G}}$ of the causal model.

Definition 2.1.11. A distribution f is consistent with a DAG $\mathcal{D}$ if the pair $(\mathcal{D}, f)$ forms a CBN.

CGM are capable of modeling changes in a system, expressing manipulations in the variables by external factors regardless of their previous probability distributions. The most common operation on CGM to represent these manipulations is the **intervention** operation, which we will discuss in later sections. The causal structure of a CGM and a CBN can be different than a DAG as we explore in the next chapters. On the other hand, in this research we are interested in the case when the joint probability distribution f for the CBN is a joint Gaussian probability distribution over a set of continuous random variables $\{X_1, \ldots, X_p\}$. So we go deeper into this type of parameterization in the following section.

2.2 Gaussian Directed Networks

In many situations, some variables are best modeled as taking values in some continuous space. Fortunately, nothing in our formulation of a Causal Bayesian network requires that we restrict attention to discrete variables. Our only requirement is that the CPDs $P(X_i \mid \boldsymbol{pa}(X_i))$ represent a distribution on a continuum of values over $\{X_1, \ldots, X_p\}$. However, as we show next, we can provide implicit representations for this type of CPDs. We focus on multivariate Gaussian distributions, which make strong assumptions but are surprisingly good approximation for many real-world distributions (see [KF09]).

2.2.1 Jointly Gaussian distributions

The most common characterization of a multivariate Gaussian distribution over a set of p random continuous random variables $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ is by an p-dimensional **mean vector** $\boldsymbol{\mu}$, and a symmetric $p \times p$ **covariance matrix** $\Sigma = (\sigma_{ij})$. This parameterization of the multivariate Gaussian distribution is call the **covariance form**, in which the multivariate Gaussian density function is defined as:

$$P(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big]. \tag{2.1}$$

The expression $(2\pi)^{p/2}|\Sigma|^{1/2}$ acts as a normalization constant, ensuring that the density integrates to 1, where $|\Sigma|$ is the determinant of the covariance matrix Σ. The matrix Σ most be positive definite, i.e., for any $\boldsymbol{x} \in \mathbb{R}^p$ such that $\boldsymbol{x} \neq 0$, we have that $\boldsymbol{x}^T \Sigma \boldsymbol{x} > 0$. Positive definite matrices are non-singular, and hence have determinant different from zero.

2.2.2 Marginalization and Conditioning

The operation of **marginalization** is it is easy to perform in the covariance form. If we have a joint normal distribution over $\{\boldsymbol{X}, \boldsymbol{Y}\}$, where $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$. The mean and covariance matrix of the Gaussian joint distribution can decompose as:

$$P(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y\end{pmatrix}; \begin{pmatrix}\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY}\end{pmatrix}\right), \tag{2.2}$$

where $\boldsymbol{\mu}_X \in \mathbb{R}^p$ and $\boldsymbol{\mu}_Y \in \mathbb{R}^q$, $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{XY}$, $\boldsymbol{\Sigma}_{YX}$, and $\boldsymbol{\Sigma}_{YY}$, are matrices of sizes $n \times n$, $n \times m$, $m \times n$ and $m \times m$, respectably. Then the marginal distribution over $\boldsymbol{Y}$ is given as the normal distribution

$$P(\boldsymbol{Y}) = \mathcal{N}(\boldsymbol{\mu}_y; \boldsymbol{\Sigma}_{YY}). \tag{2.3}$$

The other main operation that we wish to perform in joint Gaussian distribution over the set of variables $\boldsymbol{X}$ is **conditioning** the distribution given some evidence, i.e. on the observation $\boldsymbol{Z} = \boldsymbol{z}$, with $\boldsymbol{Z} \subset \boldsymbol{X}$. For the case of a joint normal distribution over $\boldsymbol{X}$ the conditional $P(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z})$ is given as the normal distribution

$$\begin{aligned} P(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z}) =& \mathcal{N}(\boldsymbol{\mu}_{X|Z}; \boldsymbol{\Sigma}_{X|Z}), \text{ where} \\ \boldsymbol{\mu}_{X|Z} =& \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{ZZ}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_Z) \text{ and} \\ \boldsymbol{\Sigma}_{X|Z} =& \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\Sigma}_{ZX} \end{aligned} \tag{2.4}$$

An alternative parameterization is the so called **information form**, where a Gaussian distribution is defined in terms of its inverse covariance matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$, called **information matrix**. Sometimes is useful and less computationally expensive to handled conditional operations in the information form in which $\boldsymbol{\mu}_{X|Z}$ and $\boldsymbol{\Sigma}_{X|Z}$ are expressed as follows:

$$\begin{aligned} \boldsymbol{\mu}_{X|Z} =& \boldsymbol{\mu}_X - \boldsymbol{\Lambda}_{XX}^{-1}\boldsymbol{\Lambda}_{XZ}(\boldsymbol{z} - \boldsymbol{\mu}_Z) \\ \boldsymbol{\Sigma}_{X|Z} =& \boldsymbol{\Lambda}_{XX}^{-1} \end{aligned} \tag{2.5}$$

2.2.3 Independencies in Multivariate Gaussians

For multivariate Gaussians, independence is easy to determine directly from the parameters of the distribution. In concrete, if $\boldsymbol{X} = \{X_1, \ldots, X_p\}$ have a joint normal distribution $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$. Then X_i and X_j are independent if and only if $\sigma_{i,j} = 0$. On the other hand, the independence structure in the distribution is apparently not in the covariance matrix, but in the information matrix, as we state next.

Proposition 2.2.1. If $P(X_1, \ldots, X_p) \sim \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, and $\boldsymbol{\Lambda} = (\lambda_{ij}) = \boldsymbol{\Sigma}^{-1}$, the information matrix. Then $\lambda_{ij} = 0$ if and only if $P \models (X_i \perp X_j | \boldsymbol{X} - \{X_i, X_j\})$

2.2.4 Linear Gaussian CPDs

Definition 2.2.1. Let Y be a continuous variable with continuous parents $\boldsymbol{pa}(Y) = \{X_1, \ldots X_q\}$. We say that Y has a **linear Gaussian CPD** if there are parameters $\beta_0, \beta_1, \ldots, \beta_q$ and σ^2 such that

$$\begin{aligned} P(Y \mid x_1, \ldots, x_q) =& \mathcal{N}(\beta_0 + \beta_1 x_1 + \ldots, +\beta_q x_q; \sigma^2), \text{ equivalently} \\ P(Y \mid \boldsymbol{pa}(y)) =& \mathcal{N}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{pa}(y); \sigma^2), \text{ in vector notation} \end{aligned} \tag{2.6}$$

This formulation can be interpreted as Y being a function of the variables $X_1, \ldots X_q$ with the addition of a Gaussian noise ϵ with mean 0 and variance σ^2. This is,

$$Y = \beta_0 + \beta_1 x_1 + \ldots, +\beta_q x_q + \epsilon$$

,

where ϵ is a Gaussian random variable with mean 0 and variance ϵ^2, representing the noise in the system. Note also that this is identical to a parameterization with linear structural equation systems of a CGM. Using the next proposition and induction, it follows that a Gaussian directed network defines a joint Gaussian distribution.

Proposition 2.2.2. If Y has a linear Gaussian CPD $P(Y \mid \boldsymbol{pa}(y)) = \mathcal{N}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{pa}(y)); \sigma^2)$ with parents $\boldsymbol{pa}(Y) = \{X_1, \ldots, X_q\}$, and $X_1, \ldots, X_q$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$. Then the distribution of Y is a normal distribution $P(Y) = \mathcal{N}(\mu_Y; \sigma_Y^2)$ where:

$$\begin{aligned} \mu_Y =& \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu} \\ \sigma_Y^2 =& \sigma^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}, \end{aligned}$$

and the joint distribution over $\{\boldsymbol{X}, Y\}$ is a normal distribution where:

$$Cov(X_i, Y) = \sum_{j=1}^{q} \beta_j \sigma_{ij}$$

From proposition 2.2.2, follows that if a BN has linear Gaussian CPDs then it defines a joint distribution that is jointly Gaussian. The converse of proposition 2.2.2 also holds, i.e., the result of conditioning is a normal distribution, where there is a linear relation on the conditioning variables. Using this, we have the following result.

Theorem 2.2.1. Let $\boldsymbol{X} = \{X_1, \dots X_p\}$ be a set of variables and let P be a joint Gaussian distribution over $\boldsymbol{X}$. We can always construct a structure DAG $\mathcal{D}$ and a **Gaussian Bayesian Network (GBN)** over $\mathcal{D}$ such that:

i $\boldsymbol{pa}(X_i) \subseteq \{X_1, \dots X_{i-1}\}$;

ii The CPD of X_i is a linear Gaussian CPD of its parents.

iii $\mathcal{D}$ is an I-map for P.

As for the case of discrete networks, the minimal I-map is not unique: different choices of orderings over the variables will lead to different network structures. From 2.2.1 theorem we have stated an equivalence between joint Gaussian distributions and **Directed Gaussian Networks (DGN)** as the Gaussian Bayesian Networks are also called. Even though we already mention the parametrization for DGNs as a system of linear structural equation in this section, in the next section we complete the theory and we reason about interventions over DGNs for use it as causal Gaussian Bayesian networks.

2.3 Interventions on DGNs

Given a CBN, it is possible to derive **post-intervention densities**. In particular, we are interested in interventions that set $\boldsymbol{X}$ to $\boldsymbol{x}$ uniformly in the population, which are denoted using the **do-calculus** as $do(\boldsymbol{X} = \boldsymbol{x})$ or shorthand $do(\boldsymbol{x})$, with $\boldsymbol{X} \subset \boldsymbol{V}$ (see [Pea09]). For this kind of interventions the post-intervention densities are given by the so called **truncated factorization formula**:

$$f(\boldsymbol{v} \mid do(\boldsymbol{x})) = \begin{cases} \prod_{\{i \mid X_i \in \boldsymbol{V} \setminus \boldsymbol{X}\}} f(x_i \mid \boldsymbol{pa}(x_i)) & \text{if } \boldsymbol{v} \text{ is consistent with } \boldsymbol{x}, \\ 0 & \text{otherwise,} \end{cases} \tag{2.7}$$

where $\boldsymbol{v}$ consistent with $\boldsymbol{x}$ means that $\boldsymbol{v}$ and $\boldsymbol{x}$ assign the same values to the variables in $\boldsymbol{V} \cap \boldsymbol{X}$.

Theorem 2.3.1. Sea $\boldsymbol{pa}(X)$ the set of parents (or direct causes) of variable X and let $\boldsymbol{Y}$ be any set of variables disjoint to $\{X \cup \boldsymbol{pa}(X)\}$. The **causal effect** of the intervention $f(\boldsymbol{y} \mid do(x))$ is given by

$$f(\boldsymbol{y} \mid do(x)) = \int_{\boldsymbol{pa}(X)} f(\boldsymbol{y} \mid x, \boldsymbol{pa}(x)) f(\boldsymbol{pa}(x)), \tag{2.8}$$

where $f(\boldsymbol{y} \mid x, \boldsymbol{pa}(x))$ and $f(\boldsymbol{pa}(x))$ are **pre-intervention densities**.

The causal effect $f(\boldsymbol{y} \mid do(x))$ is said to be **identifiable** if it can be calculated using only pre-intervention distributions, as in Equation 2.8.

Theorem 2.3.2. Given a CBN, in which a subset $\boldsymbol{V}$ of variables are measured, the causal effect $P(y \mid do(x))$ is identifiable whenever $\{X \cup Y \cup \boldsymbol{pa}(x)\} \subseteq V$, that is whenever X, Y, and all parents of variables in X are measured. The expression for computing $f(\boldsymbol{y} \mid do(x))$ is then obtained by adjusting for $\boldsymbol{pa}(x)$, as in Equation 2.8.

Corollary 2.3.1. Given a CBN, in which all variables are measured, the causal effect $P(y \mid do(x))$ is identifiable for every two subsets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ and is obtained from the truncated factorization (see Equation 2.7).

The previous corollary guarantees that we can always calculate the causal effect between any pairs of variables when all the variables in the system are measured. The proofs of Theorems 2.3.1 and 2.3.2 can be understood by consulting Chapter 3 in [Pea09]. The following definition implies that it is possible adjusting for different sets of variables, as will be seen in later chapters, and not just for the set of parents, as in Theorem 2.3.2.

Definition 2.3.1. Let $\mathcal{D}$ be a DAG, and let $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ be pairwise disjoint subsets of the set of variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$, with $\boldsymbol{X} \neq \emptyset$ and $\boldsymbol{Y} \neq \emptyset$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the manipulated and outcome variables, respectively. The set of variables $\boldsymbol{Z}$ is an **adjustment set** relative to $(\boldsymbol{X}, \boldsymbol{Y})$ in $\mathcal{D}$ if for any density f consistent with D we have that:

$$f(\boldsymbol{y} \mid do(\boldsymbol{x})) = \begin{cases} f(\boldsymbol{y} \mid \boldsymbol{x}), & \text{if } \boldsymbol{Z} = \emptyset \\ \int_{\boldsymbol{Z}} f(\boldsymbol{y} \mid \boldsymbol{xz}) f(\boldsymbol{z}) d\boldsymbol{z}, & \text{otherwise.} \end{cases} \tag{2.9}$$

Observe that adjustment sets allow post-intervention densities involving the do-operator to be identified, i.e., expressed only as specific functions of conditional densities which can be estimated from observational data, so the search for adjustment sets is fundamental for the computation of causal effects. Note that Equation 2.8 is the particular case of Equation 2.9 when $\boldsymbol{Z}$ is given by the set $\boldsymbol{pa}(X)$. The use of Equation 2.9 for computing causal effects over a CBN is what is called the **covariate adjustment** method. For the particular case of computing the covariate adjustment of pairs of single variables X and Y over multivariate Gaussian densities, (case in which we are interested in this research) we can use the fact that this kind of densities are fully defined by expectations, and for expressing conditional independencies $P(Y \mid x, \boldsymbol{z}) = P(Y \mid \boldsymbol{z})$ we can use equivalently conditional expectations $E(Y \mid x, \boldsymbol{z}) = E(Y \mid \boldsymbol{z})$. Furthermore, since conditional expectations are linear in a multivariate Gaussian distribution, the substitution of probabilities by expectations, allows to use regression for estimate $E(Y \mid x, \boldsymbol{z})$ as $E(Y \mid x, \boldsymbol{z}) = \alpha + \beta x + \gamma^T \boldsymbol{z}$, for some $\alpha, \beta \in \mathbb{R}$ and $\gamma \in \mathbb{R}^{|\boldsymbol{z}|}$.

Definition 2.3.2. The **total causal effect** of X on Y for continuous random variables setting is defined as $\frac{\partial}{\partial x} E(Y \mid do(x))$ (see [MKB09]).

So the total causal effect of X on Y in this setting is β, that is, the coefficient of X in the regression of Y on X and the adjustment set $\boldsymbol{Z}$, since by the adjustment set definition, we have that

$$E(Y \mid do(x)) = \int_{\boldsymbol{z}} E(Y \mid x, \boldsymbol{z}) f(\boldsymbol{z}) d\boldsymbol{z} = \alpha + \beta x + \gamma^T E(\boldsymbol{Z}).$$

2.4 Learning Causal Bayesian Networks

Learning causal Bayesian network includes two aspects: learning the structure and learning the parameters. The input of the learning procedure is: (i) Some knowledge or constraints about the structure $\mathcal{D}$ or the parameters ϕ of the model. (ii) A set $\boldsymbol{D}$ of data instances $\{\xi[1], \ldots \xi[n]\}$, which are independent and identically distributed (IID) samples from a distribution P.

Since we can place various constraints on the structure or on the parameters of the model, the class of models that we are allowed to consider as possible outputs of our learning algorithm, i.e., the hypothesis space extension, variate according to these input constraints. The less prior knowledge we are given, the larger the hypothesis space, and

the more possibilities we need to consider when selecting a model. We may not know the structure, and we have to learn both parameters and structure from the data. Even worse, we may not even know the complete set of variables over which the distribution P is defined. In other words, we may only observe some subset of the variables in the domain and possibly be unaware of others. The main contribution in this research uses parameter learning in the presence of unobserved sets of variables.

A central concept in learning a model from data is the likelihood function that measures the probability of the data induced by different choices of models and parameters. The likelihood function is determined by the probabilistic model we are learning. Given a choice of parameters, the model defined the probability of each instance. In the case of fully observed data, we assumed that each instance in a training data set $\boldsymbol{D}$ is simply a random sample from the model.

2.4.1 Parameter Learning

When the structure is known, parameter learning consists in estimating the CPDs from data. If the data are complete, the learning problem reduces to a set of local learning problems, one for each variable. That is, to learn the parameters for a Bayesian network with structure $\mathcal{D}$ and parameters Φ given a data set $\boldsymbol{D}$ consisting of $\xi[1], \ldots, \xi[n]$, we use the conditional likelihood of the variables $\{X_1, \ldots, X_p\}$ given it parents in the structure.

Definition 2.4.1. Let $\Phi_{X_i|\boldsymbol{pa}(X_i)}$ denote the subset of parameters in Φ that determines $P(X_i \mid \boldsymbol{pa}(X_i))$ with structure $\mathcal{D}$ of the BN as a parametric model. The **conditional likelihood** for these CPDs is given as the so called **local likelihood** function for X_i

$$L_i(\Phi_{X_i|\boldsymbol{pa}(X_i)} : \mathcal{D}) = \prod_{j=1}^{n} P(x_i[j] \mid \boldsymbol{pa}(x_i[j]) : \Phi_{X_i|\boldsymbol{pa}(X_i)}),$$

Then the likelihood function for a BN as a parametric model is given as:

$$L(\Phi : \mathcal{D}) = \prod_{i=1}^{p} L_i(\Phi_{X_i|\boldsymbol{pa}(X_i)} : \mathcal{D})$$

This shows that the likelihood of a BN decomposes as a product of independent terms, one for each CPD in the network.

Proposition 2.4.1. Let $\boldsymbol{D}$ be a complete data set for $X_1, \ldots, X_p$, and let $\mathcal{D}$ be a network structure over these variables. Let $\hat{\Phi}_{X_i|\boldsymbol{pa}(X_i)}$ be the parameters that maximize $L_i(\Phi_{X_i|\boldsymbol{pa}(X_i)}$:

$\mathcal{D}$). Then, $\hat{\mathbf{\Phi}} = (\hat{\mathbf{\Phi}}_{X_1|\boldsymbol{pa}(X_1)}, \ldots, \hat{\mathbf{\Phi}}_{X_p|\boldsymbol{pa}(X_p)})$ maximizes $L(\mathbf{\Phi} : \mathcal{D})$.

In other words, we can maximize each local likelihood function independently of the rest of the network, and then combine the solutions to get a **Maximum Likelihood Estimation (MLE)** solution.

2.4.2 Learning with Hidden Variables

For the methodology proposed in this dissertation it is important the case where there is a set of variables in the model for which there is no data at all. This common situation is known as the problem of parameter estimation with **hidden nodes** of **latent variables**. For these hidden nodes, the approach to estimate their parameters is based on the **Expectation–Maximization (EM)** technique (see [Suc15]). The EM algorithm takes the perspective that, when learning with missing data, we are actually trying to solve two problems at once: learning the parameters, and hypothesizing values for the unobserved variables in each of the data cases. Each of these tasks is fairly easy when we have the solution to the other. It consists of two steps which are repeated iteratively: (i) In the **E step**, the missing data values are estimated based on the current parameters, using probabilistic inference. (ii) In the **M step**, we then treat the completed data as if it were observed and learn a new set of parameters.

Given a dataset for variables $\boldsymbol{V} = \{X_1, \ldots X_p\}$ and a structure $\mathcal{D}$ for $\boldsymbol{V} \cup \boldsymbol{L}$ where L is the set of hidden nodes $\boldsymbol{L} = \{L_1, \ldots L_l\}$, to estimate the CPDs of the model, the EM algorithm does the following:

1. Obtain the CPDs for all the not hidden variables based on an ML estimator.
2. Initialize the unknown parameters with random values.
3. Considering the actual parameters, estimate the values of the hidden nodes based on the known variables via probabilistic inference.
4. Use the estimated values for the hidden nodes to complete/update the dataset.
5. Re-estimate the parameters for the hidden nodes with the updated data.
6. Repeat 3–5 until converge, i.e., no significant changes are observed in the parameters.
7. The EM algorithm optimizes the unknown parameters and gives a local maximum, the final estimates depend on the initialization.

2.4.3 Causal Structure Discovery

The type of datasets used in the learning of the causal structure may be: (i) **Observational** data corresponding to measurements made under natural conditions of a causal system. (ii) **Experimental** data correspond to measurements made under different disturbances of the system caused by external interventions. Ideally, experimental data should be used in the structure learning of causal BNs. Nevertheless, experimental data are not always available or can be unethical, infeasible, time consuming, or expensive. On the other hand, observational data, i.e., data associated with processes that cannot be reproduced are often abundant.

There are two classes of algorithms to learn causal structures, known as causal discovery algorithms (see [GZS19]). The **functional based algorithms** estimates the dependencies and conditional independencies of each variable over independent noises, and uses these relations to establish the direction between pairs of variables in the causal structure. This procedure is based on the idea that the independence between the noise and cause variable Y, holds for only one direction, such that it implies the causal asymmetry between X and Y. It has been shown that without some assumptions this causal direction is not identifiable because for both directions one can find an independent noise term. Nevertheless, assuming linearity and non-Gaussian noises this methods have been shown to be able to produce unique causal directions and have received practical applications (see [Shi+06]).

The **constraint-based algorithms** consider conditional dependencies on the observational distribution, to infer the causal directed acyclic graph (DAG) that generated the data. These algorithms, relies on the following assumptions:

1. **Causal Markov Condition**. Each variable in the causal structure is independent of its non-descendants given its directed causes.

2. **Causal Faithfulness**. Each true conditional independence between variables is entailed by the causal structure.

3. **Causal sufficiency**. Every common cause of two or more variables in a set of measured variables $\boldsymbol{V} = \{X_1, \dots X_p\}$, also is a measured variable in V.

4. **Acyclic structure**. There are no reflexive causal relations for all variables.

The disadvantages of this approach to causal discovery is that the solution is usually non-unique, but multiple causal structures can encode the same set of conditional independence relationships. So these algorithms try to efficiently search for a representation of the MEC of the causal structure that most closely entails the set of conditional independence relations under the Causal Markov and Faithfulness Assumptions. Despite this, constraint-based algorithms are generally applicable, i.e., can be used without restricting the type of functional relationship that exists between the variables and for any type of distributions (see [GZS19]).

Concluding Remarks

In this chapter, the necessary foundations on Causal Bayesian Networks were established. The theory for calculating of post-intervention densities for Gaussian Directed Networks and the concept of adjustment sets are of great relevance for the estimation of causal effects, the central theme of this research. Although structure learning was discussed, greater emphasis was placed on parameters learning because the main contribution of this dissertation uses the technique of expectations maximization to learn the parameters of a structure with hidden variables.

In the next chapter we will describe Ancestral Graphs. These type of graphs are DAGs generalizations for the modeling of insufficient causal systems, i.e., systems where the existence of unmeasured common causes are considered, a central theme in this dissertation.

Chapter 3

Ancestral Graphs Markov Models

It is common to assume that the set of variables $\boldsymbol{V}$ is causally sufficient i.e., it is assumed that no variables which are common direct causes of at least two measured variables are unmeasured. Under this assumption the causal structure of the system $\boldsymbol{V}$ is represented by a DAG $\mathcal{D}$. However, the assumption of no latent confounding is seldom appropriate, and it is desirable and even necessary in many situations to relax it. Unfortunately, the problem of causal reasoning and discovery becomes much more difficult when we drop the assumption, due to the fact that the causal structure may not be properly representable by a DAG unless latent variables are explicitly invoked. Not only are DAG models with latent variables hard to handle statistically, they make an infinite search space unless we seriously constrain the number of latent variables or the topology of the unknown causal network. Maximum ancestral graphs (MAG) are an alternative representation for the causal structure of a Causal Bayesian Network which allows modeling insufficient systems. In this chapter we introduce this type of graph and explain its semantics. Next, we introduce the necessary concepts that allow us to use the results of the previous chapter on insufficient systems modeled with MAG. In addition, canonical DAGs generated from a MAG are defined, the Partical Ancestral Graphs (PAGs) that are used to represent MEC of MAG, and the causal discovery Fast Causal Inferences algorithm is presented.

3.1 Maximal Ancestral Graphs

A directed **mixed graph** is a graph where there is at most one edge between any two vertices and that may contain two kinds of edges: directed edges ($\rightarrow$) and bi-directed edges ($\leftrightarrow$). The two ends of an edge are called **marks**, and there are two kinds of marks: **arrowhead** ($>$) and **tail** ($-$). We say an edge is **into** or **out of** a vertex if the mark of the edge at the vertex is an arrowhead or tail, respectively. We use the following terminology to describe the relations between vertices on a directed mixed graph $\mathcal{G}$:

$$\text{If } \left\{ \begin{array}{l} X \leftrightarrow Y \\ X \rightarrow Y \\ X \leftarrow Y \end{array} \right\} \text{in } \mathcal{M} \text{ then } X \text{ is a } \left\{ \begin{array}{l} \textbf{spouse} \\ \textbf{parent} \\ \textbf{child} \end{array} \right\} \text{of } Y \text{ and } \left\{ \begin{array}{l} X \in \boldsymbol{sp}(Y) \\ X \in \boldsymbol{pa}(Y) \\ X \in \boldsymbol{ch}(Y) \end{array} \right\}$$

In addition, we say that a vertex X is an **ancestor** of a vertex Y, denoted as $X \in \boldsymbol{an}(Y)$, if either there is a directed path $X \rightarrow \cdots \rightarrow Y$ from X to Y, or $X = Y$. Conversely, we say that Y is a descendant of X if X is an ancestor of Y.

Definition 3.1.1. A mixed (directed) graph is an **ancestral graph** if there are no directed cycles, and whenever there is an edge $X \leftrightarrow Y$, then there is no directed path from X to Y, or from Y to X.

In an ancestral graph, a non-endpoint vertex X on a path is said to be a **collider** if two arrowheads meet at X, i.e., with adjacency of the form: $\rightarrow X \leftarrow$, $\leftrightarrow X \leftrightarrow$, $\leftrightarrow X \leftarrow$. $\rightarrow X \leftrightarrow$. All other non-endpoint vertices on a path are **noncolliders**, i.e., vertices with adjacency of the form: $\rightarrow X \rightarrow$, $\leftarrow X \leftarrow$, $\leftarrow X \rightarrow$, $\leftrightarrow X \rightarrow$, $\leftarrow X \leftrightarrow$. A path along which every non-endpoint is a collider is called a **collider path**.

Definition 3.1.2. In an ancestral graph, a path π between vertices X and Y is **active** or m**-connecting** relative to a (possibly empty) set of vertices $\boldsymbol{Z}$, with $X, Y \notin \boldsymbol{Z}$ if

(i) every noncollider on π is not a member of $\boldsymbol{Z}$;

(ii) every collider on π is an ancestor of some member of $\boldsymbol{Z}$.

(iii) otherwise, $\boldsymbol{Z}$ **blocks** π.

Example: For the ancestral graph $A \rightarrow B \leftrightarrow C \leftarrow D$. The path $\pi_1 = (A, B, C, D)$ is active relative to $\boldsymbol{Z} = \{B, C\}$. The path π_1 is not m-connecting relative to $\boldsymbol{Z} = \emptyset$, $\boldsymbol{Z} = \{B\}$ or $\boldsymbol{Z} = \{C\}$, i.e., $\boldsymbol{Z} = \emptyset$, $\boldsymbol{Z} = \{B\}$ and $\boldsymbol{Z} = \{C\}$ blocks π_1.

Definition 3.1.3. X and Y are said to be $\boldsymbol{m}$**-separated** by $\boldsymbol{Z}$ if there is no active path between X and Y relative to $\boldsymbol{Z}$, i.e., if $\boldsymbol{Z}$ blocks all paths between X and Y. Two disjoint sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are m-separated by $\boldsymbol{Z}$ if every variable in $\boldsymbol{X}$ is m-separated from every variable in $\boldsymbol{Y}$ by $\boldsymbol{Z}$.

Definition 3.1.4. An ancestral graph $\mathcal{G}$ is said to be **maximal** if, for every pair of nonadjacent vertices (X, Y), there exists a set $\boldsymbol{Z}$ $(X, Y \notin \boldsymbol{Z})$ such that X and Y are m-separated conditional on $\boldsymbol{Z}$.

Definition 3.1.5. An inducing path π relative to a set $\boldsymbol{L}$, between vertices X and Y in an ancestral graph $\mathcal{G}$, is a path on which every nonendpoint vertex not in $\boldsymbol{L}$ is both a collider on π and an ancestor of at least one of the endpoints, X and Y.

Any single-edge path is trivially an inducing path relative to any set of vertices. To simplify terminology, we will henceforth refer to inducing paths relative to the empty set simply as inducing paths.

Definition 3.1.6. A mixed graph is called a **maximal ancestral graph (MAG)** if

i the graph does not contain any directed or almost directed cycles (**ancestral**); and

ii there is no inducing path between any two non-adjacent vertices (**maximal**).

Maximal ancestral graphs (MAGs) are maximal in the sense that no additional edge may be added to the graph without changing the independence model.

3.2 MAGs as Causal Models

The system $\boldsymbol{V}$ is said to be causally sufficient if there are no variables in $\boldsymbol{V}$ which are common direct causes of at least two measured variables that are unmeasured. If some of the variables in the set $\boldsymbol{V}$ are unmeasured, $\boldsymbol{V}$ can be partitioned as $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{L}$, where $\boldsymbol{O}$ is the set of observed (measured) variables and $\boldsymbol{L}$ is the set of latent variables (unmeasured). MAGs can represent conditional independence information and causal relationships in DAGs that include unmeasured (hidden or latent) variables. A MAG represents a DAG after all latent variables have been marginalized out, and it preserves all entailed conditional independence relations among the measured variables (see [Zha08a]).

Proposition 3.2.1. Given any DAG $\mathcal{D}$ over $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{L}$ there is a MAG $\mathcal{M}$ over $\boldsymbol{O}$ alone, such that for any disjoint sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subseteq \boldsymbol{O}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ in $\mathcal{D}$ if and only if they are m-separated by $\boldsymbol{Z}$ in the MAG $\mathcal{M}$.

The following construction gives us such a MAG (see Figure 3.1a for an example of this construction):

i For each pair of variables $X, Y \in \boldsymbol{O}$, X and Y are adjacent in $\mathcal{M}$ if and only if there is an inducing path between them relative to $\boldsymbol{L}$ in $\mathcal{D}$.

ii For each pair of adjacent variables X, Y in $\mathcal{M}$:

(a) orient the edge as $X \to Y$ in $\mathcal{M}$ if X is an ancestor of Y in $\mathcal{D}$;

(b) orient it as $X \leftarrow Y$ in $\mathcal{M}$ if Y is an ancestor of X in $\mathcal{D}$;

(c) orient it as $X \leftrightarrow Y$ in $\mathcal{M}$, otherwise.

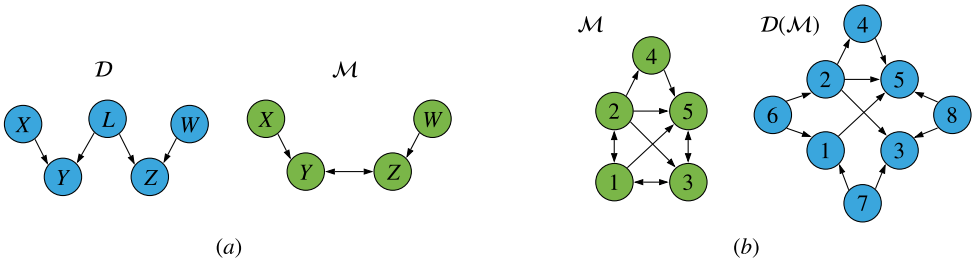


Figure 3.1 (a) The construction of a MAG $\mathcal{M}$ (green) over $\boldsymbol{O} = \{X, Y, Z, W\}$, from a DAG $\mathcal{D}$ (blue) over the set of variables $\boldsymbol{O} \cup \boldsymbol{L}$, with $\boldsymbol{L} = \{L\}$. (b) The canonical DAG $\mathcal{D}(\mathcal{M})$ (blue) associated with the MAG $\mathcal{M}$ (green), with $\boldsymbol{V} = \{1, 2, 3, 4, 5\}$ and $\boldsymbol{L}_{\mathcal{D}(\mathcal{M})} = \{\lambda_{12} = 6, \lambda_{13} = 7, \lambda_{35} = 8\}$.

On the other hand, if $\mathcal{M}$ is a MAG with vertex set $\boldsymbol{V}$, then we define the **canonical DAG** $\mathcal{D}(\mathcal{M})$ associated with $\mathcal{M}$ as follows (see [RS02] and Figure 3.1b):

Definition 3.2.1. Let $\mathcal{M}$ be a MAG with vertex set $\boldsymbol{V}$ and let $\boldsymbol{L}_{\mathcal{D}(\mathcal{M})} = \{\lambda_{XY} \mid X \leftrightarrow Y \text{ in } \mathcal{M}\}$. The canonical DAG $\mathcal{D}(\mathcal{M})$ has vertex set $\boldsymbol{V} \cup \boldsymbol{L}_{\mathcal{D}(\mathcal{M})}$ and edge set defined as:

$$\text{If } \left\{ \begin{array}{l} X \to Y \\ X \leftrightarrow Y \end{array} \right\} \text{ in } \mathcal{G} \text{ then } \left\{ \begin{array}{l} X \to Y \\ X \leftarrow \lambda_{XY} \to Y \end{array} \right\} \text{ in } \mathcal{D}(\mathcal{M}).$$

Definition 3.2.2. A probability distribution f is **consistent with a DAG** $\mathcal{D}$ if the pair $(\mathcal{D}, f)$ forms a CBN and f is **consistent with a MAG** $\mathcal{M}$ if there exists a CBN $(\mathcal{D}, g)$ such that $\mathcal{M}$ represents $\mathcal{D}$ and f is the observed marginal distribution of g, i.e., the distribution without taking into account latent variables.

Given the transformations established between DAGs and MAGs, the semantic interpretation of the latter type of graphs, and this last definitions, we can use the machinery of CBNs to calculate interventions on MAGs, such as those in the previous chapter.

3.3 Markov Equivalences Classes of MAGs

Several MAGs can also encode the same conditional independencies via m-separation. Such MAGs form a Markov Equivalence Class (MEC) which can be described uniquely by a Partial Ancestral Graph (PAG).

Definition 3.3.1. Let $[\mathcal{M}]$ be the MEC of an arbitrary MAG $\mathcal{M}$. The **Partial Ancestral Graph (PAG)** for $[\mathcal{M}]$, $\mathcal{P}_{[\mathcal{M}]}$, is a partial mixed graph with possibly three kinds of mark: arrowhead ($>$), tail ($-$) or a circle ($\circ$), such that

i $\mathcal{P}_{[\mathcal{M}]}$ has the same adjacencies as $\mathcal{M}$ (and any member of $[\mathcal{M}]$) does;

ii Every non-circle mark in $\mathcal{P}$ is an invariant mark in $[\mathcal{M}]$.

This is, a mark of arrowhead is in $\mathcal{P}_{[\mathcal{M}]}$ if and only if it is shared by all MAGs in $[\mathcal{M}]$; a mark of tail is in $\mathcal{P}_{[\mathcal{M}]}$ if and only if it is shared by all MAGs in $[\mathcal{M}]$; and a mark of circle is in $\mathcal{P}_{[\mathcal{M}]}$, otherwise. In Figure 3.2 we show a PAG and the MEC of MAGs it encodes.

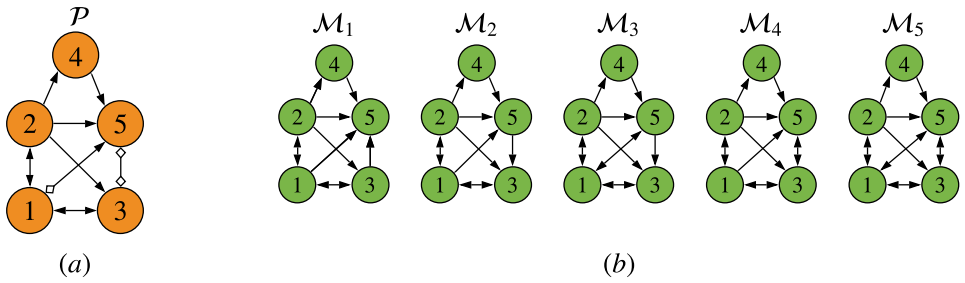


Figure 3.2 (a) A PAG $\mathcal{P}$ (orange) of five variables with three circle marks representing the MEC $[\mathcal{M}] = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5\}$. (b) The five MAG $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_5 \in [\mathcal{M}]$.

3.4 The Fast Causal Inference Algorithm

Causal discovery becomes especially challenging when the possibility of unmeasured common causes exist. Using the machinery of MAGs, there is a sound and complete causal discovery algorithm, known as the **Fast Causal Inference (FCI) algorithm**. The FCI algorithm can, under the standard assumptions of causal Markov condition, causal Faithfulness, acyclic structure and relaxing the sufficiency assumption, discover all aspects of

the causal structure that are uniquely determined by facts of probabilistic dependence and independence, i.e., a MEC for the causal MAG given observational data. The following are additional definitions needed to present the FCI algorithm.

In an ancestral graph, a path consisting of a triple of vertices (X, Y, Z) is an **unshielded triple** if and only if X is adjacent to Y but not Z, and Y and Z are adjacent. This triple is called as **unshielded collider** it both the edge between X and Y and the edge between Y and Z are into Y.

Definition 3.4.1. In a MAG, a path between X and Y, $\pi = (X, \ldots, W, V, Y)$. is a **discriminating path** for V if:

i π includes at least three edges;

ii V is a non-endpoint vertex on π, and is adjacent to Y on π; and

iii X is not adjacent to Y, and every vertex between X and V is a collider on π and is a parent of Y.

Definition 3.4.2. In a PAG, a path $p = (V_0, \ldots, V_k)$ is said to be **uncovered** if for every $1 \leq i \leq n-1$, V_{i-1} and V_{i+1} are not adjacent, i.e., if every consecutive triple on the path is unshielded.

Definition 3.4.3. In a PAG, a path $p = (V_0, \ldots, V_k)$ is said to be **potentially directed** from V_0 to V_n if for every $1 \leq i \leq n-1$, the edge between V_i and V_{i+1} is not into V_i or out to of V_{i+1}.

Intuitively, a potentially directed path is one that could be oriented into a directed path by changing the circles on the path into appropriate tails or arrowheads. As we shall see, uncovered potentially directed paths play an important role in locating invariant tails. A special case of a potentially directed path is where every edge on the path is of the form $\circ\!\!-\!\!\circ$; we call such a path a **circle path**.

We now describe the FCI algorithm according to Zhang in [Zha08b]. The algorithm consists mainly of two stages. In the first stage, the algorithm determines the adjacencies in the causal MAG. The inference of adjacencies is based on the fact that two variables are adjacent in a MAG if and only if they are not m-separated by any set of other variables in the MAG. So the basic idea is to search, for every pair of variables, a set of other variables that renders them conditionally independent. They are not adjacent if and only if such a set

is found. The second stage of the algorithm is to apply a set of seven orientation rules. A meta-symbol, asterisk ($*$), is used in the orientation rules as a wildcard that denotes any of the three marks: tail, arrowhead and circle. By this we mean the rule in question applies no matter which of the three marks actually appears in the position of $*$. It does not imply that all three marks can appear in that position.

The FCI algorithm receives a dataset $\boldsymbol{D}$ with the possibility of unmeasured common causes as input, and returns a PAG $\mathcal{P}$ representing the MEC of the MAG that we are interested in. The first stage of the FCI algorithm is the following: The algorithm first form a complete graph $\mathcal{U}$ on the set of variables, in which there is an edge $\circ\!\!-\!\!\circ$ between every pair of variables. Later, for every pair of variables X and Y, search for a set of other variables that render the two independent. If such as set $\boldsymbol{S}$ is found, remove the edge between X and Y in $\mathcal{U}$, and record $\boldsymbol{S}$ as $\boldsymbol{sepset}(X, Y)$. Let $\mathcal{P}$ be the graph resulting from the previous step. With this graph the algorithm execute the orientation Rule Cero:

Rule 0: For each unshielded triple X, Z, Y in $\mathcal{P}$, orient it as a collider $X *\!\!\rightarrow Z \leftarrow\!\!* Y$ if and only if Z is not in $\boldsymbol{sepset}(X, Y)$.

In the second stage the algorithm executes the following seven mark inference rules until none of them applies:

Rule 1: If $X *\!\!\rightarrow Y \circ\!\!-\!\!* Z$, and X and Z are not adjacent, then orient the triple as $X *\!\!\rightarrow Y \rightarrow Z$.

Rule 2: If $X \rightarrow Y *\!\!\rightarrow Z$ or $X *\!\!\rightarrow Y \rightarrow Z$, and $X *\!\!-\!\!\circ Z$, then orient $X *\!\!-\!\!\circ Z$ as $X *\!\!\rightarrow Z$.

Rule 3: If $X *\!\!\rightarrow Y \leftarrow\!\!* Z$, $X *\!\!-\!\!\circ W \circ\!\!-\!\!* Z$, X and Z are not adjacent, and $W *\!\!-\!\!\circ Y$, then orient $W *\!\!-\!\!\circ Y$ as $W *\!\!\rightarrow Y$.

Rule 4: If $u = (W, \ldots, X, Y, Z)$ is a discriminating path between W and Z for Y, and $Y \circ\!\!-\!\!* Z$; then if $Y \in \boldsymbol{sepset}(W, Z)$, orient $Y \circ\!\!-\!\!* Z$ as $Y \rightarrow Z$; otherwise orient the triple (X, Y, Z) as $X \leftrightarrow Y \leftrightarrow Z$.

Rule 5: If $X \rightarrow Y \rightarrow Z$ or $X -\!\!\circ Y \rightarrow Z$, and $X \circ\!\!\rightarrow Z$, orient $X \circ\!\!\rightarrow Z$ as $X \rightarrow Z$.

Rule 6: If $X \circ\!\!\rightarrow Z$, and $p = (X, Y, W, \ldots, Z)$ is an uncovered potentially directed path from X to Z such that Z and Y are not adjacent, then orient $X \circ\!\!\rightarrow Z$ as $X \rightarrow Z$.

Rule 7: Suppose $X \circ\!\!\rightarrow Z$, $Y \rightarrow Z \leftarrow W$, π_1 is an uncovered potentially directed path from X to Y, and π_2 is an uncovered potentially directed path from X to W. Let U be the vertex adjacent to X on π_1 (U could be Y), and T be the vertex adjacent to X on π_2 (T could be W). If U and T are distinct, and are not adjacent, then orient $X \circ\!\!\rightarrow Z$ as $X \rightarrow Z$.

As we describe before, a major virtue of ancestral graphs is that for any DAG with latent confounding, there is a unique MAG over the observed variables alone that represents the conditional independence relations and causal relations entailed by the original DAG. Instead of directly targeting the causal DAG, which for all we know might involve any number of latent variables, a more tractable goal for causal discovery is to learn as many features of the causal MAG as possible. Maximal ancestral graphs provide a neat representation of such causal systems without explicitly introducing unobserved variables, which facilitates automated search over (classes of) causal structures based on correlational information. There are several variants for the FCI algorithm such as the Greedy FCI (GFCI) (see [OSR16]), but the FCI is the standard method to learn a Markov equivalence class in causal insufficient systems.

Concluding Remarks

An important definition in this chapter is the one that establishes when a probability density is consistent with a MAG, since it allows us to directly use the theory on the calculation of interventions and the method of adjustment for covariates described in the previous chapter.

On the other hand, the transformation from DAGs to MAGs and vice versa is a recurring theme on which the main ideas of this research are based. In particular, the construction of the canonical DAG associated with a MAG is an operation that we will use in the proposed algorithm for estimating causal effects that is presented in subsequent chapters. It is important to emphasize that a MAG represents an infinity of DAGs with latent variables and the canonical DAG is just one DAG within this infinite collection, which can be built in a simple way.

Until now, it has been assumed that the causal structure is known, either as a DAG or a MAG, to calculate interventions by means of the covariate adjustment method. The next chapter describes a methodology in the area of causal inference to estimate bounds on causal effects when only the Markov equivalence class of the causal structure is known.

Part II

Related Work

Chapter 4

Estimating Bounds on Causal Effects

In this chapter, we first discuss the intervention-calculus when the DAG is absent (IDA) method for causal sufficiency systems and then its extension to insufficiency systems: the Latent Variable IDA algorithms (LV-IDA). We also discuss a procedure to determine all the MAG represented by a PAG, known as the Zhang MAG Listing or ZML algorithm.

4.1 The IDA algorithm

If we want to estimate the total causal effect of $X_i \in \boldsymbol{V} = \{X_1, \ldots, X_p\}$ on a response variable $Y = X_j \in \boldsymbol{V}$, when only a representation of the MEC of the causal structure is known, as is the case after a constraint-based causal structure learning methods are powered with only observational data, Maathuis et al. ([MKB09]) proposed the "Intervention-calculus when the DAG is Absent" (IDA) algorithm. The general idea of the IDA algorithm is as follows: After estimating a representation of the MEC given as a Completed Partial DAG (CPDAG), list all DAGs in the MEC and then apply covariate adjustment for each DAG, yielding an estimated total causal effect of X_i on Y for each possible DAG. All these total causal

effects, one for each DAG in the MEC, are collected in a multiset $\hat{\Theta}_i$, and the minimum and maximum value in $\hat{\Theta}_i$ are returned as bounds estimators for the true causal effect.

In other words, assuming that the MEC, given as a CPDAG, contains k different DAGs. For each DAG $\mathcal{D}_w$ in this MEC, we apply intervention calculus to obtain the causal effect θ_{iw} of X_i on Y. This can be represented as the estimation of a multiset $\Theta_i = \{\theta_{iw}\}$ for $w = 1, \ldots, k$ and $i = 1, \ldots, p$. Recalling that a multiset is a generalization of a set where the elements can have different multiplicities. So a set is the especial case of a multiset where all elements in it has multiplicity equal to one. If for instance all values θ_{iw}, $w = 1, \ldots, k$ in the multiset Θ_i are identical, i.e., the multiset Θ_i has one element of multiplicity k, we can establish that the causal effect of X_i on Y is uniquely determined. But, even if the multiset Θ_i contains distinct values, it still contains useful causal information to explore. For example, if $\theta_{iw} \neq 0$ for all $w = 1, \ldots, k$, then X_i must have a causal effect on Y (positive or negative). Similarly, if $\theta_{iw} > 0$ for all $w = 1, \ldots, k$, then X_i must have a positive causal effect on Y. More importantly, the minimum and the maximum value in Θ_i is a lower and upper bound, respectively, on the size of the causal effect of X_i on Y.

The IDA algorithm, shown in Algorithm 4.1, uses the fact that for the case of causal structures given as DAGs, the set of parents $\boldsymbol{pa}(X)$ is always an adjustment set for the post-intervention density $f(y \mid do(x))$, thus $f(y \mid do(x))$ is given by: (see Equation 2.9)

$$f(y \mid do(x)) = \begin{cases} f(y) & \text{if } Y \in \boldsymbol{pa}(X) \\ \int_{\boldsymbol{pa}(X)} f(y \mid x, \boldsymbol{pa}(x)) f(\boldsymbol{pa}(x))\, d\boldsymbol{pa}(x), & \text{otherwise.} \end{cases} \tag{4.1}$$

As the IDA algorithm considers the case where $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ are jointly Gaussian, we can express the later equation using expectations:

$$E(Y \mid do(x)) = \begin{cases} E(Y) & \text{if } Y \in \boldsymbol{pa}(X) \\ \int_{\boldsymbol{pa}(X)} E(Y \mid x, \boldsymbol{pa}(X)) f(\boldsymbol{pa}(X))\, d\boldsymbol{pa}(X), & \text{otherwise.} \end{cases} \tag{4.2}$$

Moreover, Gaussianity implies that $E(Y \mid x, \boldsymbol{pa}(X))$ is linear in x and $\boldsymbol{pa}(x)$ so

$$E(Y \mid x, \boldsymbol{pa}(X)) = \alpha + \beta x + \gamma^T \boldsymbol{pa}(x), \tag{4.3}$$

for some $\alpha, \beta \in \mathbb{R}$ and $\gamma \in \mathbb{R}^{|\boldsymbol{pa}(X)|}$. Combining this with Definition 2.3.2 it follows that

the causal effect of X_i on Y is given by β, the regression coefficient of X_i in the regression of Y on X_i and $\boldsymbol{pa}(X_i)$.

Algorithm 4.1 IDA

Input: A CPDAG C, conditional dependencies for $X_1, ..., X_p$, and a pair $(X_i, Y = X_j)$
Output: A multiset Θ_i of causal effects
Determine all DAGs $\mathcal{D}_1, ..., \mathcal{D}_k$ in the equivalence class of C
for $w = 1$ ***to*** k **do**
  $\theta_{iw} \leftarrow \beta$, where β is the coefficient of X_i in the regression $Y = \alpha + \beta x_i + \gamma^T \boldsymbol{pa}(x_i)$
end

4.2 The LV-IDA algorithm

The framework of the IDA algorithm was extended to insufficient causal systems by Malinsky and Spirtes in [MS17]. They named their algorithm LV-IDA (Latent Variables IDA), where they worked with PAGs as representations of MECs, and used the results in [MC15] to found adjustment set in MAGs for the covariate adjustment computations.

In order to construct an adjustment set for a pair of variables (X_i, Y) in each MAG represented by the PAG, the LV-IDA algorithm, shown in Algorithm 4.2, uses the generalized backdoor criteria given in [MC15]. This is a graphical criterion, i.e., it uses only the information given by the causal structure graph, to select a set of variables (adjustment set) on which the covariate adjustment method can be applied (See Chapter 2 Section 2.3). The following definitions are for the purpose of stating this criterion.

As seen in the previous section, the IDA algorithm uses the set of parents $\boldsymbol{pa}(X)$ of the variable X as the adjustment set. This set of parents of X always satisfies the backdoor criterion (see [Pea09]), i.e., it can be used as an adjustment set for a DAG. In the generalized backdoor criteria applied to MAGs, we have a similar result, but we need to use the set **D-SEP**(X, Y, M), defined below, instead of the parent set. It is worth mentioning that this set represents a different concept from the set of variables that d-separate a pair of variables X, Y in a DAG.

Definition 4.2.1. Let X and Y be two distinct vertices in a directed mixed graph $\mathcal{M}$. We say that $V \in$ **D-SEP** $(X, Y, \mathcal{M})$ if $V \neq X$ and there is a collider path between X and V in $\mathcal{M}$, such that every vertex on this path is an ancestor of X or Y in $\mathcal{M}$.

Definition 4.2.2. Given a MAG $\mathcal{M}$/ PAG $\mathcal{P}$, a directed edge $X \to Y$ in $\mathcal{M}/\mathcal{P}$ is **visible** if there is a vertex Z not adjacent to Y, such that there is an edge between Z and X that is into X , or there is a collider path between Z and X that is into X and every non-endpoint vertex on the path is a parent of Y. Otherwise $X \to Y$ is said to be **invisible**.

Let X be a vertex in $\mathcal{G}$, where $\mathcal{G}$ represents a causal MAG, or PAG. Let $\boldsymbol{adj}(X, \mathcal{G})$ the the adjacency set of X in $\mathcal{G}$. Let $\mathcal{R}$ be a MAG represented by $\mathcal{G}$, in the following sense. If $\mathcal{G}$ is a MAG, we simply let $\mathcal{R} = \mathcal{G}$. If $\mathcal{G}$ is a PAG, we let $\mathcal{R}$ be a MAG in the Markov equivalence class described by $\mathcal{G}$ with the same number of edges into X as $\mathcal{G}$. Let $\mathcal{R}_{\underline{X}}$ be the graph obtained from $\mathcal{R}$ by removing all directed edges out of X that are visible in $\mathcal{P}$. The set of possible descendents of X in G is denoted as $\boldsymbol{pDe}(X, \mathcal{G})$, where X_i is a possible descendents of X_j if there is a path from X_j to X_i with no arrowhead pointing towards X_j. Note that $\boldsymbol{pDe}(X, \mathcal{G})$ is equal to the set of descendents $\boldsymbol{de}(X, \mathcal{G})$ if $\mathcal{G}$ is a MAG.

Theorem 4.2.1. [MC15] Let X and Y be two distinct vertices in a MAG or PAG $\mathcal{G}$. If $Y \in \boldsymbol{adj}(X, \mathcal{R}_{\underline{X}})$ or **D-SEP**$(X, Y, \mathcal{R}_{\underline{X}}) \cap \boldsymbol{pDe}(X, \mathcal{G}) \neq \emptyset$, then $f(y \mid do(x))$ is not identifiable via the generalized back-door criterion. Otherwise **D-SEP**$(X, Y, \mathcal{R}_{\underline{X}})$ satisfies the generalized back-door criterion relative to (X, Y) and $\mathcal{G}$, i.e., **D-SEP**$(X, Y, \mathcal{R}_{\underline{X}})$ is an adjustments set for the pair (X, Y) in $\mathcal{G}$.

Then the set **D-SEP**$(X_i, Y, \mathcal{M}_{\underline{X_i}})$, when the condition $Y \in \boldsymbol{adj}(X_i, \mathcal{M}_{\underline{X_i}})$ or **D-SEP**$(X_i, Y, \mathcal{M}_{\underline{X_i}}) \cap \boldsymbol{pDe}(X_i, \mathcal{M}) \neq \emptyset$ is not met, is an adjustment set called a **black-door set** (see [MC15]) for the pair (X_i, Y) in the MAG $\mathcal{M}$.

Algorithm 4.2 LV-IDA

Input: A PDAG $\mathcal{P}$, conditional dependencies for $X_1, ..., X_p$, and a pair $(X_i, Y = X_j)$
Output: A multiset Θ_i of causal effects
Determine all MAGs $\mathcal{M}_1, \ldots, \mathcal{M}_k$ in the equivalence class of $\mathcal{P}$
for $w = 1$ ***to*** k **do**
  if $Y \in \boldsymbol{adj}(X_i, \mathcal{M}_{w,\underline{X_i}})$ or **D-SEP**$(X_i, Y, \mathcal{M}_{w,\underline{X_i}}) \cap \boldsymbol{pDe}(X_i, \mathcal{M}_w) \neq \emptyset$ **then**
    $\theta_{iw} \leftarrow NA$
  else
    $\boldsymbol{bds}(X_i, Y) \leftarrow$ **D-SEP**$(X_i, Y, \mathcal{M}_{w,\underline{X_i}})$
    $\theta_{iw} \leftarrow \beta$, where β is the coefficient of X_i in the regression $Y = \alpha + \beta x_i + \gamma^T \boldsymbol{bds}(X_i, Y)$
end

4.3 Determining the MAGs in a PAG

Listing all the MAGs represented by a PAG is more complicated than listing all the DAGs represented by a CPDAG. Such procedure would need to transform circle marks on $\circ\!\!\rightarrow$ and $\circ\!\!-\!\!\circ$ edges into tails and arrowheads, and deciding which further orientations in the graph are implied by these new tails and arrowheads, while preserving Markov equivalence. As some combinations of transformations could introduce new independence relationships among the variables.

A naive approach would be a brute force method that exhaustively tries every combination of circle mark transformations, and then checks if the resulting graph is Markov equivalent to the starting graph. For large graphs with many circle marks, there are just too many possible combinations of transformed marks and checking Markov equivalence for every resultant graph would require a lot of computation time.

Malisky and Spirtes proposed a procedure to determine all the MAGs represented by a PAG (see [MS17]). The algorithm is based on a transformational characterization of equivalence between MAGs proposed by Jiji Zhang (see [ZS05]) so they called it the ZML (Zhang MAG Listing) algorithm. Before this, there was no algorithm to enumerate all MAGs in a MEC given as a PAG. The ZML algorithm is shown in Algorithm 4.3.

Definition 4.3.1. The circle component of a PAG $\mathcal{P}$ denoted as $C(\mathcal{P})$ is the subgraph of $\mathcal{P}$ consisting of the vertices on $\circ\!\!-\!\!\circ$ edges.

Lemma 1. (see [ZS05]) Let $\mathcal{M}$ be an arbitrary MAG, and $A \rightarrow B$ an arbitrary directed edge in $\mathcal{M}$. Let $\mathcal{M}'$ be the graph identical to $\mathcal{M}$ except that the edge between A and B is $A \leftrightarrow B$. (In other words, $\mathcal{M}'$ is the result of simply changing $A \rightarrow B$ into $A \leftrightarrow B$ in $\mathcal{M}$.) $\mathcal{M}'$ is a MAG and Markov equivalent to $\mathcal{M}$ if and only if:

(i) there is no directed path from A to B other than $A \rightarrow B$ in $\mathcal{M}$;

(ii) for any $C \rightarrow A$ in $\mathcal{M}$, $C \rightarrow B$ is also in $\mathcal{M}$; and for any $D \leftrightarrow A$ in $\mathcal{M}$, either $D \rightarrow B$ or $D \leftrightarrow B$ is in $\mathcal{M}$;

(iii) there is no discriminating path for A on which B is the endpoint adjacent to A in $\mathcal{M}$.

Algorithm 4.3 ZML

Input: PAG $\mathcal{P}$
Output: A list of the MAGs represented by $\mathcal{P}$, called $[\mathcal{P}]$
Let $\mathcal{M} \leftarrow \mathcal{P}$
Transform all $\circ\!\!\rightarrow$ in $\mathcal{M}$, into $\rightarrow$
The remaining circle marks in $\mathcal{M}$ are on $\circ\!\!-\!\!\circ$ edges. For each possible orientation of $C(\mathcal{M})$ as a DAG with no new v-structures, add the resulting graph to $[\mathcal{P}]$
Let L be a list of circle mark location in $\mathcal{P}$
foreach $\mathcal{M}_k \in [\mathcal{P}]$ **do**
  for $l = 1$ ***to*** *the length of* L **do**
    foreach *sequence of circle marks in* L *of length* l **do**
      foreach *circle mark location in the sequence which is a tail in* $\mathcal{M}_k$ **do**
        (i.e., $X_i \rightarrow X_j$ in $\mathcal{M}$ but $X_i \circ\!\!\rightarrow X_j$ or $X_i \circ\!\!-\!\!\circ X_j$ in $\mathcal{P}$)
        Transform $X_i \rightarrow X_j$ in $\mathcal{M}_k$ to $X_i \leftrightarrow X_j$ if the conditions in Lemma 1 are satisfied
      end
      Add the resulting graph in $[\mathcal{P}]$ (Unless it is a duplicate)
    end
  end
end

Concluding Remarks

The IDA algorithm offers a solution to the problem of how to estimate bounds on the causal effect between pairs of variables in a system when only the Markov equivalence class (MEC) of the causal structure of the system is known. The IDA algorithm lists all the causal structures within the MEC and estimates the causal effect between the pair of variables on each of these structures. It is known that the real causal structure is one within the equivalence class, so the causal effect is bounded by the minimum and maximum value of the estimated effects for each of the structures within the MEC. The IDA algorithm assumes that all relevant variables in the system have been measured, i.e., that there are no latent confounders. Malinsky and Spirtes ([MS17]) generalized the ideas of the IDA algorithm by relaxing this assumption and proposed the LV-IDA algorithm, that we also discussed in this chapter.

The extension of IDA algorithm to insufficient systems modeled with MAG is very important, since the assumption of causal sufficiency, i.e., the assumption of the absence of latent direct common causes, is regularly violated in more realistic causal models. However, this refinement in the LV-IDA algorithm comes with an associated cost and with some limitations. In the first place, the unfolding of a PAG to list all the MAGs that belong to the MEC in the LV-IDA algorithm, is much more complicated and computationally more expensive than the unfolding of a CPDAG to list all the DAGs in the IDA algorithm. The LV-IDA algorithm uses the ZML algorithm (described above) to enumerate the MAGs in the MEC represented by a PAG. Although this algorithm is considerably more efficient than a brute force method for this enumeration, in practice is too slow even for graphs of moderate size (e.g. more than 10 variables), which is also reported in [MS17].

A considerable limitation in the extension of the IDA algorithm for insufficient causal systems is that unlike the case of causal sufficient systems, it is not always possible to estimate the causal effects using the covariant adjustment method on the members of the MEC. The reason for this is that it is not always possible to find an adjustment set for some pairs of variables in some MAGs. This is why, the LV-IDA algorithm returns missing values, i.e., NA values, in the causal effect estimation for some pairs of variables on some MAGs in the PAG.

It is not a minor limitation that the LV-IDA algorithm returns these missing values. Having at least one missing value within the multiset of causal effects estimated for a

pair of variables in the PAG, does not allow us to ensure that we know the maximum and minimum value of the causal effects in the MEC and so, not able to establish an accurate bounds of the causal effect between that pair of variables. In the next chapter we describe the LV-IDA+ algorithm, which is an extension to the LV-IDA algorithm proposed by us. The LV-IDA+ algorithm suggests a way to approximate the estimation on the causal effects when it is not possible to calculate them by the method of covariate adjustment directly over the MAGs in the Markov equivalence class.

Part III

Contribution to the Field

Chapter 5

The LV-IDA+ Framework

In this chapter we present the contribution of this dissertation: an algorithm to compute the bounds of the causal effects between any pair of variables (X, Y) in a insufficient system, i.e., considering non-measured direct causes. The proposed algorithm builds on the covariate adjustment method, so it uses causal structures represented by graphs to find an adjustment set, which is a sufficient set together with the treatment variable X, to compute the post-intervention density and estimate total causal effect between the pair of variables (X, Y).

Like the IDA and LV-IDA algorithms (see Chapter 4), the proposed algorithm that we called LV-IDA+ considers that it only has access to the Markov equivalence class of the estimable causal structure with observational data. Furthermore, it considers the case in which the system $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is jointly Gaussian, and as in the LV-IDA algorithm, the LV-IDA+ algorithm uses MAGs and PAGs, to represent the causal structure and the MEC, respectively, of the system.

Whereas in the LV-IDA algorithm is contemplated that in occasions there is no adjust-

ment set for the pair of variables (X, Y) in some, and occasionally all, MAGs in the PAG. The LV-IDA+ algorithm extends the LV-IDA algorithm using the adjustment sets of the canonical DAGs associated with the MAGs in the PAG to compute the bounds on the causal effects in these special cases. So our main contribution proposes a way to approximate the causal effect when these degenerate cases are presented.

5.1 The LV-IDA+ Algorithm

A schematic representation of the operation of the LV-IDA+ algorithm is shown in Figure 5.1. The LV-IDA+ algorithm (shown in Algorithm 5.4) computes a matrix of estimated causal effect intervals:

$$CE^* = \left([\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij}\right)_{p\times p}$$

where the interval $[\hat{\theta}_{min}, \hat{\theta}_{max}]$ is calculated for each pair (i, j) of the p variables in the system, i.e., $i, j = 1 \dots p$, and $\hat{\theta}_{min}$, $\hat{\theta}_{max}$ are the lower, upper bounds, respectively, for the estimated causal effect. Note that unlike the previous chapters and in particular Chapter 4, in this section, to describe the LV-IDA+ algorithm, we will use i and j to describe the pair of variables (i, j) on which we want to calculate the causal effect of the variable i over the variable j. Previously we were using X_i and $Y = X_j$ to denote this pair of variables, since the system was defined as $V = \{X_1, \dots X_p\}$. However, we will make this simplification to not overload the notation in this section for describing the matrix version of the LV-IDA+.

After listing all MAGs in the MEC encoded as a PAG by the ZML algorithm (see Algorithm 4.3), the LV-IDA+ algorithm computes a matrix $CE_{\mathcal{M}_w} = (\hat{\theta}_{ij}^{\mathcal{M}_w})_{p\times p}$ for each of the k MAGs in the MEC, where each total casual effect $\hat{\theta}_{ij}^{\mathcal{M}_w}$ is estimated using adjustment sets found by the sound and complete algorithms in [ZLT19] and [Per+18]. These algorithms guarantee finding adjustment sets for the variables of interest on MAGs when they exist. However, when this is not the case, the matrices $CE_{\mathcal{M}_w} = (\hat{\theta}_{ij}^{\mathcal{M}_w})_{p\times p}$ obtain NA values.

To approximate the missing values in the $CE_{\mathcal{M}_w} = (\hat{\theta}_{ij}^{\mathcal{M}_w})_{p\times p}$ matrices, LV-IDA+ generates the k canonical DAGs associated with each of the k MAGs in the MEC (See Figure 5.1 (c)). Then, it learns the parameters for each of these canonical DAGs and calculate the matrix of causal effects $CE_{\mathcal{D}_w} = (\hat{\theta}_{ij}^{\mathcal{D}_w})_{(p+l_w)\times(p+l_w)}$ for each of them, where l_w stands for the number of latent variables in the canonical DAG $\mathcal{D}_w$, for $w = 1, \dots, k$. It is always possible to find an adjustment set in a DAG, i.e., calculate a value for the estimation of the causal effect between a pair of variables using covariate adjustment (see Corollary 2.3.1).

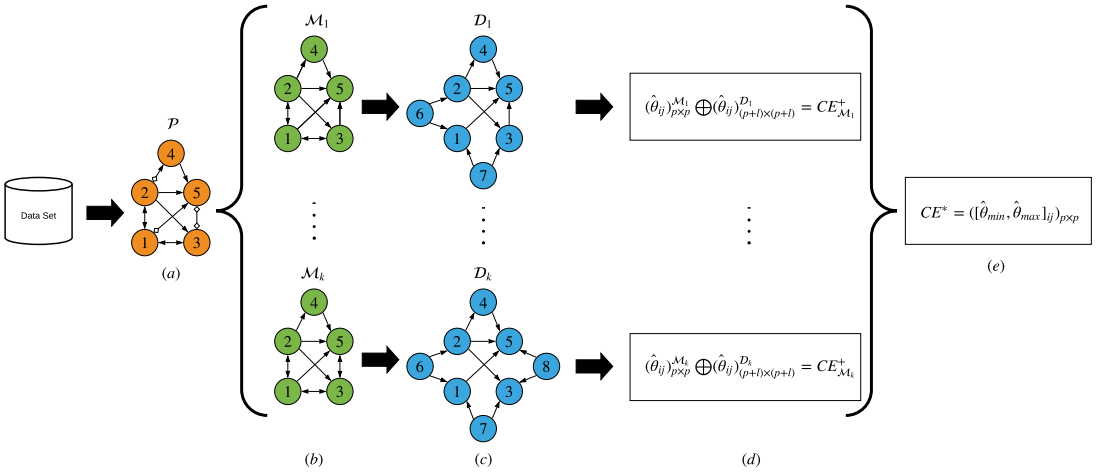


Figure 5.1 The LV-IDA+ framework. (a) A PAG $\mathcal{P}$ (orange) learned from data. (b) The set of MAGs (green) in the Markov equivalences class $[\mathcal{M}]$ encoded by $\mathcal{P}$. (c) The set of associated canonical DAGs (blue) from each of the MAGs in $[\mathcal{M}]$. (d) $CE_{\mathcal{M}_w} = (\hat{\theta}_{ij})^{\mathcal{M}_w}_{p\times p}$ and $CE_{\mathcal{D}_w} = (\hat{\theta}_{ij})^{\mathcal{D}_w}_{(p+l_w)\times(p+l_w)}$ are the square matrices of causal effects estimated from the MAG $\mathcal{M}_w$, and from the canonical DAG $\mathcal{D}_w$, respectively, where $w = 1 \ldots k$, p is the number of variables and l_w is the number of latent variables in the $\mathcal{D}_w$ DAG. Note that the $CE_{\mathcal{M}_w}$ matrices may have missing values but the $CE_{\mathcal{D}_w}$ matrices not. We denote by $\oplus$, the operation of substitute all the NA values in a $CE_{\mathcal{M}_w}$ matrix by the corresponding values in the $CE_{\mathcal{D}_w}$ matrix. (e) $CE^* = ([\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij})_{p\times p}$ denote the square matrix of interval of causal effect, where the extremes in the intervals $[\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij}$ are obtained by getting the minimum $\hat{\theta}_{min}$ and maximum $\hat{\theta}_{max}$, from the $\hat{\theta}_{ij}$ estimations in the $CE^+_{\mathcal{M}_w} = (\hat{\theta}_{ij})_{p\times p}$ matrices.

However it is important to note that even though some of the parameters can be learned from the original data, in the construction of the canonical DAG $\mathcal{D}_w$ for each MAG $\mathcal{M}_w$, hidden variables are introduced for which there are no data. This issue is solved by Expectation Maximization (EM) techniques to learn the parameters in Gaussians DAGs given the structure of the model in the presence of hidden variables (see Section 2.4.2 in Chapter 2).

With the matrices $CE_{\mathcal{M}_w}$ and $CE_{\mathcal{D}_w}$ computed, the LV-IDA+ algorithm obtains a matrix $CE^+_{\mathcal{M}_w}$ of size $p \times p$ (the same dimension as matrix $CE_{\mathcal{M}_w}$) for each of the k MAGs in the PAG by combining the matrices $CE_{\mathcal{M}_w}$ and $CE_{\mathcal{D}_w}$. This combination of the matrices $CE_{\mathcal{M}_w}$ and $CE_{\mathcal{D}_w}$ is such that if $\hat{\theta}^{\mathcal{M}_w}_{ij}$ is NA in $CE_{\mathcal{M}_w}$ it is replaced by the value of $\hat{\theta}^{\mathcal{D}_w}_{ij}$ in $CE_{\mathcal{D}_w}$, and the $\hat{\theta}^{\mathcal{M}_w}_{ij}$ values of the matrix $CE_{\mathcal{M}_w}$ are kept in $CE^+_{\mathcal{M}_w}$ in any other case. We use the symbol $\oplus$ to denote this operation of combining these matrices by substituting all the NA values in

a $CE_{\mathcal{M}_w}$ matrix by the corresponding values in the $CE_{\mathcal{D}_w}$ matrix.

Finally, the matrix of estimated causal effect intervals $CE^* = ([\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij})_{p\times p}$, i.e., the output of the LV-IDA+ algoritm, is computed by recording the minimum $(\hat{\theta}_{min})_{ij}$ and maximum $(\hat{\theta}_{max})_{ij}$ values for each of the k MAGs in the matrices $CE^+_{\mathcal{M}_w}$.

Algorithm 5.4 LV-IDA+

Input : A PAG $\mathcal{P}$, and a set of observational data D, for p variables
Output: A matrix of interval of causal effects $CE^* = ([\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij})_{p\times p}$
Determine all MAGs $\mathcal{M}_1, \ldots \mathcal{M}_k$ in the MEC $[\mathcal{M}]$ encoded as $\mathcal{P}$
for $w = 1, \ldots, k$ **do**
  Compute de canonical DAG $\mathcal{D}_w$ from $\mathcal{M}_w$
  Learn the parameters of $\mathcal{D}_w$ from data D using EM for the l_w hidden variables in $\mathcal{D}_w$
  for $i = 1, \ldots, p$ **do**
    for $j = 1, \ldots, p$ **do**
      Compute an adjustment set $\boldsymbol{Z}^{\mathcal{M}_w}_{(i,j)}$ for the pair of variables (i, j) in $\mathcal{M}_w$
      if an adjustment set $\boldsymbol{Z}^{\mathcal{M}_w}_{(i,j)}$ exists **then**
        $\hat{\theta}^{\mathcal{M}_w}_{ij} \leftarrow \beta$, where β is the coefficient of i in the regression $j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}^{\mathcal{M}_w}_{(i,j)})$ **else** $\hat{\theta}^{\mathcal{M}_w}_{ij} \leftarrow$ NA ;
      end
      Compute an adjustment set $\boldsymbol{Z}^{\mathcal{D}_w}_{(i,j)}$ for the pair of variables (i, j) in $\mathcal{D}_w$
      $\hat{\theta}^{\mathcal{D}_w}_{ij} \leftarrow \beta$, where β is the coefficient of i in the regression $j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}^{\mathcal{D}_w}_{(i,j)})$
    end
  end
  $CE^+_{\mathcal{M}_w} \leftarrow CE_{\mathcal{M}_w} \oplus CE_{\mathcal{D}_w}$
end
return $CE^* = ([\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij})_{p\times p}$

Recalling that for the case in which the system $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is jointly Gaussian $E(Y \mid do(x))$ can be calculated as the linear regression $Y = \alpha + \beta x + \gamma^T E(\boldsymbol{Z})$, were $\boldsymbol{Z}$ is an adjustments set. So the causal effect $\frac{\partial}{\partial x} E(Y \mid do(x))$ is given by β, the coefficient of x in the regression of Y on x and the adjustment set $\boldsymbol{z}$ (see Section 2.3 in Chapter 2).

We denote as $\boldsymbol{Z}^{\mathcal{M}_w}_{(i,j)}$ the adjustment set for the pair of variables (i, j) over the MAG $\mathcal{M}_w$ and by $\boldsymbol{Z}^{\mathcal{D}_w}_{(i,j)}$ the adjustment set for the pair of variables (i, j) in the DAG $\mathcal{D}_w$. So the estimated values of causal effects of the matrix $CE_{\mathcal{M}_w} = (\hat{\theta}^{\mathcal{M}_w}_{ij})_{p\times p}$ and the matrix $CE_{\mathcal{D}_w} = (\hat{\theta}^{\mathcal{D}_w}_{ij})_{(p+l_w)\times(p+l_w)}$ are computed in the LV-IDA algorithm (see Algorithm 5.4) as:

$$\hat{\theta}_{ij}^{\mathcal{M}_w} = \beta, \text{ where } \beta \text{ is the coefficient of } i \text{ in the regression } j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}_{(i,j)}^{\mathcal{M}_w})$$
$$\hat{\theta}_{ij}^{\mathcal{D}_w} = \beta, \text{ where } \beta \text{ is the coefficient of } i \text{ in the regression } j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}_{(i,j)}^{\mathcal{D}_w})$$

To finish this section, an example of the calculation of matrix $CE^{+}_{\mathcal{M}_w}$ using the matrices $CE_{\mathcal{M}_w}$ and $CE_{\mathcal{D}_w}$ is shown below. Equation 5.1 shows a matrix $CE_{\mathcal{M}_w}$ calculated for the MAG $\mathcal{M}_w$ within a MEC represented by a PAG $\mathcal{P}$ of $p = 5$ variables.

$$CE_{\mathcal{M}_w} = \begin{bmatrix} 1.00 & \text{NA} & \text{NA} & 0.01 & \text{NA} \\ \text{NA} & 1.00 & \text{NA} & 0.55 & \text{NA} \\ \text{NA} & \text{NA} & 1.00 & 0.00 & \text{NA} \\ 0.05 & \text{NA} & 0.01 & 1.00 & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & \text{NA} & 1.00 \end{bmatrix} \tag{5.1}$$

Equation 5.2 shows the matrix $CE_{\mathcal{D}_w}$ calculated from the canonical DAG $\mathcal{D}_w$, of eight variables associated with the MAG $\mathcal{M}_w$, with five variables. To estimate the causal effects of the $CE_{\mathcal{D}_w}$ matrix, the parameters for this canonical DAG structure are estimated along with the data of the three aggregated latent variables. This is precisely the problem of learning parameters with hidden variables given a structure and it is solved using expectation-maximization as seen in Chapter 2 section 2.4.2. Note that in the matrix $CE_{\mathcal{D}_w}$ the causal effects between the ordered pairs of the eight variables are estimated so it is an square eight dimensional matrix.

$$CE_{\mathcal{D}_w} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.01 & 0.67 & 0.03 & 0.15 & 0.00 \\ 0.00 & 1.00 & 0.80 & 0.55 & 1.69 & 0.06 & 0.00 & 0.00 \\ 0.00 & 0.48 & 1.00 & 0.00 & 0.00 & 0.00 & 0.03 & 0.00 \\ 0.05 & 0.42 & 0.01 & 1.00 & 0.61 & 0.00 & 0.00 & 0.00 \\ 0.14 & 0.35 & 0.00 & 0.32 & 1.00 & 0.00 & 0.00 & 0.01 \\ 0.04 & 0.06 & 0.05 & 0.03 & 0.14 & 1.00 & 0.00 & 0.00 \\ 0.15 & 0.00 & 0.05 & 0.00 & 0.01 & 0.00 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.00 & 0.00 & 1.00 \end{bmatrix} \tag{5.2}$$

Equation 5.3 shows the matrix $CE^{+}_{\mathcal{M}_w}$ calculated from the matrices $CE_{\mathcal{M}_w}$ y $CE_{\mathcal{D}_w}$. The idea is to return estimates of the causal effect between all ordered pairs of the original variables in the matrix $CE^{+}_{\mathcal{M}_w}$, i.e., the matrix $CE^{+}_{\mathcal{M}_w}$ is of the same dimension as the matrix

$CE_{\mathcal{M}_w}$. To obtain estimates of the causal effect between all the ordered pairs of the original variables, all the estimates achieved in the matrix $CE_{\mathcal{M}_w}$ are kept on the matrix $CE^+_{\mathcal{M}_w}$. Those estimates not achieved in the $CE_{\mathcal{M}_w}$ matrix, i.e., those with a NA value, are replaced by the estimates in the respective row and column of the matrix $CE_{\mathcal{D}_w}$ on the matrix $CE^+_{\mathcal{M}_w}$. For example, the element $(5, 1)$ (row 5 and column 1, from top to bottom and from left to right) of the matrix $CE_{\mathcal{M}_w}$ is a NA value so in the matrix $CE^+_{\mathcal{M}_w}$ this value is replaced by element $(5, 1)$ of the matrix $CE_{\mathcal{D}_w}$. On the other hand, the element $(2, 4)$ in the matrix $CE_{\mathcal{M}_w}$ is not a NA value and therefore, this value is kept in the matrix $CE^+_{\mathcal{M}_w}$.

$$CE^+_{\mathcal{M}_w} = CE_{\mathcal{M}_w} \oplus CE_{\mathcal{D}_w} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.01 & 0.67 \\ 0.00 & 1.00 & 0.80 & 0.55 & 1.69 \\ 0.00 & 0.48 & 1.00 & 0.00 & 0.00 \\ 0.05 & 0.42 & 0.01 & 1.00 & 0.61 \\ 0.14 & 0.35 & 0.00 & 0.32 & 1.00 \end{bmatrix} \tag{5.3}$$

5.2 Differences between LV-IDA+ and LV-IDA

The main difference between the LV-IDA algorithm with respect to the proposed LV-IDA+ algorithm is that LV-IDA+ always guarantees the calculation of the causal effect between any pair of variables in the system. Whereas LV-IDA cannot always calculate this effect for some pairs of variables and then occasionally returns missing values as output. A mayor drawback of the LV-IDA algorithm is that it is not always possible to find an adjustment set for some pairs of nodes in some MAGs to perform covariate adjustment. In such cases, the estimated multisets of total causal effects $\hat{\mathbf{\Theta}}$ contains missing values and the fundamental idea of using the minimum and maximum value in $\hat{\mathbf{\Theta}}$ to bound the true causal effect is not longer valid. We refer to this kind of missing values as **simple NAs**. Moreover, in some cases there are only missing values in $\hat{\mathbf{\Theta}}$ for some pairs of variables. We call these more problematic types of missing values, **extreme NAs**.

Suppose that the LV-IDA+ and LV-IDA algorithms receive as input a PAG $\mathcal{P}$, which represents a MEC made up of five MAGs: $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_5$. Suppose LV-IDA estimates the following causal effects for the pair of variables (X, Y): $\theta_{\mathcal{M}_1}$ = NA over the MAG $\mathcal{M}_1$, $\theta_{\mathcal{M}_2} = 0.44$ over the $\mathcal{M}_2$, $\theta_{\mathcal{M}_3}$ = NA over the MAG $\mathcal{M}_3$, $\theta_{\mathcal{M}_4} = .38$ over the MAG $\mathcal{M}_4$ y $\theta_{\mathcal{M}_5} = .44$ over the MAG $\mathcal{M}_5$, i.e., the estimated multiset $\mathbf{\Theta} = \{\theta_{\mathcal{M}_1}, \theta_{\mathcal{M}_2}, \theta_{\mathcal{M}_3}, \theta_{\mathcal{M}_4}, \theta_{\mathcal{M}_5}\}$ for the pair (X, Y) is $\mathbf{\Theta} = \{\text{NA}\ , 0.44, \text{NA}\ , 0.38, 0.44\}$ and the interval of causal effect estimated

by LV-IDA is $[0.38, 0.44]$. When at least one of the values in the multiset estimated by LV-IDA is different from NA, LV-IDA can return an interval of causal effects for some pair of variables. We call the NA values on the estimated multiset simple NAs. On the other hand, suppose that for the pair of variables (X, Z), with $Z \neq Y$, LV-IDA estimates the multi-set of causal effects $\Theta = \{\theta_{\mathcal{M}_1}, \theta_{\mathcal{M}_2}, \theta_{\mathcal{M}_3}, \theta_{\mathcal{M}_4}, \theta_{\mathcal{M}_5}\}$, given by $\Theta = \{\text{NA}\ , \text{NA}\ , \text{NA}\ , \text{NA}\ , \text{NA}\ \}$. In this case, LV-IDA does not return a causal effect interval for the pair (X, Z), but rather a value of NA on the returned interval matrix. We call these types of NAs, extreme NAs. That is, only when the estimated multiset of causal effects is completely made up with simple NAs, then we have an extreme NA. Which translates to a value of NA in the matrix of intervals returned by LV-IDA.

The LV-IDA+ algorithm (Algorithm 5.4) is proposed to calculate the causal effect between all pairs of variables $(X_i, Y = X_j)$ in a system $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ and returns an interval matrix, while LV-IDA (as described in Algorithm 5.5) returns the multiset Θ of causal effects between a single pair of variables $(X_i, Y = X_j)$. However, LVIDA+ can be rewritten to return such multiset of causal effects without missing values, i.e., a multiset version of LV-IDA+ as shown in Algorithm 5.5.

In this version of the LV-IDA+ algorithm, it is shown that only when there is no adjustment set for one of the k MAGs in the PAG then the canonical DAG associated with this MAG is constructed and an adjustment set for this DAG is used to approximate the causal effect of the (i, j) pair in this MAG.

Another difference between these algorithms is that LV-IDA uses the generalized backdoor criterion to find adjustments sets in the MAGs of the Markov equivalence class associated with the input PAG (see Section 4.2). The generalized backdoor criterion is a sound but not a complete criterion. On the other hand, the LV-IDA+ algorithm uses the sound and completeness criterion given by Perkovic and van der Zander in [Per+18] and [ZLT19]. This graphical criterion proposed by Perkovic and van der Zander is based on the generalized backdoor criterion (See Section 4.2 Theorem 4.2.1), but unlike this, this criteria guarantees to find an adjustment set for a given pair of variables and a given MAG when it exists.

Next we describe the graphic criteria of Perkovic et al. given in [Per+18] that is equivalent to the one given in [ZLT19]. To this end, we introduce some additional terminology. Remember that a direct path from X to Y is a path from X to Y in which all edges are directed towards Y. We also refer to this as a **causal path**. A **possibly directed path** or

Algorithm 5.5 LV-IDA+ Multiset version

Input : A PAG $\mathcal{P}$, a set of observational data D, and a pair of variables (i, j) $i, j = 1, \ldots, p$
Output: A multiset Θ_i of causal effects
Determine all MAGs $\mathcal{M}_1, \ldots \mathcal{M}_k$ in the MEC $[\mathcal{M}]$ encoded as $\mathcal{P}$
for $w = 1, \ldots, k$ **do**
  Compute an adjustment set $\boldsymbol{Z}_{(i,j)}^{\mathcal{M}_w}$ for the pair of variables (i, j) in $\mathcal{M}_w$
  if an adjustment set $\boldsymbol{Z}_{(i,j)}^{\mathcal{M}_w}$ exists **then**
    $\hat{\theta}_{iw} \leftarrow \beta$, where β is the coefficient of i in the regression $j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}_{(i,j)}^{\mathcal{M}_w})$
    else
      Compute de canonical DAG $\mathcal{D}_w$ from $\mathcal{M}_w$
      Learn the parameters of $\mathcal{D}_w$ from data D using EM for the l_w hidden variables in $\mathcal{D}_w$
      Compute an adjustment set $\boldsymbol{Z}_{(i,j)}^{\mathcal{D}_w}$ for the pair of variables (i, j) in $\mathcal{D}_w$
      $\hat{\theta}_{iw} \leftarrow \beta$, where β is the coefficient of i in the regression $j = \alpha + \beta i + \gamma^T E(\boldsymbol{Z}_{(i,j)}^{\mathcal{D}_w})$
    end
end

possibly causal path from X to Y is a path from X to Y that does not contains an arrowhead pointing in the direction of X. For example $X \to \cdots \leftarrow\!\!-\!\circ \cdots \to Y$ is not a possibly causal path, but $X \to \cdots \circ\!\!-\!\!\to \cdots \to Y$ it is. If there is a directed (possibly directed) path from X to Y, then X is a ancestor (**possible ancestor**) of Y, and Y is a descendant (**possible descendant**) of X. We also use the convention that every node is a descendant, possible descendant, ancestor and possible ancestor of itself.

Definition 5.2.1. Let X and Y a pair of vertices in a MAG $\mathcal{M}$. Then $\mathcal{M}$ is said to be **amenable** relative to (X, Y) if every proper possible directed path from X to Y in $\mathcal{M}$ starts with a visible edge out of X (see Definition 4.2.2 for the definition of visible edge).

Definition 5.2.2. Let X y Y be two different vertices in a MAG $\mathcal{M}$. The **proper back-door graph** $\mathcal{M}_{XY}^{pbd}$ is obtained from $\mathcal{M}$ by removing all visible edges out of X that are possibly directed paths from X to Y in $\mathcal{M}$.

Definition 5.2.3. Let X and Y be two different vertices in the set of vertices $\boldsymbol{V}$ in a MAG $\mathcal{M}$. Then the **forbidden set** relative to the pair (X, Y) is defied as:

$$\boldsymbol{Forb}(X, Y, \mathcal{M}) = \{W' \in \boldsymbol{V} : W' \in \boldsymbol{PossDe}(W, G), \text{ for some } W \neq X \text{ which lies on a possibly directed path from } X \text{ to } Y \text{ in } \mathcal{M}\}.$$

Definition 5.2.4. Let X y Y be two different vertices and $\boldsymbol{Z}$ a set of vertices such as $X, Y \notin \boldsymbol{Z}$ in a MAG $\mathcal{M}$. Then Z satisfied the generalized adjustment criterion relative to (X, Y) in $\mathcal{M}$ if the following three conditions hold:

i (**Amenability**) $\mathcal{M}$ is amenable relative to (X, Y), and

ii (**Forbidden set**) $\boldsymbol{Z} \cap \boldsymbol{Forb}(X, Y, \mathcal{M}) = \emptyset$, and

iii (**Separation**) $\boldsymbol{Z}$ m-separates X y Y in $\mathcal{M}_{XY}^{pbd}$.

Theorem 5.2.1. Let X y Y be two different vertices and $\boldsymbol{Z}$ a set of vertices such as $X, Y \notin \boldsymbol{Z}$ in a causal MAG $\mathcal{M}$. Then $\boldsymbol{Z}$ is an adjustment set relative to (X, Y) in $\mathcal{M}$ if and only if $\boldsymbol{Z}$ satisfies the generalized adjustment criteria relative to (X, Y) in $\mathcal{M}$. (See [Per+18].)

In Figure 5.2 four MAGs are exemplified for which an adjustment set relative to the pair of variables (X, Y) is sought.

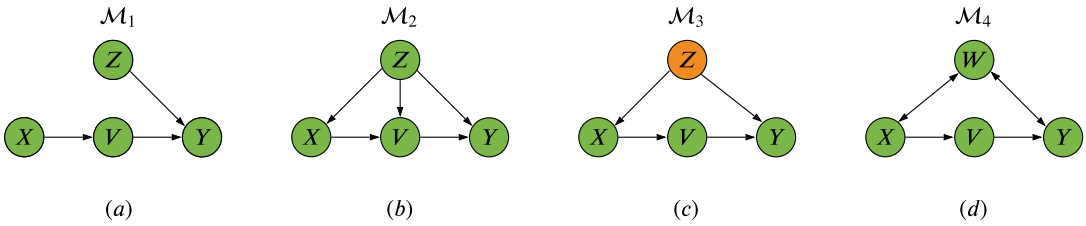


Figure 5.2 (a)-(b) The MAGs $\mathcal{M}_1$ and $\mathcal{M}_2$ are not amenable relative to (X, Y) since the edge $X \rightarrow Y$ is not visible, so no adjustment sets exists for pair (X, Y) in this two MAGs. (c) The only valid adjustment set in $\mathcal{M}_3$ is $\{Z\}$. (d) The empty set is the only valid adjustment set $\mathcal{M}_4$.

5.3 Rationale and Limitations

The rationale behind our method is that it is reasonable to calculate the causal effect on one of the DAGs that belongs to the set of DAGs represented by the MAG as an approximation to the causal effect on a pair of variables when this cannot be calculated directly on the MAG in question. With this idea we have answer the question about what would be a basic approximation to estimate the effect on a pair of variables when there are no adjustment sets on a MAG, and avoid to answer that such effect cannot be calculated.

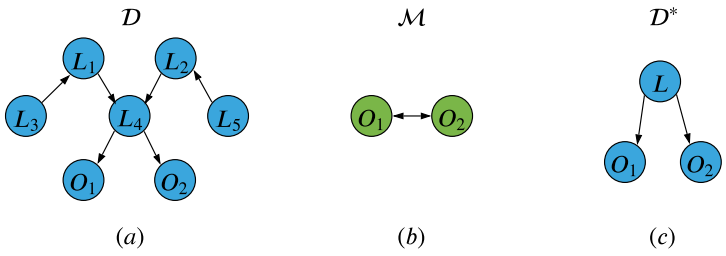


Figure 5.3 (a) A DAG $\mathcal{D}$ with a complex substructure of several unmeasured variables, $\boldsymbol{L} = \{L_1, L_2, L_3, L_4, L_5\}$ and $L^* = L_4$. (b) The MAG $\mathcal{M}$ obtained from $\mathcal{D}$ by marginalizing the latent variables L_1, L_2, L_3, L_4 and L_5. (c) The canonical DAG $\mathcal{D}^*$ associated to MAG $\mathcal{M}$.

If we assume that we know the DAG $\mathcal{D}$ with latent variables that generated the data and we marginalize the latent variables in it, we find the MAG that represents this DAG $\mathcal{D}$. In practice, from the observed variables data in the system we can generate a MAG and avoid worrying about specifying the DAG $\mathcal{D}$ with latent variables that generates the data. Since trying to find directly this DAG with latent variables from the observed variables data is a much more difficult problem, and in fact an open problem [SGS00]. An important disadvantage on the simplification of using MAGs to represent DAGs with latent variables is that the transformation of this type of DAGs to MAGs is not injective, i.e. several DAGs with latent variables are transformed into the same MAG. Furthermore, an infinite number of DAGs with latent variables are transformed into the same MAG. In this sense, there is no inverse transformation with which we could recover the DAG with latent variables directly. However, we can at least rebuild some of the DAGs represented by a MAG. For the proposed LV-IDA+ algorithm, we chose a very particular DAG within this infinite set of DAGs represented by a MAG: the canonical DAG associated with the MAG. This canonical DAG has the advantage that it can be efficiently built from a MAG (as shown in Chapter 3). In addition, this type of DAGs encompasses the most significant aspects of the causal structure of an infinity subset of DAGs represented by a MAG.

From the exercise of marginalizing the latent variables of a DAG $\mathcal{D}$ to construct a MAG $\mathcal{M}$ and then construct the canonical DAG $\mathcal{D}^*$ associated with this MAG $\mathcal{M}$, we were able to identify two types of substructure patterns in which although a canonical DAG $\mathcal{D}^*$ is not an exact portrayal of the original DAG $\mathcal{D}$, a canonical DAG offers a good representation for estimating causal effects (see Figures 5.3 and 5.4). The first type of

substructure pattern is one in which the original $\mathcal{D}$ DAG contains a complex substructure of a set $\boldsymbol{L}$ of several unmeasured variables and one of the latent variables L^* is a common cause of a pair of observed variables, say O_1 and O_2. When the latent variables in this substructure are marginalized, a MAG $\mathcal{M}$ is created where none of the latent variables of the substructure appear and the MAG $\mathcal{D}$ contains a bidirectional edge between O_1 and O_2. Then, when constructing the canonical DAG $\mathcal{D}^*$ for this MAG $\mathcal{M}$, a single variable L is added as a common cause for the observed variables O_1 and O_2 (see Figure 5.3 for an instance of this substructure pattern). For this first substructure pattern, it can be seen that the weight of all latent variables in the substructure can be absorbed by a single variable L, just as in the reconstruction of the original DAG $\mathcal{D}$ by mean of a canonical DAG $\mathcal{D}^*$. Therefore, a canonical DAG is a good representation for estimating the causal effect on the class of DAGs that present this type of substructure pattern.

The second type of structural pattern occurs when there is a latent variable L that is a common cause of not only a couple of measured variables but several more, this translates into several bi-directed edges in the MAG $\mathcal{M}$ when this variable is marginalized. When the canonical DAG associated to this MAG is built, a new variable is added for each bi-directed edge, although in reality it is a single unmeasured variable which is the common cause for several variables on the bi-directed edges (see Figure 5.4 for an instance of this substructure pattern). This structural pattern is not problematic for the calculation of the causal effect with LV-IDA+ as it seems, since what happens is that this unmeasured common cause is simply repeated for each pair of variables measured in the canonical DAG. For the LV-IDA+ algorithm, the essential is that an unmeasured common cause has been identified at each bidirectional edge, which is then solved by expectation maximization.

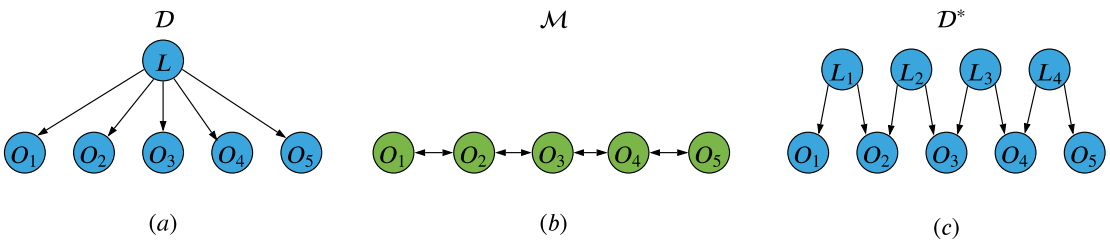


Figure 5.4 (a) A DAG $\mathcal{D}$ with a latent variable L that is a common cause for several observed variables $\{O_1, O_2, O_3, O_4, O_5\}$. (b) The MAG $\mathcal{M}$ obtained from $\mathcal{D}$ by marginalizing the latent variable L. (c) The canonical DAG $\mathcal{D}^*$ associated to MAG $\mathcal{M}$.

On the other hand, a class of DAGs that we identify are not well represented by a canonical DAG are those we call anti-canonical DAGs. We call anti-canonical DAG to a DAG that is built from a MAG $\mathcal{M}$ in which at least one directed edge from a variable O_1 to a variable O_2 in the MAG $\mathcal{M}$ is transformed into the directed edge from O_1 to O_2 and augmenting a latent variable L as the common cause of O_1 and O_2 in the DAG (see Figure 5.5 (c)). We call to this transformation from a directed edge in a MAG to a directed edge with a latent common cause in the DAG an anti-canonical pattern. If all the directed edges in the MAG are transformed as anti-canonical patterns we call the resulting DAG, a total anti-canonical DAG (see Figure 5.5 (d)).

Anti-canonical DAGs are problematic because, if in the original DAG $\mathcal{D}$ that generates the data there is an unmeasured common cause L for two variables O_1 and O_2, and in addition, O_1 is the direct cause of O_2. When forming a MAG $\mathcal{M}$ by marginalizing L from the DAG $\mathcal{D}$, the MAG $\mathcal{M}$ has the directed edge from O_1 to O_2 (not a bi-directed edge indicating the latent variable L) and then the canonical DAG $\mathcal{D}^*$ associated with $\mathcal{M}$ would have simply the same directed edge from O_1 to O_2. This means that if we construct the canonical DAG $\mathcal{D}^*$ associated to $\mathcal{M}$ we are ruling out the existence of the unmeasured common cause L. When a direct common cause is not contemplated in a system, this causes greater values than what they actually are on the estimations of some causal effects among the variables in the system. So we identify this case as a possible limitation on the proposal to approximate the calculation of the causal effect using the canonical DAG when it is not possible to calculate it for a couple of variables directly on the MAG.

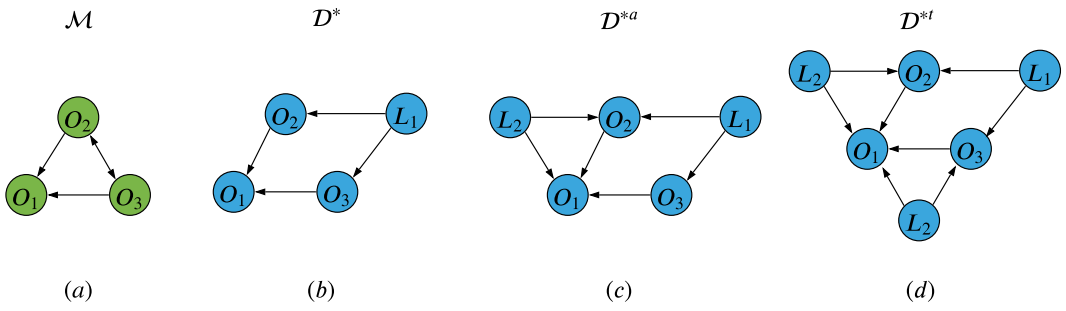


Figure 5.5 (a) A MAG $\mathcal{M}$ with two directed edges and one bi-directed edge.(b) The canonical DAG associated to the MAG $\mathcal{M}$. (c) An anti-canonical DAG $\mathcal{D}^{*a}$ associated to the MAG $\mathcal{M}$ with an anti-canonical pattern over the directed edge from O_1 to O_2. (d) The total anti-canonical DAG $\mathcal{D}^{*t}$ associated to the MAG $\mathcal{M}$.

Concluding Remarks

The main contribution of the LV-IDA+ algorithm is to provide a way to approximate the estimation of the causal effects when these cannot be calculated by means of covariate adjustment directly on some of the MAGs in the Markov equivalence class. The idea to approximate these values is to estimate the causal effect on a representative DAG on the collection of DAGs represented by a MAG when this estimation cannot be made directly on the MAG in question, as it is always possible to estimate the causal effect on a DAG. In particular, we use the canonical DAG associated with a MAG for this purpose. When the canonical DAG is built for a MAG, variables are added for which their data is not known. In the LV-IDA + algorithm, we solved this problem by reducing it to a parameter learning problem with hidden variables and we use expectation-maximization to obtain the parametrization of this structure, with which we have all the ingredients to apply covariate adjustment over the canonical DAG.

As we explained in this chapter, we chose the canonical DAG as representative of the infinite collection of DAGs associated with a MAG to make this approximation in the first place because it is easy to build but also because it is a good representation for a large collection of DAGs associated with a MAG. After understanding in detail the implications of approximating the causal effect using canonical DAGs, we identified that the main limitation of the LV-IDA algorithm is when what we call the anti-canonical patterns are present in the construction of the MAG. For these cases, the information of the existence of latent cofactors is lost and this cause an overestimation in the causal effects for some pairs of variables in the system.

We recognize that more sophisticated approximations can be formulated but they would imply a higher computational cost. For example, one could average the causal effects calculated on: the canonical DAG, a set of anti-canonical DAGs and the total anti-canonical DAG. Another interesting line of research would be to design heuristics that based on the data, could recognize anti-canonical patterns and add these substructures to the canonical DAG, to get better approximations in these cases. However, the proposed form of approximation is in principle an efficient baseline that can motivate the search for better approximations. Since, to our knowledge, no other way of approximating causal effects has been proposed when they cannot be calculated by covariate adjustments directly in a MAG.

Chapter 6

LV-IDA+ Experimental Evaluation

In this chapter we experimentally evaluate the proposed LV-IDA+ algorithm. We restrict the evaluation of the LV-IDA+ algorithm to synthetic models given by canonical DAGs and compare the accuracy of the intervals of causal effect found with those build by the LV-IDA algorithm.

6.1 Evaluation Metrics.

To evaluate the accuracy of the causal effect interval matrix returned by LV-IDA+ and compare this against the calculated by LV-IDA, we use the Average Interval Mean Square Error, based on the Interval Mean Square Error metric (see [MS17]).

Let $CEReal = (\theta_{ij})$ be the real total causal effects matrix, where θ_{ij} is the true total causal effect from the pair of variables (i, j) for $i, j = 1, \ldots, p$. The Interval Mean Square Error metric (IntMSE) between the true total effect θ_{ij} from $CEReal$, and the estimate interval of causal effect $[\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij}$ from the CE^* matrix, returned by the LV-IDA+ or the LV-IDA algorithms, is defined as:

$$\text{IntMSE} = \begin{cases} 0 & \text{if } \theta_{ij} \in [\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij} \\ \min\{|\theta_{ij} - (\hat{\theta}_{min})_{ij}|, |\theta_{ij} - (\hat{\theta}_{max})_{ij}|\} & \text{otherwise,} \end{cases} \tag{6.1}$$

Let $IEM = (\epsilon_{ij})$ be the Interval Error Matrix where the ϵ_{ij} are defined as the IntMSE between the true total effect θ_{ij} in $CEReal$, and the estimate interval of causal effect $[\hat{\theta}_{min}, \hat{\theta}_{max}]_{ij}$ in CE^*. Let Σ be the sum of all the elements ϵ_{ij} of the Interval Error Matrix $IEM = (\epsilon_{ij})$. The Average Interval Mean Square Error (AIntMSE), is defined as:

$$\text{AIntMSE} = \frac{\Sigma}{p^2 - (p + \text{eNA})},$$

where p is the number of variables and eNA is the number of extreme NAs. Note that the

denominator corresponds to the total number of elements of the matrices CE^*, minus the elements of the diagonal (which always has value [1, 1]), and the number of extreme NAs, which are the returned NA values on the matrices CE^* when it is not possible to estimated interval of causal effect by the algorithms. This corresponds to the number of estimated interval actually calculated by the LV-IDA+ or LV-IDA algorithms, so this evaluation gives us an average of the error in the calculation of the estimated intervals of causal effect in the CE^* matrices.

The AIntMSE measure does not advantage the LV-IDA+ algorithm in any way over the LV-IDA algorithm. The LV-IDA+ algorithm does not return intervals matrices with NA values unlike LV-IDA. This implies that the denominator in the fraction of AIntMSE when evaluating the interval matrix returned by LV-IDA+ for a model $\mathcal{G}$ and data set $\boldsymbol{D}$ is always greater than or equal to the denominator of AIntMSE for the interval matrix returned by LV-IDA for the same model $\mathcal{G}$ and the set $\boldsymbol{D}$. However, when it is the case that this denominator is greater, say by r units of difference, it means that r more estimates were also added in the numerator Σ for LV-IDA+. So as we said before, AIntMSE is the average of the IntMSE error over the estimates actually computed by the LV-IDA+ and LV-IDA algorithms, without penalizing the extreme NA values.

To clarify how the AIntMSE measure is calculated, below is an example of its calculation for a real case during experimentation. Below we show the real total causal effect matrix $CEReal = (\theta_{ij})$ calculated from a synthetic model $\mathcal{G}$ with $p = 5$ variables, and present the interval matrices CE^* returned by the LV-IDA+ and LV-IDA algorithms form the synthetic model $\mathcal{G}$. Then, the Interval Error Matrices $IEM = (\epsilon_{ij})$ matrices for each of the LV-IDA+ and LV-IDA CE^* matrices are shown. Finally, we show the calculation of the AIntMSE for each of the CE^* matrices, to show the way the AIntMSE metric is computed.

Equation 6.2 shows the $CEReal$ matrix that is calculated from a synthetic model $\mathcal{G}$ of $p = 5$ variables.

$$CEReal = (\theta_{ij}) = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.67 \\ 0.00 & 1.00 & 0.79 & 0.58 & 1.74 \\ 0.00 & 0.30 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.53 & 0.00 & 1.00 & 0.62 \\ 0.40 & 0.31 & 0.00 & 0.25 & 1.00 \end{bmatrix} \tag{6.2}$$

Equation 6.3 shows the CE^* matrix returned by the LV-IDA+ algorithm for the G

model.

$$CE^* = \begin{bmatrix} [1.00, 1.00] & [0.00, 0.00] & [0.00, 0.16] & [0.01, 0.01] & [0.00, 0.67] \\ [0.00, 0.00] & [1.00, 1.00] & [0.68, 0.80] & [0.55, 0.55] & [1.69, 2.18] \\ [0.00, 0.00] & [0.43, 0.48] & [1.00, 1.00] & [-0.04, 0.00] & [0.00, 1.02] \\ [0.05, 0.05] & [0.42, 0.42] & [0.02, 0.15] & [1.00, 1.00] & [0.61, 0.64] \\ [0.00, 0.14] & [0.34, 0.36] & [0.00, 0.32] & [0.29, 0.33] & [1.00, 1.00] \end{bmatrix} \tag{6.3}$$

Equation 6.4 shows the CE^* matrix returned by the LV-IDA algorithm for the G model.

$$CE^* = \begin{bmatrix} [1.00, 1.00] & [0.00, 0.00] & [0.00, 0.00] & [0.00, 0.00] & [0.00, 0.00] \\ [0.00, 0.00] & [1.00, 1.00] & \text{NA} & [0.55, 0.55] & \text{NA} \\ [0.00, 0.00] & [0.00, 0.00] & [1.00, 1.00] & [0.00, 0.00] & [0.00, 0.00] \\ [0.00, 0.00] & [0.00, 0.00] & [0.00, 0.00] & [1.00, 1.00] & \text{NA} \\ [0.00, 0.00] & [0.00, 0.00] & [0.00, 0.27] & [0.00, 0.00] & [1.00, 1.00] \end{bmatrix} \tag{6.4}$$

Equation 6.5 shows the IEM matrix calculated using the $CEReal$ matrix and the CE^* matrix returned by the LV-IDA+ algorithm for the G model. Based on this, the AIntMSE measure is calculated below.

$$IEM = (\epsilon_{ij}) = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.03 & 0.00 \\ 0.00 & 0.13 & 0.00 & 0.00 & 0.00 \\ 0.05 & 0.11 & 0.02 & 0.00 & 0.00 \\ 0.26 & 0.03 & 0.00 & 0.04 & 0.00 \end{bmatrix} \tag{6.5}$$

$$\text{AIntMSE} = \frac{\Sigma}{p^2 - (p + \text{eNA})} = \frac{0.68}{25 - (5 + 0)} = \frac{.68}{20} = 0.034$$

Equation 6.6 shows the IEM matrix calculated using the $CEReal$ matrix and the CE^* matrix returned by the LV-IDA algorithm for the G model. Note that since the matrix CE^* contains NAs, the matrix IEM also contains NAs. Based on this IEM matrix, the AIntMSE measure is calculated below.

$$IEM = (\epsilon_{ij}) = \begin{bmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.67 \\ 0.00 & 0.00 & \text{NA} & 0.03 & \text{NA} \\ 0.00 & 0.30 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.53 & 0.00 & 0.00 & \text{NA} \\ 0.40 & 0.31 & 0 & 0.25 & 0.00 \end{bmatrix} \tag{6.6}$$

$$\text{AIntMSE} = \frac{\Sigma}{p^2 - (p + \text{eNA})} = \frac{2.49}{25 - (5 + 3)} = \frac{2.49}{17} = 0.146$$

6.2 Data Generation Process.

We evaluate the LV-IDA+ and LV-IDA algorithms using data simulated from synthetic models. We describe next the data generation process for insufficient systems, (see Figure 6.1). First, a random PAG $\mathcal{P}$ is generated. Later, by unfolding the $\mathcal{P}$ the set of MAGs in the MEC $[\mathcal{M}]$ are listed, and we selected randomly one MAG $\mathcal{M}^*$ from $[\mathcal{M}]$. Then, the canonical DAG $\mathcal{D}^*$ is obtained from $\mathcal{M}^*$. Finally, a data set $\boldsymbol{D}'$ is generated simulating the DAG $\mathcal{D}^*$. We parameterized the DAG $\mathcal{D}^*$ with linear Gaussian structural equations, where the coefficients are distributed as ±Uniform([0.5, 1.5]), and the random disturbances according to a normal distribution with mean zero and standard deviations taken from a Uniform([1, 3])). These values in the parameterization of a Gaussian linear DAG are standard in the literature on causal inference and are those used for the experimental evaluation of the LV-IDA algorithm in the work of Malinsky and Spirtes [MS17]. So, to generate each data set $\boldsymbol{D}'$, every vertex X_i in the DAG $\mathcal{D}^*$ is visited in topological order and its value is given by function $X_i = f_i(\boldsymbol{pa}(X_i); e_i)$ where $\boldsymbol{pa}(X_i)$ is the set of parents of the X_i vertex and e_i is an error term that distributes according to a Gaussian distribution with mean zero and standard deviations taken from a Uniform([1, 3])). Every function f_i function is defined as $f_i = w_1 X_1 + \ldots + w_h X_h + e_i$, where h indicates the number of parents of X_i, and the weights $w_1 \ldots w_h$ are sampled from a uniform distribution ±Uniform([0.5, 1.5]). Observations in $\boldsymbol{D}'$ are sampled by evaluating each f_i as many time as needed.

6.3 Experimentation Protocol.

From the data generation process we keep the PAG $\mathcal{P}$, and remove the columns of the variables that were added in the construction of the canonical DAG $\mathcal{D}^*$ over the data set $\boldsymbol{D}'$, to form the data set $\boldsymbol{D}$. This PAG $\mathcal{P}$ and the data set $\boldsymbol{D}$ are used as the input for

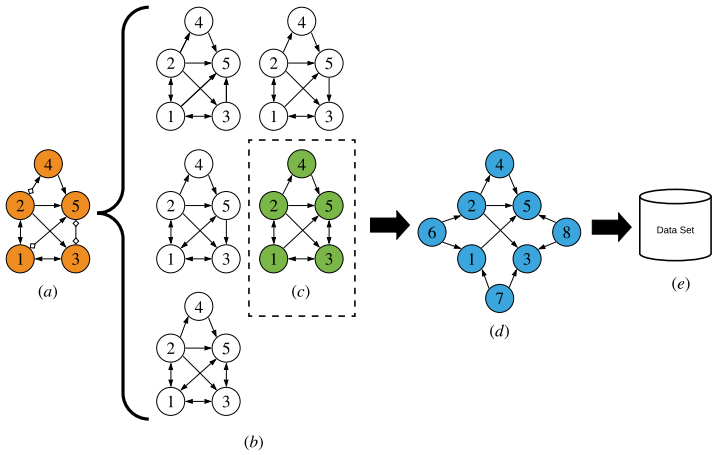


Figure 6.1 The data generation process. (a) A random generated PAG $\mathcal{P}$ (orange). (b) The set of MAGs in the Markov equivalences class $[\mathcal{M}]$. (c) The random selected MAG $\mathcal{M}^*$ (green) from $[\mathcal{M}]$. (d) The canonical DAG $\mathcal{D}^*$ obtained from $\mathcal{M}^*$ (blue). (e) The data set is generated simulating the DAG $\mathcal{D}^*$.

the LV-IDA+ and LV-IDA algorithms. Besides that, we preserve the parameterized DAG $\mathcal{D}^*$ and the complete data set $\boldsymbol{D'}$ to calculate the true causal effects in the *CEReal* matrix. We generate 80 synthetic causal models by varying the number of variables p using $p \in \{5, 8, 11, 14\}$, 20 different synthetic models for each value of p and their respective data sets, generated with samples of size 1000. For a total of 7,360 causal effect estimations, with each of the two algorithms: LV-IDA+ and LV-IDA. We use the LV-IDA implementation on the R programming language, given in [MS17], and we implement the LV-IDA+ algorithm also in the R language. Both implementations of the algorithms were used to calculate the causal effect on all pairs of variables in each of the synthetic models. All experiments were run on a computer with an Intel Core i5 processor running at 2.0 GHz and 8 GB of RAM.

6.4 Results

Table 6.1 summarizes the estimation results (mean ± standard deviation) for the 80 different synthetic data sets, over the evaluation metric AIntMSE. The mean and standard deviation of the number of simple and extreme NA values generated by the algorithms are also exhibited, as the running times per experiment in seconds for the LV-IDA+ and LV-IDA algorithms. The first thing to notice is that the LV-IDA+ algorithm outperforms LV-IDA according to the AIntM evaluation metric for each of the synthetic data sets we

have tested. Which tells us that LV-IDA+ can get higher accuracy in estimating bounds on causal effects in causal insufficient systems.

Since we only consider the non-missing estimated bounds of both algorithms over the AIntM evaluation, we can say that the accuracy on the estimations is the result of the effectiveness in the proposed scheme for the approximation of the missing values. On the other hand, ignoring missing values in the multisets of causal effects (simple NA values) in the LV-IDA algorithm leads to misleading bounds in causal effects. We detected that this was the main reason why LV-IDA sometimes generates intervals of causal effects with greater error than those calculated by LV-IDA+.

Furthermore, with LV-IDA+ is possible to calculate more bounds of causal effects than with LV-IDA. The count for extremes NA values in the results gives us an idea of how many more estimation bounds on causal effects we can calculate for pairs of variables in a model with LV-IDA+.

As we expected, the execution times of the LV-IDA+ algorithm exceed those of LV-IDA, although the increment is relatively low. This is basically because for each MAG in the MEC we add the calculation for the construction of the associated canonical DAG, the search for adjustment sets for this DAG given a par of variables, and the parameter learning process using expectation maximization for the hidden variables.

	AIntMSE	**SimpleNAs**	**ExtremeNAs**	**Time (sec)**
$p = 5$				
LV-IDA+	**0.0223** ± **0.0182**	**0**	**0**	6.6 ± 2.4
LV-IDA	0.0357 ± 0.0413	28.88 ± 16.04	3.50 ± 2.27	**1.4** ± **0.7**
$p = 8$				
LV-IDA+	**0.0441** ± **0.0333**	**0**	**0**	14.4 ± 7.5
LV-IDA	0.0710 ± 0.0611	107.8 ± 28.32	20.3 ± 4.44	**4.6** ± **1.6**
$p = 11$				
LV-IDA+	**0.1040** ± **0.1130**	**0**	**0**	17.0 ± 11.1
LV-IDA	0.1797 ± 0.2133	177.3 ± 78.2	49.4 ± 26.6	**16.4** ± **7.1**
$p = 14$				
LV-IDA+	**0.2353** ± **0.2028**	**0**	**0**	18.4 ± 18.2
LV-IDA	0.4101 ± 0.3605	340.3 ± 98.69	74.1 ± 8.37	**12.2** ± **21.6**

Table 6.1 Performance comparisons between LV-IDA+ and LV-IDA (mean ± standard deviation) over AIntMSE, for $p = 5, 8, 11$ and 14 where p is the number of variables. The best results for each evaluation are highlighted in bold type. The mean and standard deviation of simple and extreme NAs, and the running time in seconds for each algorithm are shown too.

Concluding Remarks

In this chapter we validate our algorithm with simulated data from synthetic models in the same way as it was done for the validation of the LV-IDA algorithm in [MS17]. For this purpose we use practically the same metric as the one used in Malinsky and Spirtes ([MS17]), i.e., the IntMSE metric. The only difference between this and the AIntMSE metric, chosen to evaluate the algorithms, is that AIntMSE considers the average of the IntMSE over all the calculated intervals of causal effect (without taking into account the NAs and the elements on the diagonal of the matrix) in the matrix of intervals returned by the algorithms. In this sense, we are not penalizing the returned NA values in any way but simply do not consider them to calculate this average. The calculation of the real causal effects in made according to Malinsky and Spirtes ([MS17]), i.e., we used the covariate adjustment method over the synthetic DAG that generated the data.

The approach of generating the data from a random PAG is very important to us. This allowed us to evaluate the algorithms LV-IDA+ y LV-IDA without adding noise from the PAGs learning algorithms using the data, such as the FCI (See Chapter 3) or GFCI (See [OSR16]) algorithms. We chose to simulate canonical DAGs to generate the synthetic data, as we decided to test our algorithm where we think it might actually work. However, from the analysis carried out in Chapter 5 we are aware that there are other types of DAGs, such as anti-canonical DAGs, which could be problematic for our algorithm. Therefore, to explore the generality of our algorithm, it is necessary to extend the spectrum of synthetic models in the experimentation, i.e., perform tests with simulated data from, for example, anti-canonical DAGs.

The experiments show a result of great interest to us, which is that with the proposed approximation, the accuracy of the intervals returned by LV-IDA+ was better than the obtained with the ones estimated using LV-IDA. With these experimental results, we are verifying the effectiveness of the parameter learning with hidden variables by mean of expectation-maximization and, on the other hand, we confirm the assumption that the non-estimated effects that cause the missing values, make the bounds returned by LV-IDA to be misleading.

Chapter 7

Conclusions and Future Work

7.1 Summary

In this dissertation we have addressed the problem of computing the causal effect between pairs of variables in the domain of causally insufficient systems, i.e., systems with possible unmeasured common direct causes. For this problem we use the covariate adjustment method over Causal Bayesian Networks where the causal structure is a Maximal Ancestral Graph, considering only the equivalence class of the causal structure of the system, represented as a Partial Ancestral Graph, and assuming that the variables in the system are jointly Gaussian.

While in previous solutions for this problem, the impossibility of estimating bounds on the causal effects between some pairs of variables was contemplated, in this research we contributed with a new algorithm, that we called the LV-IDA+ algorithm, capable of estimating bounds on the causal effects for all pairs of the measured variables in the system, by approximating the causal effect for these special cases in which other algorithms are limited to returning missing values. Knowing that the causal effects between each ordered pair of variables in a system can always be estimated in a DAG, if all the data of the variables are known together with the parameterization of the DAG, our proposal to approximate these causal effects is to calculate them in a DAG that belongs to the set of DAGs represented by the MAG in question.

7.2 Conclusions

The aforementioned proposal to approximate the estimation of causal effects for special cases in which other algorithms are limited to returning missing values, presented the following two major challenges:

1. The first is the fact that the DAGs that belong to the set of DAGs represented by a

MAG are DAGs with latent variables. In this sense, if we want to approximate the estimation of some causal effects of a MAG using a DAG with l latent variables, first we must have a complete parameterization of this DAG and the data for the l latent variables. So, an important question to be answered was whether this problem could be solved by reducing it to the problem of parameter learning with hidden variables. As a first conclusion of this research, we have identified that this part can be solved using parameter learning techniques with hidden variables for directed probabilistic graphical models. In particular, we found that expectation-maximization techniques solves this problem effectively.

2. Having solved the previous challenge, opens the possibility of doing research on the most suitable type of structures to approximate the causal effects in this way. The second challenge with this proposed approximation was to select a DAG with latent variables that could be constructed efficiently, with a small number of latent variables with respect to the number of observed variables and also, being representative for the purpose of estimate causal effects for the entire set of DAGs, represented by the MAG. This last requirement is too ambitious and we do not believe that we have fully resolved it in this dissertation. However we proposed to use the canonical DAG associated to the MAG for this purpose. This type of DAGs has the advantage that it are easy to build from a MAG and generates few latent variables with respect to the number of observed variables, which facilitates the process of parameters learning mentioned in the previous point.

In this research work, we have evaluated the aforementioned approach on the case in which the data were generated by a canonical DAG, which is a particular case over the entire possible spectrum. For this case, we were able to evaluate the quality of the proposed approach by comparing the intervals of causal effect returned by the LV-IDA+ algorithm, which uses the proposed approximation, with those of LV-IDA, using synthetic models. The experiments show a result of great interest to us, which is that with the proposed approximation, the accuracy of the intervals returned by LV-IDA+ was better than the obtained with the ones estimated using LV-IDA. With these experimental results, we are verifying the effectiveness of the parameter learning with hidden variables by mean of expectation-maximization and, on the other hand, we confirm the assumption that the non-estimated effects that cause the missing values, make the bounds returned by LV-IDA

to be misleading. Based on the results of this experimental evaluation, we can conclude that our proposal is useful to establish better bounds in the estimates of the causal effect for insufficient systems modeled with MAGs for the case in which the data are generated from canonical DAGs.

In terms of efficiency, the execution times of the LV-IDA+ algorithm exceed those of LV-IDA although the increment is relatively low. This is basically because for each MAG in the MEC we add the calculation for the construction of the associated canonical DAG, the search for adjustment sets for this DAG given a par of variables, and the parameter learning process using expectation maximization for the hidden variables. Furthermore, we have identified that the true bottleneck in the running time of both LV-IDA and LV-IDA+ is in the PAG unfolding, i.e., the process of listing all MAGs in the MEC, using the ZML algorithm.

7.3 Contribution and Relevance.

The main contribution in this dissertation is the proposal of a method for approximating the causal effect in cases where other works limit themselves to answering that it is not possible to find the causal effect among the pair of variables on the causal system, represented by a MAG. Our proposal is founded on the idea that the causal effects calculated using the adjustment sets on the canonical DAGs associated with the MAGs, in the Markov equivalence class, are good approximations for the causal effects when there are no adjustment sets for some pairs of variables over these MAGs. With this approximation, the proposed LV-IDA+ algorithm can estimate the causal effects among all pairs of variables (or at least approximate the estimate in some cases) in each of the MAGs in the Markov equivalence class and thus can obtain approximate lower and upper bounds of the real value of the causal effect for each pair of system variables.

Estimating the causal effects that variables have on each other, given the causal structure of a system using only observational data, is one of the two main problems in the area of causal inference. The importance of this kind of estimations is that with it we can measure how much a variable Y changes after manipulating another variable X, without the need to carry out controlled experiments. It is well known that the case in which unmeasured common causes are considered in the causal structure, the estimation of causal effects is especially challenging. In this work we contribute to solve this important problem in the area, assuming that the causal structure of the system is only partially known.

The second main problem of causal inference is finding the causal structure given observational data about the system. The most general and widely used algorithms to solve this problem are the constraint-based algorithms. However, these algorithms are only capable of finding the Markov equivalence class of the causal structure given observational data. In the work of Montero-Hernandez et al. in [MOS18] it is shown how to use bounds on the causal effects between the variables of a system, such as those estimated by the LV-IDA+ algorithm, to find the causal structure within a Markov equivalence class returned by a constraint-based algorithm. The result in Montero's et al. works is in the context of systems that assume causal sufficiency. The missing values returned by the LV-IDA algorithm in the estimated bounds on causal effects between the variables of a system do not allow to extend the Montero's et al. algorithm in systems with causal insufficiency. On the other hand, the approximation approach proposed in the LVIDA+ algorithm opens the door to extend the ideas of the Montero's et al. algorithm to systems where the causal sufficiency is relaxed, i.e., more applicable in realistic contexts, which adds relevance to this research.

7.4 Future Work

The experimental results for the LV-IDA+ algorithm are promising but in the future it would be important to extend the experimentation. In section 5.3 it was justified why it is thought that canonical DAGs are a good representation for a large sub-collection of DAGs with latent variables represented by a MAG, for the purpose of estimating causal effects. However, it is necessary to carry out more experimentation with these structures, to see if indeed a canonical DAG is a good representation for them. In the same section, a collection of DAGs with latent variables for which we conjecture they are not well represented by canonical DAGs for the same purpose, which we call anti-canonical DAGs, was also described, but also it is necessary to evaluate with more experiments how much a representation through a canonical DAG fits models with anti-canonical patterns. Therefore, to explore the generality of our algorithm, it is necessary to extend the spectrum of synthetic models in the experimentation, i.e., perform tests with simulated data from, for example, anti-canonical DAGs. Furthermore, in practice, the performance of the LV-IDA+ algorithm will be affected by the precision of the underlying PAG search method and more experiments in which the input PAG is learned from data are needed. For this purpose, it would be interesting to test the LV-IDA+ algorithm using the PAGs returned by the FCI (see Zhang2008) and the GFCI [OSR16]) algorithms. On the other hand, we consider important

to test the algorithm on real world Bayesian Gaussian networks as well.

As future work, we would like to analyze the performance of LV-IDA+ using different optimization techniques, such as the Gradient Descent algorithm, to find maximum values in the likelihood functions, as alternative to expectation maximization in the process of learning the parameters with hidden variables on the canonical DAGs. Extending the work in [MOS18] to insufficient systems, is another direction that we are interested in exploring. In the latter, intervals of causal effects are used to singled out a unique model from the Markov equivalence class on sufficient causal systems. For the aforementioned, it is essential to be able to compute intervals of causal effects for any pairs of variables in the causal system. Therefore, the proposed LV-IDA+ algorithm is fundamental to continue the ideas of this work and bring them to the domain of insufficient causal systems.

Bibliography

[Fis74] R. A. Fisher. *The design of experiments*. New York: Hafner Press, 1974. ISBN: 0028446909.

[SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search, 2nd edition*. Vol. 39. 1. 2000, pp. 137–140.

[RS02] Thomas Richardson and Peter Spirtes. "Ancestral graph Markov models". In: *Ann. Statist.* 30.4 (Aug. 2002), pp. 962–1030.

[ZS05] Jiji Zhang and Peter Spirtes. "A Transformational Characterization of Markov Equivalence for Directed Acyclic Graphs with Latent Variables". In: *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*. Arlington, Virginia: AUAI Press, 2005, pp. 667–674.

[Shi+06] Shohei Shimizu et al. "A Linear Non-Gaussian Acyclic Model for Causal Discovery". In: *J. Mach. Learn. Res.* 7 (Dec. 2006), pp. 2003–2030.

[Zha08a] Jiji Zhang. "Causal Reasoning with Ancestral Graphs". In: *J. Mach. Learn. Res.* 9 (June 2008), pp. 1437–1474.

[Zha08b] Jiji Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". In: *Artificial Intelligence* 172.16-17 (2008), pp. 1873–1896.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN: 0262013193.

[MKB09] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. "Estimating high-dimensional intervention effects from observational data". In: *Ann. Statist.* 37.6A (Dec. 2009), pp. 3133–3164.

[Pea09] Judea Pearl. *Causality: Models, Reasoning and Inference*. 2nd. USA: Cambridge University Press, 2009. ISBN: 052189560X.

[Maa+10] Marloes H. Maathuis et al. “Predicting causal effects in large-scale systems from observational data”. In: *Nature Methods* 7.4 (2010), pp. 247–248.

[MC15] Marloes H. Maathuis and Diego Colombo. “A generalized back-door criterion”. In: *Ann. Statist.* 43.3 (June 2015), pp. 1060–1088.

[Suc15] Luis Enrique Sucar. *Probabilistic Graphical Models: Principles and Applications*. Springer Publishing Company, Incorporated, 2015. ISBN: 1447166981.

[OSR16] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. “A Hybrid Causal Search Algorithm for Latent Variable Models.” In: *JMLR workshop and conference proceedings* 52 (2016), pp. 368–379.

[MS17] Daniel Malinsky and Peter Spirtes. “Estimating bounds on causal effects in high-dimensional and possibly confounded systems”. In: *International Journal of Approximate Reasoning* 88 (2017), pp. 371–384.

[MOS18] Samuel Montero-Hernandez, Felipe Orihuela-Espina, and Luis Enrique Sucar. “Intervals of Causal Effects for Learning Causal Graphical Models”. In: *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*. Vol. 72. Proceedings of Machine Learning Research. Prague, Czech Republic: PMLR, Nov. 2018, pp. 296–307.

[Per+18] Emilija Perković et al. “Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs”. In: *Journal of Machine Learning Research* 18 (2018), pp. 1–62.

[GZS19] Clark Glymour, Kun Zhang, and Peter Spirtes. “Review of Causal Discovery Methods Based on Graphical Models”. In: *Frontiers in Genetics* 10 (2019), p. 524.

[JZB19] Amin Jaber, Jiji Zhang, and Elias Bareinboim. “Causal Identification under Markov Equivalence: Completeness Results”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, Sept. 2019, pp. 2981–2989.

[ZLT19] Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. “Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework”. In: *Artificial Intelligence* 270 (2019), pp. 1–40.