IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2008

Model testing for causal models

Changsung Kang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

 Part of the Computer Sciences Commons

Recommended Citation
Kang, Changsung, "Model testing for causal models" (2008). *Graduate Theses and Dissertations*. 11145.
https://lib.dr.iastate.edu/etd/11145

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

www.manaraa.com

Model testing for causal models

by

Changsung Kang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Computer Science

Program of Study Committee:
Jin Tian, Major Professor
Vasant Honavar
Jack Lutz
Dimitris Margaritis
Alicia Carriquiry

Iowa State University

Ames, Iowa

2008

Copyright © Changsung Kang, 2008. All rights reserved.

www.manaraa.com

To Yoonee and my parents.

www.manaraa.com

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER 1. INTRODUCTION 1

1.1 Linear Structural Equation Models 1

1.2 Causal Bayesian Networks 2

1.3 Thesis Outline 3

CHAPTER 2. RELATED WORK 4

2.1 Linear Causal Models 4

2.2 Polynomial Constraints in Causal Bayesian Networks 5

2.3 Inequality Constraints in Causal Bayesian Networks 6

2.4 Characterizing Interventional Distributions 7

CHAPTER 3. NOTATION AND DEFINITIONS 8

3.1 Linear Causal Models 8

3.2 Causal Bayesian Networks and Interventions 10

3.3 Algebraic Sets, Semi-algebraic Sets and Ideals 12

CHAPTER 4. MARKOV PROPERTIES FOR LINEAR CAUSAL MODELS WITH CORRELATED ERRORS 14

4.1 Preliminaries and Motivation 15

4.1.1 Model Testing and Markov Properties 15

4.1.2 A Local Markov Property for ADMGs 18

4.2 Markov Properties for ADMGs without Directed Mixed Cycles 20

www.manaraa.com

4.2.1 The Reduced Local Markov Property 21
4.2.2 The Ordered Reduced Local Markov Property 25
4.2.3 The Pairwise Markov Property 29
4.2.4 Relation to Other Work 31
4.3 Markov Properties for General ADMGs 33
4.3.1 Reducing the Ordered Local Markov Property 33
4.3.2 An Example 41
4.3.3 Comparison of (LMP,<) and (S-MP,<) 44
CHAPTER 5. POLYNOMIAL CONSTRAINTS IN CAUSAL BAYESIAN NETWORKS . 47
5.1 Problem Statement 48
5.2 Causal Bayesian Network with No Hidden Variables 50
5.2.1 One Interventional Distribution 51
5.2.2 All Interventional Distributions 53
5.2.3 Two Interventional Distributions 53
5.3 Causal Bayesian Network with Hidden Variables 56
5.3.1 Two-step Method 57
5.3.2 Reducing the Implicitization Problem Using Known Constraints 57
5.3.3 Constraints in Subgraphs 62
5.4 Model Testing Using Polynomial Constraints 66
CHAPTER 6. INEQUALITY CONSTRAINTS IN CAUSAL BAYESIAN NETWORKS . 70
6.1 Constraints on Interventional Distributions 70
6.1.1 Inequality Constraints 73
6.2 Inequality Constraints On a Subset of Interventional Distributions 75
6.2.1 Bounds on Causal Effects 79
6.2.2 Inequality Constraints on Nonexperimental Distribution 80
CHAPTER 7. CONCLUSION 81
7.1 Markov Properties for Linear Causal Models with Correlated Errors 81
7.2 Polynomial Constraints in Causal Bayesian Networks 82

www.manaraa.com

7.3 Inequality Constraints in Causal Bayesian Networks 83

BIBLIOGRAPHY 84

ACKNOWLEDGMENTS 89

www.manaraa.com

LIST OF TABLES

Table 5.1 The Type I and Type II errors in testing G_1 against G_2 68

Table 5.2 Comparison of the error rates of two model selection methods 69

www.manaraa.com

LIST OF FIGURES

Figure 2.1 U is a hidden variable. 7

Figure 3.1 Causal diagram illustrating the effect of smoking on lung cancer 10

Figure 4.1 A causal diagram 16

Figure 4.2 An ADMG and its compressed graph 18

Figure 4.3 Directed mixed cycles 21

Figure 4.4 (a) An ADMG with directed mixed cycles (b) Illustration of the procedure **GetOrdering**. The modified graph after the first step is shown. 33

Figure 4.5 The relationship between A and A' that satisfy the conditions in Lemma 2. The induced subgraph G_A is shown. The vertices of G_A are decomposed into two disjoint subsets $\mathrm{de}_{G_A}(T)$ and A'. 35

Figure 4.6 A procedure to generate a reduced set of conditional independence relations for an ADMG G and a consistent ordering $\prec$ 38

Figure 4.7 The c-component $\{V_1, V_2, V_3, V_4\}$ has the root set $\{V_1, V_2\}$ 40

Figure 4.8 A greedy algorithm to generate a good consistent ordering on the vertices of an ADMG G 41

Figure 4.9 An example ADMG for which using (S-MP,$\prec$) is most beneficial. There is no directed mixed cycle and each c-component is a clique joined by bi-directed edges. 46

Figure 5.1 Two causal BNs. 50

Figure 5.2 A procedure for listing polynomial relations among interventional distributions 59

Figure 5.3 Two causal BNs with one hidden variable 60

www.manaraa.com

Figure 5.4 Testing a subgraph that includes the vertices V_1, V_2 and V_3 63

Figure 5.5 A procedure for finding a subgraph in which the local Markov property is satisfied 65

Figure 5.6 Two causal BNs that are Markov equivalent 67

Figure 5.7 A model testing procedure for a causal BN using a polynomial constraint . . . 67

Figure 6.1 U_1,U_2 and U_3 are hidden variables. 72

Figure 6.2 A Procedure for Listing Inequality Constraints On a Subset of Interventional Distributions 77

www.manaraa.com

CHAPTER 1. INTRODUCTION

Finding cause-effect relationships is the central aim of many studies in the physical, behavioral, social and biological sciences. There have been many attempts to theorize about causality. We consider two well-known mathematical causal models: *Structural equation models (SEMs)* and *causal Bayesian networks (BNs)*. When we hypothesize a causal model, that model often imposes constraints on the statistics of the data collected. These constraints enable us to test or falsify the hypothesized causal model. We develop efficient and reliable methods to test a causal model using various types of constraints. For linear SEMs, we investigate the problem of generating a small number of constraints in the form of zero partial correlations, providing an efficient way to test hypothesized models. For causal BNs, we study equality and inequality constraints imposed on data and analyze the structure of the constraints and investigate a way to use these constraints for model testing.

1.1 Linear Structural Equation Models

Linear SEMs are widely used for causal reasoning in social sciences, economics, and artificial intelligence (Goldberger, 1972; Bollen, 1989; Spirtes et al., 2001; Pearl, 2000). One important problem in the applications of linear causal models is testing a hypothesized model against the given data. We seek an efficient method to test linear SEMs with correlated errors. We adopt a local testing method that involves testing for the vanishing partial correlations instead of the conventional method that involves fitting the covariance matrix.

Since conditional independence relations correspond to zero partial correlations, the problem reduces to that of finding a small set of conditional independence relations that imply all other conditional independence relations encoded in an *acyclic directed mixed graph (ADMG)*. Such set of conditional independence relations is called, *local Markov property* for the ADMG. Using a set of axioms that con-

www.manaraa.com

ditional independence relations satisfy, we investigate a way to reduce the local Markov property for ADMGs representing linear SEMs. An additional axiom, called *composition*, which holds for normal distributions, turns out to be a key to reducing the local Markov property.

1.2 Causal Bayesian Networks

In linear SEMs, the causal relationships are expressed in the form of functional equations. In contrast, causal BNs express causal relationships in a stochastic way. We study various types of constraints implied by a causal BN for the purpose of model testing.

First, assuming that we have obtained a collection of interventional distributions by manipulating various sets of variables and observing others, we can ask the following question: it this collection compatible with some underlying causal Bayesian network (even if we do not know its structure)? We show that the interventional distributions are completely characterized by a set of equalities and inequalities. Our result enables us to reject the entire set of models under consideration. The violation of any of these equalities and inequalities leads us to conclude that the underlying model is not semi-Markovian (e.g., there may be feedback loops).

Second, we seek the polynomial equality constraints imposed by a causal BN on both non-experimental and interventional distributions. We propose to use the implicitization procedure to generate polynomial equality constraints. This approach places causal BNs into the realm of algebraic geometry. There are two main challenges in this problem: (i) Computational complexity. (ii) Understanding structures of constraints. To deal with challenge (i), we develop methods to reduce the complexity of the implicitization problem utilizing the structural properties of causal BNs. To deal with challenge (ii), we present some preliminary results on the algebraic structure of the constraints. We also propose a model testing method using polynomial equality constraints.

Third, we study a class of inequality constraints imposed by a causal BN with hidden variables on both non-experimental and interventional distributions. We derive bounds on causal effects in terms of non-experimental distributions and given interventional distributions. We derive instrumental inequality type of constraints upon non-experimental distributions. Although the constraints we give are not complete, they constitute necessary conditions for a hypothesized model to be compatible with the

www.manaraa.com

data. The constraints also provide information (bounds) on the effects of interventions that have not been tried experimentally, from observational data and given experimental data.

1.3 Thesis Outline

This thesis is organized as follows. Chapter 2 discusses related work in linear SEMs and causal BNs. Chapter 3 formally defines causal models. Chapter 4 considers the problem of testing linear SEMs with correlated errors. Chapter 5 considers the problem of efficiently computing polynomial equality constraints in causal BNs. Chapter 6 investigates inequality constraints in causal BNs. Chapter 7 is the conclusion.

www.manaraa.com

CHAPTER 2. RELATED WORK

In this chapter, we overview related work in causal models. We focus on various constraints implied by causal models.

2.1 Linear Causal Models

The conventional method of testing a linear SEM involves maximum likelihood estimation of the covariance matrix. An alternative approach has been proposed recently which involves testing for the conditional independence relationships implied by the model (Spirtes et al., 1998; Pearl, 1998; Pearl and Meshkat, 1999; Pearl, 2000; Shipley, 2000, 2003). The advantages of using this new test method instead of the traditional global fitting test have been discussed in Pearl (1998); Shipley (2000); McDonald (2002); Shipley (2003). The method can be applied in small data samples and it can test "local" features of the model.

To apply this test method, one needs to be able to identify the conditional independence relationships implied by an SEM. This can be achieved by representing the SEM with a graph called a path diagram (Wright, 1934) and then reading independence relations from the path diagram. For a linear SEM without correlated errors, the corresponding path diagram is a directed acyclic graph (DAG). The set of all conditional independence relations holding in any model associated with a DAG, often called a global Markov property for the DAG, can be read by the d-separation criterion (Pearl, 1988). However, it is not necessary to test for all the independencies implied by the model as a subset of those independencies may imply all others. A local Markov property specifies a much smaller set of conditional independence relations which will imply (using the laws of probability) all other conditional independence relations that hold under the global Markov property. A well-known local Markov property for DAGs is that each variable is conditionally independent of its non-descendants given its parents (Lau-

www.manaraa.com

ritzen et al., 1990; Lauritzen, 1996). Based on this local Markov property, Pearl and Meshkat (1999) and Shipley (2000) proposed testing methods for linear SEMs without correlated errors that involve at most one conditional independence test for each pair of variables.

On the other hand, the path diagrams for linear SEMs with correlated errors are DAGs with bi-directed edges ($\leftrightarrow$) where bi-directed edges are used to represent correlated errors. A DAG with bi-directed edges is called an *acyclic directed mixed graph (ADMG)* in Richardson (2003). The set of all conditional independence relations encoded in an ADMG can still be read by (a natural extension of) the d-separation criterion (called m-separation in Richardson, 2003) which provides the global Markov property for ADMGs (Spirtes et al., 1998; Koster, 1999; Richardson, 2003). A local Markov property for ADMGs is given in Richardson (2003), which, in the worst case, may invoke an exponential number of conditional independence relations, a sharp difference with the local Markov property for DAGs, where only one conditional independence relation is associated with each variable. Shipley (2003) suggested a method for testing linear SEMs with correlated errors but the method may or may not, depending on the actual models, be able to find a subset of conditional independence relations that imply all others.

2.2 Polynomial Constraints in Causal Bayesian Networks

There has been much research on identifying constraints on the non-experimental distributions implied by a BN with hidden variables (Verma and Pearl, 1990; Robins and Wasserman, 1997; Desjardins, 1999; Spirtes et al., 2001; Tian and Pearl, 2002b). In algebraic methods, BNs are defined parametrically by a polynomial mapping from a set of parameters to a set of distributions. The distributions compatible with a BN correspond to a *semi-algebraic set*, which can be described with a finite number of polynomial equalities and inequalities. In principle, these polynomial equalities and inequalities can be derived by the quantifier elimination method presented in Geiger and Meek (1999). However, due to high computational demand (doubly exponential in the number of probabilistic parameters), in practice, quantifier elimination is limited to models with few number of probabilistic parameters. Geiger and Meek (1998); Garcia (2004); Garcia et al. (2005) used a procedure called *implicitization* to generate independence and non-independence constraints on the observed non-experimental distributions.

www.manaraa.com

These constraints consist of a set of polynomial equalities that define the smallest *algebraic set* that contains the semi-algebraic set. Garcia et al. (2005) analyzed the algebraic structure of constraints for a class of small BNs.

Algebraic approaches have been applied in causal BNs to deal with the problem of the identifiability of causal effects (Riccomagno and Smith, 2003, 2004). However, to the best of our knowledge, the implicitization method has not been applied to the problem of identifying constraints on interventional distributions induced by causal BNs.

2.3 Inequality Constraints in Causal Bayesian Networks

It is well-known that the observational implications of a BN are completely captured by conditional independence relationships among the variables when all the variables are observed (Pearl et al., 1990). When a BN invokes unobserved variables, called *hidden* or *latent* variables, the network structure may impose other equality and/or inequality constraints on the distribution of the observed variables (Verma and Pearl, 1990; Robins and Wasserman, 1997; Desjardins, 1999; Spirtes et al., 2001). Methods for identifying equality constraints were given in Geiger and Meek (1998); Tian and Pearl (2002b). Pearl (1995) gave an example of inequality constraints in the model shown in Figure 2.1. The model imposes the following inequality, called the *instrumental inequality* by Pearl, for discrete variables X, Y, and Z,

$$\max_x \sum_y \max_z P(xy|z) \leq 1. \tag{2.1}$$

This model has been further analysed using convex analysis approach in Bonet (2001). In principle, all (equality and inequality) constraints implied by BNs with hidden variables can be derived by the quantifier elimination method presented in Geiger and Meek (1999). However, due to high computational demand (doubly exponential in the number of probabilistic parameters), in practice, quantifier elimination is limited to BNs with few number of probabilistic parameters. For example, the current quantifier elimination algorithms cannot deal with the simple model in Figure 2.1 for X, Y, and Z being binary variables.

When all variables are observed, a complete characterization of constraints on interventional distributions imposed by a given causal BN has been given in (Pearl, 2000, pp.23-4). When a causal BN

www.manaraa.com

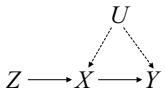


Figure 2.1 U is a hidden variable.

contains unobserved variables, there may be inequality constraints on interventional distributions Tian and Pearl (2002a). For the model in Figure 2.1, bounds on causal effects $P_x(y)$ in terms of the nonexperimental distribution $P(x, y, z)$ was derived in Balke and Pearl (1994); Chickering and Pearl (1996) using linear programming method for X, Y, and Z being binary variables.

2.4 Characterizing Interventional Distributions

Another related problem is the characterization of the interventional distributions generated from a causal Bayesian network of "unknown structure". Assuming that we have obtained a collection of interventional distributions by manipulating various sets of variables and observing others, we can ask the following question: it this collection compatible with *some* underlying causal Bayesian network (even if we do not know its structure)? Tian et al. (2006) showed that the interventional distributions are completely characterized by a set of equalities and inequalities. While the purpose of Kang and Tian (2006, 2007) is to test a single model (with a fixed structure), the result in Tian et al. (2006) enables us to reject the entire set of models under consideration. The violation of any of these equalities and inequalities leads us to conclude that the underlying model is not *semi-Markovian* (e.g., there may be feedback loops).

www.manaraa.com

CHAPTER 3. NOTATION AND DEFINITIONS

In this chapter, we give a formal definition of causal models. Also we introduce some concepts related to algebraic geometry needed to obtain our results.

3.1 Linear Causal Models

The SEM technique was developed by geneticists (Wright, 1934) and economists (Haavelmo, 1943) for assessing cause-effect relationships from a combination of statistical data and qualitative causal assumptions. It is an important causal analysis tool widely used in social sciences, economics, and artificial intelligence (Goldberger, 1972; Duncan, 1975; Bollen, 1989; Spirtes et al., 2001).

In an SEM, the causal relationships among a set of variables are often assumed to be linear and expressed by linear equations. Each equation describes the dependence of one variable in terms of the others. For example, an equation

$$Y = aX + \epsilon \tag{3.1}$$

represents that X may have a *direct* causal influence on Y and that no other variables have (direct) causal influences on Y except those factors (represented by the error term ϵ traditionally assumed to have normal distribution) that are omitted from the model. The parameter a quantifies the (direct) causal effect of X on Y. An equation like (3.1) with a causal interpretation represents an autonomous causal mechanism and is said to be *structural*.

As an example, consider the following model from Pearl (2000) that concerns the relations between

www.manaraa.com

smoking (X) and lung cancer (Y), mediated by the amount of tar (Z) deposited in a person's lungs:

$$X = \epsilon_1$$

$$Z = aX + \epsilon_2$$

$$Y = bZ + \epsilon_3$$

The model assumes that the amount of tar deposited in the lungs depends on the level of smoking (and external factors) and that the production of lung cancer depends on the amount of tar in the lungs but smoking has no effect on lung cancer except as mediated through tar deposits. To fully specify the model, we also need to decide whether those omitted factors (ϵ_1, ϵ_2, ϵ_3) are correlated or not. We may assume that no other factor that affects tar deposit is correlated with the omitted factors that affect smoking or lung cancer ($Cov(\epsilon_1, \epsilon_2) = Cov(\epsilon_2, \epsilon_3) = 0$). However, there might be unobserved factors (say some unknown carcinogenic genotype) that affect both smoking and lung cancer ($Cov(\epsilon_1, \epsilon_3) \neq 0$), but the genotype nevertheless has no effect on the amount of tar in the lungs except indirectly (through smoking). Often, it is illustrative to express our qualitative causal assumptions in terms of a graphical representation, as shown in Figure 3.1.

We now formally define the model that we will consider in this thesis. A *linear causal model* (or *linear SEM*) over a set of random variables $V = \{V_1, \ldots, V_n\}$ is given by a set of structural equations of the form

$$V_j = \sum_i c_{ji} V_i + \epsilon_j, \;\; j = 1, \ldots, n, \tag{3.2}$$

where the summation is over the variables in V judged to be immediate causes of V_j. c_{ji}, called a *path coefficient*, quantifies the direct causal influence of V_i on V_j. ϵ_j's represent "error" terms due to omitted factors and are assumed to have normal distribution. We consider recursive models and assume that the summation in Eq. (3.2) is for $i < j$, that is, $c_{ji} = 0$ for $i \geq j$.

We denote the covariances between observed variables $\sigma_{ij} = Cov(V_i, V_j)$, and between error terms $\psi_{ij} = Cov(\epsilon_i, \epsilon_j)$. We denote the following matrices, $\Sigma = [\sigma_{ij}]$, $\Psi = [\psi_{ij}]$, and $C = [c_{ij}]$. The parameters of the model are the non-zero entries in the matrices C and Ψ. A parameterization of the model assigns a value to each parameter in the model, which then determines a unique covariance matrix Σ given by

www.manaraa.com

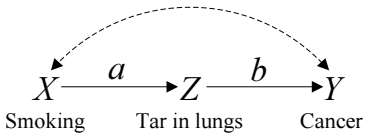


Figure 3.1 Causal diagram illustrating the effect of smoking on lung cancer

(see, for example, Bollen (1989))

$$\Sigma = (I - C)^{-1}\Psi(I - C)^{t-1}. \tag{3.3}$$

The structural assumptions encoded in the model are the zero path coefficients and zero error covariances. The model structure can be represented by a DAG G with (dashed) bi-directed edges (an ADMG), called a *causal diagram* (or *path diagram*), as follows: the nodes of G are the variables $V_1, \ldots, V_n$; there is a directed edge from V_i to V_j in G if V_i appears in the structural equation for V_j, that is, $c_{ji} \neq 0$; there is a bi-directed edge between V_i and V_j if the error terms ϵ_i and ϵ_j have non-zero correlation. For example, the smoking-and-lung-cancer SEM is represented by the causal diagram in Figure 3.1, in which each directed edge is annotated by the corresponding path coefficient.

We note that linear SEMs are often used without explicit causal interpretation. In such cases, linear SEMs can be regarded as an extension of regression models. A linear SEM in which error terms are uncorrelated consists of a set of regression equations. Note that an equation as given by (3.2) is a regression equation if and only if ϵ_j is uncorrelated with each V_i ($Cov(V_i, \epsilon_j) = 0$). Hence, an equation in an SEM with correlated errors may not be a regression equation. Linear SEMs provide a more powerful way to model data than the regression models taking into account correlated error terms.

3.2 Causal Bayesian Networks and Interventions

A causal Bayesian network, also known as a *Markovian model*, consists of two mathematical objects: (i) a DAG G, called a *causal graph*, over a set $V = \{V_1, \ldots, V_n\}$ of vertices, and (ii) a probability distribution $P(v)$, over the set V of discrete variables that correspond to the vertices in G.[1] In this

[1] We only consider discrete random variables in this thesis.

www.manaraa.com

thesis, we will assume a topological ordering $V_1 > \ldots > V_n$ in G. V_1 is always a sink and V_n is always a source. The interpretation of such a graph has two components, probabilistic and causal. The probabilistic interpretation views G as representing conditional independence restrictions on P: Each variable is independent of all its non-descendants given its direct parents in the graph. These restrictions imply that the joint probability function $P(v) = P(v_1, \ldots, v_n)$ factorizes according to the product

$$P(v) = \prod_i P(v_i|pa_i) \tag{3.4}$$

where pa_i are (values of) the parents of variable V_i in G.

The causal interpretation views the arrows in G as representing causal influences between the corresponding variables. In this interpretation, the factorization of (3.4) still holds, but the factors are further assumed to represent autonomous data-generation processes, that is, each conditional probability $P(v_i|pa_i)$ represents a stochastic process by which the values of V_i are assigned in response to the values pa_i (previously chosen for V_i's parents), and the stochastic variation of this assignment is assumed independent of the variations in all other assignments in the model. Moreover, each assignment process remains invariant to possible changes in the assignment processes that govern other variables in the system. This modularity assumption enables us to predict the effects of interventions, whenever interventions are described as specific modifications of some factors in the product of (3.4). The simplest such intervention, called *atomic*, involves fixing a set T of variables to some constants $T = t$, which yields the post-intervention distribution

$$P_t(v) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \tag{3.5}$$

Eq. (3.5) represents a truncated factorization of (3.4), with factors corresponding to the manipulated variables removed. This truncation follows immediately from (3.4) since, assuming modularity, the post-intervention probabilities $P(v_i|pa_i)$ corresponding to variables in T are either 1 or 0, while those corresponding to unmanipulated variables remain unaltered. If T stands for a set of treatment variables and Y for an outcome variable in $V \setminus T$, then Eq. (3.5) permits us to calculate the probability $P_t(y)$ that event $Y = y$ would occur if treatment condition $T = t$ were enforced uniformly over the population.

When some variables in a Markovian model are unobserved, the probability distribution over the observed variables may no longer be decomposed as in Eq. (3.4). Let $V = \{V_1, \ldots, V_n\}$ and $U =$

www.manaraa.com

$\{U_1, \ldots, U_{n'}\}$ stand for the sets of observed and unobserved variables respectively. If no U variable is a descendant of any V variable, then the corresponding model is called a *semi-Markovian model*. We only consider semi-Markovian models. However, the results can be generalized to models with arbitrary unobserved variables as shown in Tian and Pearl (2002b). In a semi-Markovian model, the observed probability distribution, $P(v)$, becomes a mixture of products:

$$P(v) = \sum_u \prod_i P(v_i|pa_i, u^i)P(u) \tag{3.6}$$

where PA_i and U^i stand for the sets of the observed and unobserved parents of V_i, and the summation ranges over all the U variables. The post-intervention distribution, likewise, will be given as a mixture of truncated products

$$P_t(v) = \begin{cases} \sum_u \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i)P(u) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \tag{3.7}$$

Assuming that v is consistent with t, we can write

$$P_t(v) = P_t(v \setminus t) \tag{3.8}$$

In the rest of the thesis, we will use $P_t(v)$ and $P_t(v \setminus t)$ interchangeably, always assuming v being consistent with t.

3.3 Algebraic Sets, Semi-algebraic Sets and Ideals

The set of all polynomials in $x_1, \ldots, x_n$ with real coefficients is called a *polynomial ring* and denoted by $\mathbb{R}[x_1, \ldots, x_n]$. Let $f_1, \ldots, f_s$ be the polynomials in $\mathbb{R}[x_1, \ldots, x_n]$. A *variety* or an *algebraic set* $V(f_1, \ldots, f_s)$ is the set $\{(a_1, \ldots, a_n) \in \mathbb{R}^n : f_i(a_1, \ldots, a_n) = 0 \text{ for all } 1 \leq i \leq s\}$. Thus, an algebraic set is the set of all solutions of a system of polynomial equations.

A subset V of $\mathbb{R}^n$ is called a *semi-algebraic set* if $V = \cup_{i=1}^{s} \cap_{j=1}^{r_i} \{x \in \mathbb{R}^n : P_{i,j}(x) \Leftrightarrow_{ij} 0\}$ where P_{ij} are polynomials in $\mathbb{R}[x_1, \ldots, x_n]$ and $\Leftrightarrow_{ij}$ is one of the comparison operators $\{<, =, >\}$. Informally, a semi-algebraic set is a set that can be described by a finite number of polynomial equalities and inequalities.

A subset $I \subset \mathbb{R}[x_1, \ldots, x_n]$ is called an *ideal* if it satisfies:

www.manaraa.com

(i) $0 \in I$.

(ii) If $f, g \in I$, then $f + g \in I$.

(iii) If $f \in I$ and $h \in \mathbb{R}[x_1, \ldots, x_n]$, then $hf \in I$.

The ideal generated by a set of polynomials $g_1, \ldots, g_n$ is the set of polynomials h that can be written as $h = \sum_{i=1}^{n} f_i g_i$ where f_i are polynomials in the ring and is denoted by $\langle g_1, \ldots, g_n \rangle$. The sum of two ideals I and J is the set $I + J = \{f + g : f \in I, \ g \in J\}$ and it holds that if $I = \langle f_1, \ldots, f_r \rangle$ and $J = \langle g_1, \ldots, g_s \rangle$, then $I + J = \langle f_1, \ldots, f_r, g_1, \ldots, g_s \rangle$. See Cox et al. (1996) for more details.

www.manaraa.com

CHAPTER 4. MARKOV PROPERTIES FOR LINEAR CAUSAL MODELS WITH CORRELATED ERRORS

In this chapter, we seek to improve the local Markov property given in Richardson (2003) for linear SEMs with correlated errors. The local Markov property in Richardson (2003) is applicable for ADMGs associated with arbitrary probability distributions. Specifically, only semi-graphoid axioms which must hold in all probability distributions (Pearl, 1988) are used in showing that the set of conditional independence relations specified by the local Markov property will imply all those specified by the global Markov property. On the other hand, in linear SEMs, variables are assumed to have normal distributions, and it is known that normal distributions also satisfy the so-called composition axiom. Therefore, in this chapter, we look for local Markov properties for ADMGs associated with probability distributions that satisfy the composition axiom. We will show that for a class of ADMGs, the local Markov property will invoke only one conditional independence relation for each variable, and therefore the testing for the corresponding linear SEMs will involve at most one conditional independence test for each pair of variables. For general ADMGs, we provide a procedure that reduces the number of conditional independencies invoked by the local Markov property given in Richardson (2003), and therefore reduces the complexity of testing linear SEMs with correlated errors.

In the test of conditional independence relations, the efficiency of the test is influenced by the size of the conditioning set (that is, the number of conditioning variables) with a small conditioning set having advantage over a large one. The conditional independence relations invoked by the standard local Markov property for DAGs use a parent set as the conditioning set. Pearl and Meshkat (1999) have shown for linear SEMs without correlated errors how to find a set of conditional independence relations that may involve fewer conditioning variables. In this chapter, we also generalize this result to linear SEMs with correlated errors.

www.manaraa.com

The chapter is organized as follows. In Section 4.1, we introduce basic notation and definitions, and present the local Markov property developed in Richardson (2003). In Section 4.2, we show that for a class of ADMGs, there is a local Markov property for probability distributions satisfying the composition axiom that invokes only a linear number of conditional independence relations. We also show a local Markov property that may involve fewer conditioning variables. In Section 4.3, we consider general ADMGs (for probability distributions satisfying the composition axiom) and show a local Markov property that invokes fewer conditional independencies than that in Richardson (2003).

4.1 Preliminaries and Motivation

4.1.1 Model Testing and Markov Properties

One important task in the applications of linear SEMs is to test a model against data. One approach for this task is to test for the conditional independence relationships implied by the model, which can be read from the causal diagram by the d-separation criterion as defined in the following. [1] A *path* between two vertices V_i and V_j in an ADMG consists of a sequence of consecutive edges of any type (directed or bi-directed). A vertex V_i is said to be an *ancestor* of a vertex V_j if there is a path $V_i \to \cdots \to V_j$. A non-endpoint vertex W on a path is called a *collider* if two arrowheads on the path meet at W, i.e. $\to W \leftarrow$, $\leftrightarrow W \leftrightarrow$, $\leftrightarrow W \leftarrow$, $\to W \leftrightarrow$; all other non-endpoint vertices on a path are *non-colliders*, i.e. $\leftarrow W \to$, $\leftarrow W \leftarrow$, $\to W \to$, $\leftrightarrow W \to$, $\leftarrow W \leftrightarrow$. A path between vertices V_i and V_j in an ADMG is said to be *d-connecting given a set* of vertices Z if

1. every non-collider on the path is not in Z, and

2. every collider on the path is an ancestor of a vertex in Z.

If there is no path d-connecting V_i and V_j given Z, then V_i and V_j are said to be *d-separated* given Z. Sets X and Y are said to be *d-separated* given Z, if for every pair V_i, V_j, with $V_i \in X$ and $V_j \in Y$, V_i and V_j are d-separated given Z. Let $I(X, Z, Y)$ denote that X is conditionally independent of Y given Z. The set of all the conditional independence relations encoded by a causal diagram G is specified by the following global Markov property.

[1] The d-separation criterion was originally defined for DAGs (Pearl, 1988) but can be naturally extended for ADMGs and is called m-separation in Richardson (2003).

www.manaraa.com

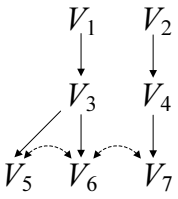


Figure 4.1 A causal diagram

Definition 1 (The Global Markov Property (GMP)) *A probability distribution P is said to satisfy the global Markov property for G if for arbitrary disjoint sets X, Y, Z,*

$$\text{(GMP)} \qquad X \text{ is d-separated from } Y \text{ given } Z \text{ in } G \Longrightarrow I(X, Z, Y). \tag{4.1}$$

The global Markov property typically involves a vast number of conditional independence relations and it is possible to test for a subset of those independencies that will imply all others. A local Markov property specifies a much smaller set of conditional independence relations which will imply by the laws of probability all other conditional independence relations that hold under the global Markov property. For example, a well-known local Markov property for DAGs is that each variable is conditionally independent of its non-descendants given its parents. The causal diagram for a linear SEM with correlated errors is an ADMG and a local Markov property for ADMGs is given in Richardson (2003).

Note that in linear SEMs, the conditional independence relations will correspond to zero partial correlations (Lauritzen, 1996):

$$\rho_{V_i V_j . Z} = 0 \Longleftrightarrow I(\{V_i\}, Z, \{V_j\}). \tag{4.2}$$

As an example, for the linear SEM with the causal diagram in Figure 4.1, if we use the local Markov property in Richardson (2003), then we need to test for the vanishing of the following set of partial correlations (for ease of notation, we write $\rho_{ij.Z}$ to denote $\rho_{V_i V_j . Z}$):

$$\{\rho_{21}, \rho_{32.1}, \rho_{43.2}, \rho_{41.2}, \rho_{54.3}, \rho_{52.3}, \rho_{51.3}, \rho_{64.53}, \rho_{62.53}, \rho_{61.53}, \rho_{64.3}, \rho_{62.3}, \rho_{61.3}, \rho_{72.6543},$$
$$\rho_{71.6543}, \rho_{72.643}, \rho_{71.643}, \rho_{75.4}, \rho_{73.4}, \rho_{72.4}, \rho_{71.4}\}. \tag{4.3}$$

www.manaraa.com

The local Markov property in Richardson (2003) is valid for any probability distributions. In fact, the equivalence of the global and local Markov properties is proved using the following so-called *semi-graphoid axioms* (Pearl, 1988) that probabilistic conditional independencies must satisfy:

- Symmetry

$$I(X, Z, Y) \Longleftrightarrow I(Y, Z, X)$$

- Decomposition

$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z, Y) \ \& \ I(X, Z, W)$$

- Weak Union

$$I(X, Z, Y \cup W) \Longrightarrow I(X, Z \cup W, Y)$$

- Contraction

$$I(X, Z, Y) \ \& \ I(X, Z \cup Y, W) \Longrightarrow I(X, Z, Y \cup W)$$

where X, Y, Z, and W are disjoint sets of variables.

On the other hand, in linear SEMs the variables are assumed to have normal distributions, and normal distributions also satisfy the following *composition* axiom:

- Composition

$$I(X, Z, Y) \ \& \ I(X, Z, W) \Longrightarrow I(X, Z, Y \cup W).$$

Therefore, we expect a local Markov property for linear SEMs to invoke fewer conditional independence relations than that for arbitrary distributions. In this chapter, we will derive reduced local Markov properties for linear SEMs by making use of the composition axiom. As an example, for the linear SEM in Figure 4.1, a local Markov property which we will present in this chapter (see Section 4.2.3) says that we only need to test for the vanishing of the following set of partial correlations:

$$\{\rho_{21}, \rho_{32}, \rho_{43}, \rho_{41}, \rho_{54}, \rho_{52}, \rho_{51.3}, \rho_{64}, \rho_{62}, \rho_{61.3}, \rho_{75}, \rho_{73}, \rho_{71}, \rho_{72.4}\}. \tag{4.4}$$

The number of tests needed and the size of the conditioning set Z are both substantially reduced compared with (4.3), thus leading to a more economical way of testing the given model.

www.manaraa.com

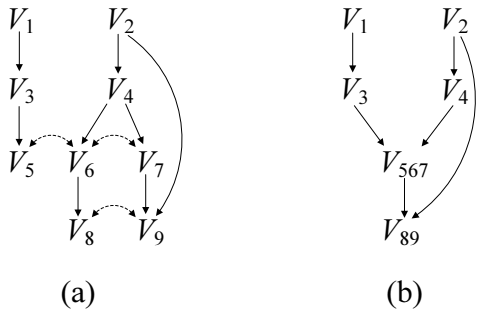


Figure 4.2 An ADMG and its compressed graph

4.1.2 A Local Markov Property for ADMGs

In this section, we describe the local Markov property for ADMGs associated with arbitrary probability distributions presented in Richardson (2003). In this chapter, this Markov property will be used as an important tool to prove the equivalence of our local Markov properties and the global Markov property.

First, we define some graphical notations. For a vertex X in an ADMG G, $\mathrm{pa}_G(X) \equiv \{Y | Y \to X \text{ in } G\}$ is the set of *parents* of X. $\mathrm{sp}_G(X) \equiv \{Y | Y \leftrightarrow X \text{ in } G\}$ is the set of *spouses* of X. $\mathrm{an}_G(X) \equiv \{Y | Y \to \cdots \to X \text{ in } G \text{ or } Y = X\}$ is the set of *ancestors* of X. And $\mathrm{de}_G(X) \equiv \{Y | Y \leftarrow \cdots \leftarrow X \text{ in } G \text{ or } Y = X\}$ is the set of *descendants* of X. These definitions will be applied to sets of vertices, so that, for example, $\mathrm{pa}_G(A) \equiv \cup_{X \in A} \mathrm{pa}_G(X)$, $\mathrm{sp}_G(A) \equiv \cup_{X \in A} \mathrm{sp}_G(X)$, etc.

Definition 2 (C-component) *A c-component of G is a maximal set of vertices in G such that any two vertices in the set are connected by a path on which every edge is of the form ↔; a vertex that is not connected to any bi-directed edge forms a c-component by itself.*

For example, the ADMG in Figure 4.2 (a) is composed of 6 c-components $\{V_1\}$, $\{V_2\}$, $\{V_3\}$, $\{V_4\}$, $\{V_5, V_6, V_7\}$ and $\{V_8, V_9\}$. The *district* of X in G is the c-component of G that includes X. Thus,

$$\mathrm{dis}_G(X) \equiv \{Y | Y \leftrightarrow \cdots \leftrightarrow X \text{ in } G \text{ or } Y = X\}.$$

For example, in Figure 4.2 (a), we have $\mathrm{dis}_G(V_5) = \{V_5, V_6, V_7\}$ and $\mathrm{dis}_G(V_8) = \{V_8, V_9\}$. A set A is said to be *ancestral* if it is closed under the ancestor relation, i.e. if $\mathrm{an}_G(A) = A$. Let G_A denote the induced

www.manaraa.com

subgraph of G on the vertex set A, formed by removing from G all vertices that are not in A, and all edges that do not have both endpoints in A.

Definition 3 (Markov Blanket) [2] *If A is an ancestral set in an ADMG G, and X is a vertex in A that has no children in A then the* Markov blanket of vertex X with respect to the induced subgraph on A, *denoted* $\mathrm{mb}(X, A)$ *is defined to be*

$$\mathrm{mb}(X, A) \equiv \mathrm{pa}_{G_A}\left(\mathrm{dis}_{G_A}(X)\right) \cup \left(\mathrm{dis}_{G_A}(X) \setminus \{X\}\right).$$

For example, for an ancestral set $A = \mathrm{an}_G(\{V_5, V_6\}) = \{V_1, V_2, V_3, V_4, V_5, V_6\}$ in Figure 4.2 (a), we have

$$\mathrm{mb}(V_5, A) = \{V_3, V_4, V_6\}.$$

An ordering ($\prec$) on the vertices of G is said to be consistent with G if $X \prec Y \Rightarrow Y \notin \mathrm{an}_G(X)$. Given a consistent ordering $\prec$, let $\mathrm{pre}_{G,\prec}(X) \equiv \{Y | Y \prec X \text{ or } Y = X\}$.

Definition 4 (The Ordered Local Markov Property (LMP,$\prec$)) *A probability distribution P satisfies the ordered local Markov property for G with respect to a consistent ordering $\prec$, if, for any X and ancestral set A such that $X \in A \subseteq \mathrm{pre}_{G,\prec}(X)$,*

$$\text{(LMP,}\prec\text{)} \qquad I(\{X\}, \mathrm{mb}(X, A), A \setminus (\mathrm{mb}(X, A) \cup \{X\})). \tag{4.5}$$

Theorem 1 (Richardson, 2003) *If G is an ADMG and $\prec$ is a consistent ordering, then a probability distribution P satisfies the ordered local Markov property for G with respect to $\prec$ if and only if P satisfies the global Markov property for G.*

We will write (GMP) $\Longleftrightarrow$ (LMP,$\prec$) to denote the equivalence of the two Markov properties. Therefore the (smaller) set of conditional independencies specified in the ordered local Markov property will imply all other conditional independencies which hold under the global Markov property. It is possible to further reduce the number of conditional independence relations in the ordered local Markov property. An ancestral set A, with $X \in A \subseteq \mathrm{pre}_{G,\prec}(X)$ is said to be *maximal with respect to the Markov blanket* $\mathrm{mb}(X, A)$ if, whenever there is a set B such that $A \subseteq B \subseteq \mathrm{pre}_{G,\prec}(X)$ and $\mathrm{mb}(X, A) = \mathrm{mb}(X, B)$, then $A = B$. For example, suppose that we are given an ordering $\prec$: $V_1 \prec$

[2] *The definition of Markov blanket here follows that in Richardson (2003) and is compatible with that in Pearl (1988).*

www.manaraa.com

$V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$ for the graph G in Figure 4.2 (a). While an ancestral set $A = \text{an}_G(\{V_3, V_6, V_7\}) = \{V_1, V_2, V_3, V_4, V_6, V_7\}$ is maximal with respect to the Markov blanket $\text{mb}(V_7, A) = \{V_4, V_6\}$, an ancestral set $A' = \text{an}_G(\{V_6, V_7\}) = \{V_2, V_4, V_6, V_7\}$ is not. It was shown that we only need to consider ancestral sets A which are maximal with respect to $\text{mb}(X, A)$ in the ordered local Markov property (Richardson, 2003). Thus, we will consider only maximal ancestral sets A when we discuss (LMP,$\prec$) for the rest of this chapter. The following lemma characterizes maximal ancestral sets.

Lemma 1 (Richardson, 2003) *Let X be a vertex and A an ancestral set in G with consistent ordering $\prec$ such that $X \in A \subseteq \text{pre}_{G,\prec}(X)$. The set A is maximal with respect to the Markov blanket* mb*(X,A) if and only if*

$$A = \text{pre}_{G,\prec}(X) \setminus \text{de}_G(\text{h}(X, A))$$

where

$$\text{h}(X, A) \equiv \text{sp}_G\Big(\text{dis}_{G_A}(X)\Big) \setminus \Big(\{X\} \cup \text{mb}(X, A)\Big).$$

Even though we only consider maximal ancestral sets, the ordered local Markov property may still invoke an exponential number of conditional independence relations. For example, for a vertex X, if $\text{dis}_G(X) \subseteq \text{pre}_{G,\prec}(X)$ and $\text{dis}_G(X)$ has a clique of n vertices joined by bi-directed edges, then there are at least $O(2^{n-1})$ different Markov blankets.

It should be noted that only the semi-graphoid axioms were used to prove Theorem 1 on the equivalence of the two Markov properties and no assumptions about probability distributions were made. Next we will show that the ordered local Markov property can be further reduced if we use the composition axiom in addition to the semi-graphoid axioms. The local Markov properties we obtained (in Sections 4.2 and 4.3) are not restricted to linear causal models in that they are actually valid for any probability distributions that satisfy the composition axiom.

4.2 Markov Properties for ADMGs without Directed Mixed Cycles

In this section, we introduce three local Markov properties for a class of ADMGs and show that they are equivalent to the global Markov property. Also, we discuss related work in maximal ancestral graphs and chain graphs. First, we give some definitions.

www.manaraa.com

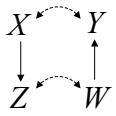


Figure 4.3 Directed mixed cycles

Definition 5 (Directed Mixed Cycle) *A path is said to be a directed mixed path from X to Y if it contains at least one directed edge and every edge on the path is either of the form $Z \leftrightarrow W$, or $Z \rightarrow W$ with W between Z and Y. A directed mixed path from X to Y together with an edge $Y \rightarrow X$ or $Y \leftrightarrow X$ is called a directed mixed cycle.*

For example, the path $X \rightarrow Z \leftrightarrow W \rightarrow Y \leftrightarrow X$ in the graph in Figure 4.3 forms a directed mixed cycle. In this section, we will consider only ADMGs without directed mixed cycles.

Definition 6 (Compressed Graph) *Let G be an ADMG. The compressed graph of G is defined to be the graph $G' = (V', E')$, $V' = \{V_C \mid C \text{ is a c-component of } G\}$, $E' = \{V_{C_i} \rightarrow V_{C_j} \mid \text{there is an edge } X \rightarrow Y \text{ in } G \text{ such that } X \in C_i, Y \in C_j\}$.*

Figure 4.2 shows an ADMG and its compressed graph. If there exists a directed mixed cycle in an ADMG G, there will be a cycle or a self-loop in the compressed graph of G. For example, if for two vertices X and Y in a c-component C of G there exists an edge $X \rightarrow Y$, then the compressed graph of G contains a self-loop $\circlearrowright_{V_C}$. The following proposition holds.

Proposition 1 *Let G be an ADMG. The compressed graph of G is a DAG if and only if G has no directed mixed cycles.*

4.2.1 The Reduced Local Markov Property

In this section, we introduce a local Markov property for ADMGs without directed mixed cycles which only invokes a linear number of conditional independence relations and show that it is equivalent to the global local Markov property.

www.manaraa.com

Definition 7 (The Reduced Local Markov Property (RLMP)) *Let G be an ADMG without directed mixed cycles. A probability distribution P is said to satisfy the reduced local Markov property for G if*

$$\text{(RLMP)} \qquad \forall X \in V, \quad I(\{X\}, \mathrm{pa}_G(X), V \setminus \mathrm{f}(X, G)) \tag{4.6}$$

where $\mathrm{f}(X, G) \equiv \mathrm{pa}_G(X) \cup \mathrm{de}_G(\{X\} \cup \mathrm{sp}_G(X))$.

The reduced local Markov property states that *a variable is independent of the variables that are neither its descendants nor its spouses' descendants given its parents*.

Theorem 2 *If a probability distribution P satisfies the composition axiom and an ADMG G has no directed mixed cycles, then*

$$\text{(GMP)} \Longleftrightarrow \text{(RLMP)}. \tag{4.7}$$

Proof: (GMP) $\Longrightarrow$ (RLMP)

We need to prove that any variable X is d-separated from $V \setminus \mathrm{f}(X, G)$ given $\mathrm{pa}_G(X)$ in G with no directed mixed cycle. Consider a vertex $\alpha \in V \setminus \mathrm{f}(X, G)$. We will show that there is no path d-connecting X and α given $\mathrm{pa}_G(X)$. There are four possible cases for any path between X and α.

1. $X \leftarrow \beta \cdots \alpha$
2. $X \rightarrow \cdots \rightarrow \delta \leftarrow * \cdots \alpha$
3. $X \leftrightarrow \gamma \leftarrow * \cdots \alpha$
4. $X \leftrightarrow \gamma \rightarrow \cdots \rightarrow \delta \leftarrow * \cdots \alpha$

A symbol $*$ serves as a wildcard for an end of an edge. For example, $\leftarrow *$ represents both $\leftarrow$ and $\leftrightarrow$. In case 1, $\beta \in \mathrm{pa}_G(X)$. In case 2, the collider δ is not an ancestor of a vertex in $\mathrm{pa}_G(X)$ (otherwise, there would be a cycle). In cases 3 and 4, neither γ nor δ is an ancestor of a vertex in $\mathrm{pa}_G(X)$ (otherwise, there would be directed mixed cycles). In any case, the path is not d-connecting. ■

Proof: (RLMP) $\Longrightarrow$ (GMP)

www.manaraa.com

We will show that for some consistent ordering $\prec$, (RLMP) $\Longrightarrow$ (LMP, $\prec$). Then, by Theorem 1, we have (RLMP) $\Longrightarrow$ (GMP).

We construct a consistent ordering with the desired property as follows.

1. Construct the compressed graph G' of G.
2. Let $\prec'$ be any consistent ordering on G'. Construct a consistent ordering $\prec$ from $\prec'$ by replacing each V_C (corresponding to each c-component C of G) in $\prec'$ with the vertices in C (the ordering of the vertices in C is arbitrary).

We now prove that (RLMP) $\Longrightarrow$ (LMP,$\prec$). Assume that a probability distribution P satisfies (RLMP). Consider the set of conditional independence relations invoked by (LMP,$\prec$) for each variable X given in (4.5). First, observe that for any vertex Y in $\mathrm{dis}_{G_A}(X)$, we have

$$A \setminus (\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_G(Y)) \subseteq V \setminus \mathrm{f}(Y, G), \tag{4.8}$$

since

$$\begin{aligned} & A \setminus (\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_G(Y)) \\ & = A \setminus \Big(\big(\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_G(Y)\big) \cup \big(\mathrm{de}_G(\{Y\} \cup \mathrm{sp}_G(Y)) \setminus (\{Y\} \cup \mathrm{sp}_G(Y))\big) \Big) \\ & = A \setminus \mathrm{f}(Y, G). \end{aligned} \tag{4.9}$$

The equality (4.9) holds since the vertices in $\mathrm{de}_G(\{Y\} \cup \mathrm{sp}_G(Y)) \setminus (\{Y\} \cup \mathrm{sp}_G(Y))$ do not appear in A (because of the way $\prec$ is constructed, no descendant of $\mathrm{dis}_{G_A}(X)$ is in A). Thus, by (4.6), for all Y in $\mathrm{dis}_{G_A}(X)$, we have

$$I(\{Y\}, \mathrm{pa}_G(Y), A \setminus (\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_{G_A}(Y))). \tag{4.10}$$

Let $S_1 = \mathrm{pa}_G(\mathrm{dis}_{G_A}(X)) \setminus \mathrm{pa}_G(Y)$ and $S_2 = A \setminus (\mathrm{mb}(X, A) \cup \{X\})$. It follows that

$$S_1 \subseteq A \setminus (\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_G(Y)) \text{ and} \tag{4.11}$$

$$S_2 \subseteq A \setminus (\mathrm{pa}_G(Y) \cup \{Y\} \cup \mathrm{sp}_G(Y)). \tag{4.12}$$

Also, we have

$$S_1 \cap S_2 = \emptyset, \tag{4.13}$$

www.manaraa.com

since $S_1 \subseteq \mathrm{mb}(X, A)$. Therefore,

$$I(\{Y\}, \mathrm{pa}_G(Y), S_1 \cup S_2) \quad \text{by decomposition} \tag{4.14}$$

$$I(\{Y\}, \mathrm{pa}_G(Y) \cup S_1, S_2) \quad \text{by weak union} \tag{4.15}$$

$$I(\mathrm{dis}_{G_A}(X), \mathrm{pa}_G(\mathrm{dis}_{G_A}(X)), A \setminus (\mathrm{mb}(X, A) \cup \{X\})) \quad \text{by composition} \tag{4.16}$$

$$I(\{X\}, \mathrm{pa}_G(\mathrm{dis}_{G_A}(X)) \cup (\mathrm{dis}_{G_A}(X) \setminus \{X\}),$$

$$A \setminus (\mathrm{mb}(X, A) \cup \{X\})) \quad \text{by weak union.} \tag{4.17}$$

Thus, by the definition of the Markov blanket of X with respect to A, we have

$$I(\{X\}, \mathrm{mb}(X, A), A \setminus (\mathrm{mb}(X, A) \cup \{X\})). \tag{4.18}$$

■

As an example, consider the ADMG G in Figure 4.2 (a) which has no directed mixed cycles. The graph in Figure 4.2 (b) is the compressed graph G' of G described in the proof. From the ordering $\prec'$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_{567} \prec V_{89}$, we obtain the ordering $\prec$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$. The ordered local Markov property (LMP,$\prec$) involves the following conditional independence relations:

$$I(\{V_2\}, \emptyset, \{V_1\}), \qquad I(\{V_3\}, \{V_1\}, \{V_2\}),$$

$$I(\{V_4\}, \{V_2\}, \{V_1, V_3\}), \qquad I(\{V_5\}, \{V_3\}, \{V_1, V_2, V_4\}),$$

$$I(\{V_6\}, \{V_3, V_4, V_5\}, \{V_1, V_2\}), \qquad I(\{V_6\}, \{V_4\}, \{V_1, V_2, V_3\}),$$

$$I(\{V_7\}, \{V_3, V_4, V_5, V_6\}, \{V_1, V_2\}), \qquad I(\{V_7\}, \{V_4, V_6\}, \{V_1, V_2, V_3\}),$$

$$I(\{V_7\}, \{V_4\}, \{V_1, V_2, V_3, V_5\}), \qquad I(\{V_8\}, \{V_6\}, \{V_1, V_2, V_3, V_4, V_5, V_7\}),$$

$$I(\{V_9\}, \{V_2, V_6, V_7, V_8\}, \{V_1, V_3, V_4, V_5\}), \qquad I(\{V_9\}, \{V_2, V_7\}, \{V_1, V_3, V_4, V_5, V_6\}). \tag{4.19}$$

www.manaraa.com

(RLMP) invokes the following conditional independence relations:

$$
\begin{aligned}
&I(\{V_1\}, \emptyset, \{V_2, V_4, V_6, V_7, V_8, V_9\}), && I(\{V_2\}, \emptyset, \{V_1, V_3, V_5\}),\\
&I(\{V_3\}, \{V_1\}, \{V_2, V_4, V_6, V_7, V_8, V_9\}), && I(\{V_4\}, \{V_2\}, \{V_1, V_3, V_5\}),\\
&I(\{V_5\}, \{V_3\}, \{V_1, V_2, V_4, V_7, V_9\}), && I(\{V_6\}, \{V_4\}, \{V_1, V_2, V_3\}),\\
&I(\{V_7\}, \{V_4\}, \{V_1, V_2, V_3, V_5\}), && I(\{V_8\}, \{V_6\}, \{V_1, V_2, V_3, V_4, V_5, V_7\}),\\
&I(\{V_9\}, \{V_2, V_7\}, \{V_1, V_3, V_4, V_5, V_6\}) &&
\end{aligned} \tag{4.20}
$$

which, by Theorem 2, imply all the conditional independence relations in (4.19).

For the special case of graphs containing only bi-directed edges,[3] Kauermann (1996) provides a local Markov property for probability distributions obeying the composition axiom as follows:

$$\forall X \in V, \quad I(\{X\}, \emptyset, V \setminus (\{X\} \cup \mathrm{sp}_G(X))). \tag{4.21}$$

Since a graph containing only bi-directed edges is a special case of ADMGs without directed mixed cycles, the reduced local Markov property (RLMP) is applicable, and it turns out that (RLMP) reduces to (4.21) for graphs containing only bi-directed edges. Therefore (RLMP) includes the local Markov property given in Kauermann (1996) as a special case.

4.2.2 The Ordered Reduced Local Markov Property

The set of zero partial correlations corresponding to a conditional independence relation $I(X, Z, Y)$ is

$$\{\rho_{V_i V_j . Z} = 0 \mid V_i \in X, V_j \in Y\}. \tag{4.22}$$

Although (RLMP) gives only a linear number of conditional independence relations, the number of zero partial correlations may be larger than that invoked by (LMP,$\prec$) in some cases. For example, 12 conditional independence relations in (4.19) involve 37 zero partial correlations while 9 conditional independence relations in (4.20) involve 41 zero partial correlations. In this section, we will show an ordered local Markov property such that at most one zero partial correlation is invoked for each pair of variables.

[3] Kauermann (1996) actually used undirected graphs with dashed edges which are Markov equivalent to graphs with only bi-directed edges (see Richardson, 2003, for discussions).

www.manaraa.com

Definition 8 (C-ordering) *Let G be an ADMG. A consistent ordering $\prec$ on the vertices of G is said to be a c-ordering if all the vertices in each c-component of G are continuously ordered in $\prec$.*

For example, the ordering $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$ is a c-ordering on the vertices of G in Figure 4.2 (a). The following holds.

Proposition 2 *There exists a c-ordering on the vertices of G if G does not have directed mixed cycles.*

We can easily construct a c-ordering from the compressed graph of G. We introduce the following Markov property.

Definition 9 (The Ordered Reduced Local Markov Property (RLMP,$\prec_c$)) *Let G be an ADMG without directed mixed cycles and $\prec_c$ be a c-ordering on the vertices of G. A probability distribution P is said to satisfy the ordered reduced local Markov property for G with respect to $\prec_c$ if*

$$\text{(RLMP,}\prec_c\text{)} \qquad \forall X \in V,\ I(\{X\}, \mathrm{pa}_G(X), \mathrm{pre}_{G,\prec_c}(X) \setminus (\{X\} \cup \mathrm{pa}_G(X) \cup \mathrm{sp}_G(X))). \tag{4.23}$$

The ordered reduced local Markov property states that *a variable is independent of its predecessors, excluding its spouses, in a c-ordering given its parents*. We now establish the equivalence of (GMP) and (RLMP,$\prec_c$).

Theorem 3 *If a probability distribution P satisfies the composition axiom and an ADMG G has no directed mixed cycles, then for a c-ordering $\prec_c$ on the vertices of G,*

$$\text{(GMP)} \Longleftrightarrow \text{(RLMP,}\prec_c\text{)}. \tag{4.24}$$

Proof: (GMP) $\Longrightarrow$ (RLMP,$\prec_c$)

The set $\mathrm{pre}_{G,\prec_c}(X)$ does not include any descendant of $\mathrm{dis}_G(X)$ since $\prec_c$ is a c-ordering. We have

$$\begin{aligned}
&\mathrm{pre}_{G,\prec_c}(X) \setminus (\{X\} \cup \mathrm{pa}_G(X) \cup \mathrm{sp}_G(X)) \\
&= \mathrm{pre}_{G,\prec_c}(X) \setminus \Big(\big(\{X\} \cup \mathrm{pa}_G(X) \cup \mathrm{sp}_G(X)\big) \cup \big(\mathrm{de}_G(\{X\} \cup \mathrm{sp}_G(X)) \setminus (\{X\} \cup \mathrm{sp}_G(X))\big)\Big) \\
&= \mathrm{pre}_{G,\prec_c}(X) \setminus \mathrm{f}(X, G) \\
&\subseteq V \setminus \mathrm{f}(X, G).
\end{aligned} \tag{4.25}$$

www.manaraa.com

Hence, (RLMP,$\prec_c$) follows from (RLMP). ■

Proof: (RLMP,$\prec_c$) $\Longrightarrow$ (GMP)

We will show that (RLMP,$\prec_c$) $\Longrightarrow$ (LMP,$\prec_c$). Assume that a probability distribution P satisfies (RLMP,$\prec_c$). Let $\mathrm{g}(Y) = \mathrm{pre}_{G,\prec_c}(Y) \setminus (\{Y\} \cup \mathrm{pa}_G(Y) \cup \mathrm{sp}_G(Y)$. Consider the set of conditional independence relations invoked by (LMP,$\prec_c$) for each variable X given in (4.5). By (4.23), for all Y in $\mathrm{dis}_{G_A}(X)$, we have

$$I(Y, \mathrm{pa}_G(Y), \mathrm{g}(Y)). \tag{4.26}$$

Let $S_1 = \mathrm{pa}_G(\mathrm{dis}_{G_A}(X)) \setminus \mathrm{pa}_G(Y)$ and $S_2 = A \setminus (\mathrm{mb}(X, A) \cup \{X\})$. We have that

$$S_1 \subseteq \mathrm{g}(Y). \tag{4.27}$$

Note that $S_2 \setminus \mathrm{g}(Y)$ may be non-empty. Let $S_3 = S_2 \setminus \mathrm{g}(Y)$. It suffices to show that

$$I(Y, \mathrm{pa}_G(Y), S_3), \tag{4.28}$$

which implies $I(Y, \mathrm{pa}_G(Y), S_2)$. Then, the rest of the proof would be identical to that of Theorem 2.

We first characterize the vertices in S_3. We will show that

$$S_3 = (\mathrm{pre}_{G,\prec_c}(X) \setminus \mathrm{pre}_{G,\prec_c}(Y)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X)). \tag{4.29}$$

By Lemma 1, we have

$$S_2 = \mathrm{pre}_{G,\prec_c}(X) \setminus \Big(\mathrm{de}_G(\mathrm{h}(X, A)) \cup \mathrm{mb}(X, A) \cup \{X\}\Big). \tag{4.30}$$

Since $\prec_c$ is a c-ordering, no descendant of $\mathrm{dis}_G(X)$ will appear in A. Hence,

$$S_2 = \mathrm{pre}_{G,\prec_c}(X) \setminus \Big(\mathrm{sp}_G(\mathrm{dis}_{G_A}(X)) \cup \mathrm{pa}_G(\mathrm{dis}_{G_A}(X))\Big). \tag{4.31}$$

To identify some common elements of S_2 and $\mathrm{g}(Y)$, we will reformulate S_2 and $\mathrm{g}(Y)$ as follows.

$$S_2 = \Big(B \setminus \mathrm{pa}_G(\mathrm{dis}_{G_A}(X))\Big) \cup \Big((\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(X)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X))\Big) \tag{4.32}$$

$$\mathrm{g}(Y) = \Big(B \setminus \mathrm{pa}_G(Y)\Big) \cup \Big((\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(Y)) \setminus (\{Y\} \cup \mathrm{sp}_G(Y))\Big) \tag{4.33}$$

www.manaraa.com

where $B = \mathrm{pre}_{G,\prec_c}(X) \setminus \mathrm{dis}_G(X)$. This can be verified by noting that $A_1 = A_2 \setminus (A_3 \cup A_4) = (A_{11} \setminus A_2) \cup (A_{12} \setminus A_3)$ if $A_1 = A_{11} \cup A_{12}, A_{11} \cap A_{12} = \emptyset, A_2 \subseteq A_{11}, A_3 \subseteq A_{12}$. From $\mathrm{pa}_G(Y) \subseteq \mathrm{pa}_G(\mathrm{dis}_{G_A}(X))$, it follows that $B \setminus \mathrm{pa}_G(\mathrm{dis}_{G_A}(X)) \subseteq B \setminus \mathrm{pa}_G(Y)$ and

$$\begin{aligned} S_3 =& S_2 \setminus \mathrm{g}(Y) \\ =& \Big((\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(X)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X))\Big) \\ & \setminus \Big((\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(Y)) \setminus (\{Y\} \cup \mathrm{sp}_G(Y))\Big). \end{aligned} \tag{4.34}$$

We can rewrite the first part of this expression as follows.

$$\begin{aligned} & (\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(X)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X)) \\ &= \Big((\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(Y)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X))\Big) \\ &\cup \Big((\mathrm{pre}_{G,\prec_c}(X) \setminus \mathrm{pre}_{G,\prec_c}(Y)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X))\Big) \end{aligned} \tag{4.35}$$

From $(\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(Y)) \setminus \mathrm{sp}_G(\mathrm{dis}_{G_A}(X)) \subseteq (\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec_c}(Y)) \setminus (\{Y\} \cup \mathrm{sp}_G(Y))$, (4.29) follows. Thus, the vertices in S_3 are those in the set $\mathrm{pre}_{G,\prec_c}(X) \setminus \mathrm{pre}_{G,\prec_c}(Y)$ and not in the set $\mathrm{sp}_G(\mathrm{dis}_{G_A}(X))$.

Now we are ready to prove $I(Y, \mathrm{pa}_G(Y), S_3)$. For any $Z \in S_3$, we have $Y \prec Z$ and $Z \notin \mathrm{sp}_G(Y)$. Hence,

$$I(\{Z\}, \mathrm{pa}_G(Z), \mathrm{g}(Z)) \tag{4.36}$$

$$I(\{Z\}, \mathrm{pa}_G(Z), \{Y\} \cup (\mathrm{pa}_G(Y) \setminus \mathrm{pa}_G(Z))) \quad \text{by decomposition} \tag{4.37}$$

$$I(\{Z\}, \mathrm{pa}_G(Z) \cup \mathrm{pa}_G(Y), \{Y\}) \quad \text{by weak union} \tag{4.38}$$

$$I(\{Y\}, \mathrm{pa}_G(Y), \mathrm{pa}_G(Z) \setminus \mathrm{pa}_G(Y)) \tag{4.39}$$

$$I(\{Y\}, \mathrm{pa}_G(Y), \{Z\}) \quad \text{by contraction.} \tag{4.40}$$

Therefore, by composition, $I(Y, \mathrm{pa}_G(Y), S_3)$ holds. ■

(RLMP,$\prec_c$) invokes one zero partial correlation for each pair of nonadjacent variables. For example, for the ADMG G in Figure 4.2 (a) and a c-ordering $\prec_c$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$,

www.manaraa.com

(RLMP,$\prec_c$) invokes the following conditional independence relations:

$$\begin{aligned}
&I(\{V_2\},\emptyset,\{V_1\}), && I(\{V_3\},\{V_1\},\{V_2\}),\\
&I(\{V_4\},\{V_2\},\{V_1,V_3\}), && I(\{V_5\},\{V_3\},\{V_1,V_2,V_4\}),\\
&I(\{V_6\},\{V_4\},\{V_1,V_2,V_3\}), && I(\{V_7\},\{V_4\},\{V_1,V_2,V_3,V_5\}),\\
&I(\{V_8\},\{V_6\},\{V_1,V_2,V_3,V_4,V_5,V_7\}), && I(\{V_9\},\{V_2,V_7\},\{V_1,V_3,V_4,V_5,V_6\})
\end{aligned} \tag{4.41}$$

which involve 25 zero partial correlations while (4.19) involve 37 zero partial correlations.

4.2.3 The Pairwise Markov Property

In this section, we give a pairwise Markov property which specifies conditional independence relations between pairs of variables and show that it is equivalent to the global Markov property. In previous sections, we focused on minimizing the number of zero partial correlations. We now take into account the size of the conditioning set Z in each zero partial correlation ρ_{XYZ}. When the size of $\mathrm{pa}_G(X)$ for a vertex X in (RLMP,$\prec_c$) is large, it might be advantageous to use a different conditioning set with smaller size (if the equivalence of the Markov properties still holds). Pearl and Meshkat (1999) introduced a pairwise Markov property for DAGs (without bi-directed edges) which may involve fewer conditioning variables and thus lead to more economical tests. The result can be easily generalized to ADMGs with no directed mixed cycles.

Let $d(X,Y)$ denote the shortest distance between two vertices X and Y, that is, the number of edges in the shortest path between X and Y. Two vertices X and Y are nonadjacent if X and Y are not connected by a directed nor a bi-directed edge.

Definition 10 (The Pairwise Markov Property (PMP,$\prec_c$)) *Let G be an ADMG without directed mixed cycles and $\prec_c$ be a c-ordering on the vertices of G. A probability distribution P is said to satisfy the pairwise Markov property for G with respect to $\prec_c$ if for any two nonadjacent vertices V_i, V_j, $V_j \prec_c V_i$*

$$(\mathrm{PMP},\prec_c) \qquad I(\{V_i\},Z_{ij},\{V_j\}) \tag{4.42}$$

where Z_{ij} is any set of vertices such that Z_{ij} d-separates V_i from V_j and $\forall Z \in Z_{ij}, d(V_i,Z) < d(V_i,V_j)$.

www.manaraa.com

Note that, in ADMGs with no directed mixed cycles, there always exists such a Z_{ij} for any two non-adjacent vertices. For example, the parent set of V_i always satisfies the condition for Z_{ij}. If the empty set d-separates V_i from V_j, then the empty set is defined to satisfy the condition for Z_{ij}. Therefore we can always choose a Z_{ij} with the smallest size, providing a more economical way to test zero partial correlations.

Theorem 4 *If a probability distribution P satisfies the composition axiom and an ADMG G has no directed mixed cycles, then*

$$\text{(GMP)} \Longleftrightarrow \text{(PMP,}\prec_c\text{)}. \tag{4.43}$$

Proof: Noting that two vertices X and Y are adjacent if $X \leftarrow Y$, $X \rightarrow Y$ or $X \leftrightarrow Y$, the proof of Theorem 1 by Pearl and Meshkat (1999) is directly applicable to ADMGs and it effectively proves that (RLMP,$\prec_c$) $\Longleftrightarrow$ (PMP,$\prec_c$). We will not reproduce the proof here. ■

As an example, for the ADMG G in Figure 4.2 (a) and a c-ordering $\prec_c$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$, the following conditional independence relations (for convenience, we combined the relations for each vertex that have the same conditioning set) can be given by (PMP,$\prec_c$):

$$\begin{aligned}
&I(\{V_2\}, \emptyset, \{V_1\}), && I(\{V_3\}, \emptyset, \{V_2\}),\\
&I(\{V_4\}, \emptyset, \{V_3, V_1\}), && I(\{V_5\}, \emptyset, \{V_4, V_2\}),\\
&I(\{V_5\}, \{V_3\}, \{V_1\}), && I(\{V_6\}, \emptyset, \{V_3, V_1\}),\\
&I(\{V_6\}, \{V_4\}, \{V_2\}), && I(\{V_7\}, \emptyset, \{V_5, V_3, V_1\}),\\
&I(\{V_7\}, \{V_4\}, \{V_2\}), && I(\{V_8\}, \{V_6\}, \{V_7, V_5, V_4, V_2\}),\\
&I(\{V_8\}, \emptyset, \{V_3, V_1\}), && I(\{V_9\}, \{V_2, V_7\}, \{V_6, V_4\}),\\
&I(\{V_9\}, \emptyset, \{V_5, V_3, V_1\}) &&
\end{aligned} \tag{4.44}$$

which involve the same number of zero partial correlations as (4.41) but involve smaller conditioning sets than those in (4.41).

www.manaraa.com

4.2.4 Relation to Other Work

In this section, we contrast the class of ADMGs without directed mixed cycles to maximal ancestral graphs and chain graphs in terms of Markov properties.

4.2.4.1 Maximal Ancestral Graphs

It is easy to see that an ADMG without directed mixed cycles is a *maximal ancestral graph (MAG)* (Richardson and Spirtes, 2002). An ADMG is said to be *ancestral* if, for any edge $X \leftrightarrow Y$, X is not an ancestor of Y (and vice versa). Note that an edge $X \leftrightarrow Y$ and a directed path from X to Y (or Y to X) form a directed mixed cycle. Hence, an ADMG without directed mixed cycles is ancestral. An ancestral graph is said to be *maximal* if, for any pair of nonadjacent vertices X and Y, there exists a set $Z \subseteq V \setminus \{X, Y\}$ that d-separates X from Y. From Theorem 4, it follows that an ADMG without directed mixed cycles is maximal. On the other hand, there exist MAGs which have directed mixed cycles (see Figure 4.3). Thus, the class of ADMGs without directed mixed cycles is a strict subclass of MAGs.

Richardson and Spirtes (2002) (pp.979) showed the following pairwise Markov property for a MAG G:

$$I(\{V_i\}, \mathrm{an}_G(\{V_i, V_j\}) \setminus \{V_i, V_j\}, \{V_j\})$$

for any two nonadjacent vertices V_i and V_j. Richardson and Spirtes (2002) proved that this pairwise Markov property implies the global Markov property assuming a Gaussian parametrization. This does not trivially imply our results in Section 4.2.3 and our results cannot be considered as a special case of the results on MAGs. The two pairwise Markov properties involve two different forms of conditioning sets. The pairwise Markov property for MAGs involves considerably larger conditioning sets than our pairwise Markov property: the conditioning set includes all ancestors of V_i and V_j, which is undesirable for our purpose of using the zero partial correlations to test a model.

Also, it should be stressed that our results do not depend on a specific parameterization. We only require the composition axiom to be satisfied. In contrast, Richardson and Spirtes (2002) consider only Gaussian parameterizations. It requires further study whether the pairwise Markov property for MAGs can be generalized to the class of distributions satisfying the composition axiom.

www.manaraa.com

In the next section, we consider general ADMGs and try to eliminate redundant conditional independence relations from (LMP,$\prec$). The class of MAGs is clearly a (strict) subclass of ADMGs. Hence, given a MAG, we have two options: either we use the result in the next section or the pairwise Markov property for MAGs. Although the pairwise Markov property for MAGs gives fewer zero partial correlations (one for each nonadjacent pair of vertices), it is possible that in some cases we are better off using the result in the next section (because of the cost incurred by the large conditioning sets in the pairwise Markov property for MAGs). An example of this situation will be given in the next section.

Richardson and Spirtes (2002) also proved that for a Gaussian distribution encoded by a MAG all the constraints on the distribution (that is, on the covariance matrix) are implied by the vanishing partial correlations given by the global Markov property. Hence, this also holds in a linear SEM represented by an ADMG without directed mixed cycles which is a special type of MAG.

4.2.4.2 Chain Graphs

The graph that results from replacing bi-directed edges with undirected edges in an ADMG without directed mixed cycles is a *chain graph*. The class of chain graphs has been studied extensively (see Lauritzen, 1996, for a review).

Some Markov properties have been proposed for chain graphs. The first Markov property for chain graphs has been proposed by Lauritzen and Wermuth (1989) and Frydenberg (1990). Andersson et al. (2001) have introduced another Markov property. These two Markov properties do not correspond to the Markov property for ADMGs. Let G be an ADMG without directed mixed cycles and G' be the chain graph obtained by replacing bi-directed edges with undirected edges. In general, the set of conditional independence relations given by the Markov property for G is not equivalent to that given by either of the two Markov properties for chain graphs. However, there are other Markov properties for chain graphs that correspond to the Markov property for ADMGs without directed mixed cycles (Cox and Wermuth, 1993; Wermuth and Cox, 2001, 2004)[4].

[4] In their terminology, ADMGs without directed mixed cycles correspond to chain graphs with dashed arrows and dashed edges.

www.manaraa.com

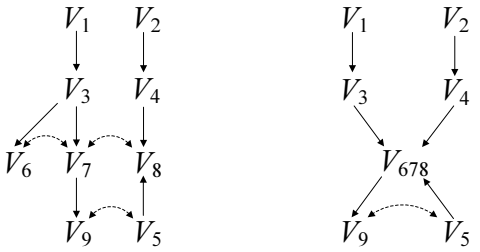


Figure 4.4 (a) An ADMG with directed mixed cycles (b) Illustration of the procedure **GetOrdering**. The modified graph after the first step is shown.

4.3 Markov Properties for General ADMGs

4.3.1 Reducing the Ordered Local Markov Property

When an ADMG G has directed mixed cycles, (RLMP), (RLMP,$\prec_c$), and (PMP,$\prec_c$) are no longer equivalent to (GMP) while (LMP,$\prec$) still is. In this section, we show that the number of conditional independence relations given by (LMP,$\prec$) for an arbitrary ADMG that might have directed mixed cycles can still be reduced. First, we introduce a lemma that gives a condition by which a conditional independence relation renders another conditional independence relation redundant.

Lemma 2 *Given an ADMG G, a consistent ordering $\prec$ on the vertices of G and a vertex X, assume that a probability distribution P satisfies the global Markov property for $G_{\mathrm{pre}_{G,\prec}(X)\setminus\{X\}}$. Let $A = \mathrm{pre}_{G,\prec}(X)$ and A' be a maximal ancestral set such that $X \in A' \subset A$, $A' \cap \mathrm{dis}_{G_A}(X) = \mathrm{dis}_{G_{A'}}(X)$ and $\mathrm{pa}_G(\mathrm{dis}_{G_A}(X) \setminus \mathrm{dis}_{G_{A'}}(X)) \subseteq \mathrm{mb}(X, A')$. Then,*

$$I(\{X\}, \mathrm{mb}(X, A), A \setminus (\mathrm{mb}(X, A) \cup \{X\})) \tag{4.45}$$

implies

$$I(\{X\}, \mathrm{mb}(X, A'), A' \setminus (\mathrm{mb}(X, A') \cup \{X\})). \tag{4.46}$$

We define $\mathrm{rd}_{G,\prec}(X)$ to be the set of all A' satisfying this condition.

www.manaraa.com

Proof: First, we show the relationships among $A, \text{dis}_{G_A}(X), \text{mb}(X, A)$ and $A', \text{dis}_{G_{A'}}(X), \text{mb}(X, A')$. By Lemma 1, we have

$$A' = A \setminus \text{de}_{G_A}(\text{h}(X, A')) \tag{4.47}$$

where

$$\text{h}(X, A') \equiv \text{sp}_{G_A}\Big(\text{dis}_{G_{A'}}(X)\Big) \setminus \Big(\{X\} \cup \text{mb}(X, A')\Big).$$

$\text{dis}_{G_{A'}}(X)$ and $\text{h}(X, A')$ are subsets of $\text{dis}_{G_A}(X)$. Since $\text{dis}_{G_{A'}}(X) \subseteq \{X\} \cup \text{mb}(X, A')$ (by the definition of the Markov blanket), $\text{dis}_{G_{A'}}(X) \cap \text{h}(X, A') = \emptyset$. Thus, we can decompose the set $\text{dis}_{G_A}(X)$ into 3 disjoint subsets as follows.

$$\text{dis}_{G_A}(X) = \text{dis}_{G_{A'}}(X) \cup \text{h}(X, A') \cup B \tag{4.48}$$

where

$$B \equiv \text{dis}_{G_A}(X) \setminus \Big(\text{dis}_{G_{A'}}(X) \cup \text{h}(X, A')\Big).$$

We have

$$\begin{aligned} A' \cap \text{dis}_{G_A}(X) &= A' \cap \Big(\text{dis}_{G_{A'}}(X) \cup \text{h}(X, A') \cup B\Big) \\ &= \text{dis}_{G_{A'}}(X) \cup B \end{aligned}$$

since $\text{dis}_{G_{A'}}(X) \subseteq A', B \subseteq A'$ and $A' \cap \text{h}(X, A') = \emptyset$. From the assumption in Lemma 2 that $A' \cap \text{dis}_{G_A}(X) = \text{dis}_{G_{A'}}(X)$, it follows that $B = \emptyset$. Thus, from (4.48), we have

$$\text{dis}_{G_A}(X) \setminus \text{dis}_{G_{A'}}(X) = \text{h}(X, A'). \tag{4.49}$$

Let $T = \text{dis}_{G_A}(X) \setminus \text{dis}_{G_{A'}}(X) = \text{h}(X, A')$. Then,

$$\begin{aligned} \text{mb}(X, A) &= \text{mb}(X, A') \cup T \cup \text{pa}_G(T) \\ &= \text{mb}(X, A') \cup T \end{aligned} \tag{4.50}$$

since $\text{pa}_G(T) \subseteq \text{mb}(X, A')$ by our assumption. Thus A decomposes into

$$A = A' \cup \text{de}_{G_A}(T) \tag{4.51}$$

www.manaraa.com

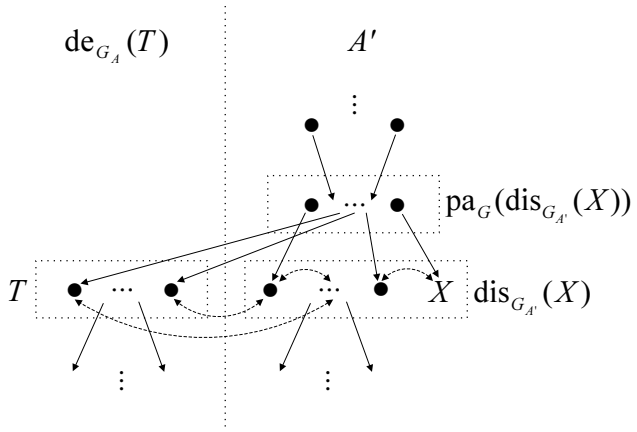


Figure 4.5 The relationship between A and A' that satisfy the conditions in Lemma 2. The induced subgraph G_A is shown. The vertices of G_A are decomposed into two disjoint subsets $\mathrm{de}_{G_A}(T)$ and A'.

since $\mathrm{de}_{G_A}(T) \subseteq A$ and (4.47).

The key relationships among A, $\mathrm{dis}_{G_A}(X)$, $\mathrm{mb}(X, A)$ and A', $\mathrm{dis}_{G_{A'}}(X)$, $\mathrm{mb}(X, A')$ are given by (4.49)–(4.51). Figure 4.5 shows these relationships. We are now ready to prove that $I(\{X\}, \mathrm{mb}(X, A'), A' \setminus (\mathrm{mb}(X, A') \cup \{X\}))$ can be derived from $I(\{X\}, \mathrm{mb}(X, A), A \setminus (\mathrm{mb}(X, A) \cup \{X\}))$. From (4.50) and (4.51), it follows that

$$A \setminus (\mathrm{mb}(X, A) \cup \{X\}) = (A' \cup \mathrm{de}_{G_A}(T)) \setminus (\mathrm{mb}(X, A') \cup \{X\} \cup T)$$

Since $A' \cap \mathrm{de}_{G_A}(T) = \emptyset$, $(\mathrm{mb}(X, A') \cup \{X\}) \cap T = \emptyset$, $\mathrm{mb}(X, A') \cup \{X\} \subseteq A'$ and $T \subseteq \mathrm{de}_{G_A}(T)$, we have

$$A \setminus (\mathrm{mb}(X, A) \cup \{X\}) = \Big(A' \setminus (\mathrm{mb}(X, A') \cup \{X\})\Big) \cup \Big(\mathrm{de}_{G_A}(T) \setminus T\Big). \tag{4.52}$$

Plugging (4.50) and (4.52) into (4.45), we get

$$I\Big(\{X\}, \mathrm{mb}(X, A') \cup T, \Big(A' \setminus (\mathrm{mb}(X, A') \cup \{X\})\Big) \cup \Big(\mathrm{de}_{G_A}(T) \setminus T\Big)\Big).$$

From the decomposition axiom, it follows that

$$I(\{X\}, \mathrm{mb}(X, A') \cup T, A' \setminus (\mathrm{mb}(X, A') \cup \{X\})). \tag{4.53}$$

www.manaraa.com

The last step is to remove T from the conditioning set to obtain $I(\{X\}, \mathrm{mb}(X, A'), A' \setminus (\mathrm{mb}(X, A') \cup \{X\}))$. We claim that

$$I(T, \mathrm{mb}(X, A'), A' \setminus (\mathrm{mb}(X, A') \cup \{X\})). \tag{4.54}$$

We first argue that T is d-separated from $A' \setminus (\mathrm{mb}(X, A') \cup \{X\})$ given $\mathrm{mb}(X, A')$. Consider a vertex $t \in T$ and a vertex $\alpha \in A' \setminus (\mathrm{mb}(X, A') \cup \{X\})$. Note that for any bi-directed edge $t \leftrightarrow \beta$ in G_A, β is either in T or $\mathrm{dis}_{G_{A'}}(X)$. There are only four possible cases for any path in G_A from t to α.

1. $t \leftarrow \gamma \cdots \alpha$
2. $t \rightarrow \cdots \rightarrow \gamma \leftarrow * \cdots \alpha$
3. $t \leftrightarrow \leftrightarrow \cdots \leftrightarrow \delta \leftarrow \gamma \cdots \alpha$
4. $t \leftrightarrow \leftrightarrow \cdots \leftrightarrow \delta \rightarrow \cdots \rightarrow \gamma \leftarrow * \cdots \alpha$

In case 1, $\gamma \in \mathrm{mb}(X, A')$ since $\mathrm{pa}_G(T) \subseteq \mathrm{mb}(X, A')$. Thus, the path is not d-connecting. In case 2, γ is a descendant of t. Since $\mathrm{mb}(X, A')$ does not contain any descendant of t, the path is not d-connecting. Case 3 is similar to case 1, but there are one or more bi-directed edges after t. δ is either in T or $\mathrm{dis}_{G_{A'}}(X)$. It follows that $\gamma \in \mathrm{mb}(X, A')$, so the path is not d-connecting. Case 4 is similar to case 2, but there are one or more bi-directed edges after t. If δ is in T, the argument for case 2 can be applied. If δ is in $\mathrm{dis}_{G_{A'}}(X)$, then $\delta \in \mathrm{mb}(X, A')$, which implies that the path is not d-connecting. This establishes that T is d-separated from $A' \setminus (\mathrm{mb}(X, A') \cup \{X\})$ given $\mathrm{mb}(X, A')$. By the assumption that P satisfies the global Markov property for $G_{\mathrm{pre}_{G,\prec}(X)\setminus\{X\}}$, (4.54) holds. Finally, from (4.53),(4.54) and the contraction axiom, it follows that $I(\{X\}, \mathrm{mb}(X, A'), A' \setminus (\mathrm{mb}(X, A') \cup \{X\}))$. ■

For example, consider the ADMG G in Figure 4.1 and a consistent ordering $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7$. Assume that the global Markov property for $G_{\mathrm{pre}_{G,\prec}(V_6)}$ is satisfied . Let $A = \{V_1, V_2, V_3, V_4, V_5, V_6, V_7\}$ and $A' = \{V_1, V_2, V_3, V_4, V_6, V_7\}$. Then,

$$\mathrm{dis}_{G_A}(V_7) = \{V_5, V_6, V_7\} \tag{4.55}$$

$$\mathrm{dis}_{G_{A'}}(V_7) = \{V_6, V_7\} \tag{4.56}$$

$$A' \cap \mathrm{dis}_{G_A}(V_7) = \{V_6, V_7\} = \mathrm{dis}_{G_{A'}}(V_7) \tag{4.57}$$

$$\mathrm{pa}_G(\mathrm{dis}_{G_A}(V_7) \setminus \mathrm{dis}_{G_{A'}}(V_7)) = \{V_3\} \subseteq \{V_3, V_4, V_6\} = \mathrm{mb}(V_7, A'). \tag{4.58}$$

www.manaraa.com

Thus, by Lemma 2, $I(\{V_7\},\{V_3,V_4,V_6\},\{V_1,V_2\})$ can be derived by $I(\{V_7\},\{V_3,V_4,V_5,V_6\},\{V_1,V_2\})$. Note that in the proof of Lemma 2, the composition axiom is not used. Thus, Lemma 2 can be used to reduce the ordered local Markov property for ADMGs associated with an arbitrary probability distribution.

We now introduce a key concept in eliminating redundant conditional independence relations from (LMP,$\prec$).

Definition 11 (C-ordered Vertex) *Given a consistent ordering $\prec$ on the vertices of an ADMG G, a vertex X is said to be c-ordered in $\prec$ if*

1. *all vertices in* $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$ *are consecutive in $\prec$ and*

2. *for any two vertices Y and Z in* $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$*, there is no directed edge between Y and Z.*

For example, consider the ADMG G in Figure 4.4 (a). $\prec$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$ is a consistent ordering on the vertices of G. $V_1, V_2, \ldots, V_8$ are c-ordered in $\prec$ but V_9 is not since V_5 and V_9 are not consecutive in $\prec$.

The key observation, which will be proved, is that c-ordered vertices contribute to eliminating many redundant conditional independence relations invoked by the ordered local Markov property (LMP,$\prec$). We provide two procedures. The first procedure **ReduceMarkov** in Figure 4.6 constructs a list of conditional independence relations in which some redundant conditional independence relations from (LMP,$\prec$) are not included. **ReduceMarkov** takes as input a fixed ordering $\prec$. The second procedure **GetOrdering** in Figure 4.8 gives a good ordering that might have many c-ordered vertices.

We first describe the procedure **ReduceMarkov**. Given an ADMG G and a consistent ordering $\prec$, **ReduceMarkov** gives a set of conditional independence relations which will be shown to be equivalent to the global Markov property for G. For each vertex V_i, **ReduceMarkov** generates a set of conditional independence relations. If V_i is c-ordered, the relations that correspond to the pairwise Markov property are generated. Otherwise, the relations that correspond to the ordered local Markov property are generated. Also, Lemma 2 is used to remove some redundant relations (by $\mathrm{rd}_{G,\prec}(V_i)$). The output

www.manaraa.com

procedure ReduceMarkov

INPUT: An ADMG G and a consistent ordering $\prec$ on the vertices of G
OUTPUT: A set of conditional independence relations S
$S \leftarrow \emptyset$
for $i = 1, \ldots, n$ **do**
  $I_i \leftarrow \emptyset$
  if V_i is c-ordered in $\prec$ **then**
    for $V_j \prec V_i$ **do**
      $I_i \leftarrow I_i \cup I(\{V_i\}, Z_{ij}, \{V_j\})$ where Z_{ij} is any set of vertices such that Z_{ij} d-separates V_i from V_j and $\forall Z \in Z_{ij}, d(V_i, Z) < d(V_i, V_j)$
    end for
  else
    for all maximal ancestral sets A such that $V_i \in A \subseteq \text{pre}_{G,\prec}(V_i)$, $A \notin \text{rd}_{G,\prec}(V_i)$ **do**
      $I_i \leftarrow I_i \cup I(\{V_i\}, \text{mb}(V_i, A), A \setminus (\text{mb}(V_i, A) \cup \{V_i\}))$
    end for
  end if
  $S \leftarrow S \cup I_i$
end for

Figure 4.6 A procedure to generate a reduced set of conditional independence relations for an ADMG G and a consistent ordering $\prec$

S = **ReduceMarkov**$(G, \prec)$ can be described as follows:

$$S = \bigcup_{X:X \text{ is c-ordered in } \prec} \Big(\bigcup_{Y:Y \prec X} I(\{X\}, Z_{XY}, \{Y\}) \Big) \bigcup \bigcup_{X:X \text{ is not c-ordered in } \prec} \Big(\bigcup_{\substack{\text{all maximal sets } A: \\ X \in A \subseteq \text{pre}_{G,\prec}(X), \\ A \notin \text{rd}_{G,\prec}(X)}} I(\{X\}, \text{mb}(X, A), A \setminus (\text{mb}(X, A) \cup \{X\})) \Big) \tag{4.59}$$

where Z_{XY} is any set of vertices such that Z_{XY} d-separates X from Y and $\forall Z \in Z_{XY}$, $d(X, Z) < d(X, Y)$.

If a vertex X is c-ordered, $O(n)$ conditional independence relations (or zero partial correlations) are added to S. Otherwise, $O(2^n)$ conditional independence relations may be added to S and $O(n2^n)$ zero partial correlations may be invoked. Furthermore, a c-ordered vertex typically involves a smaller conditioning set. $I(\{X\}, Z_{XY}, \{Y\})$ has the conditioning set $|Z_{XY}| \leq |\text{pa}_G(X)|$ while $I(\{X\}, \text{mb}(X, A), A \setminus (\text{mb}(X, A) \cup \{X\}))$ has the conditioning set $|\text{mb}(X, A)| \geq |\text{pa}_G(X)|$.

We now prove that the conditional independence relations produced by **ReduceMarkov** can derive all the conditional independence relations invoked by the global Markov property.

Definition 12 (S-Markov Property (S-MP,$\prec$)) *Let G be an ADMG and $\prec$ be a consistent ordering on*

www.manaraa.com

the vertices of G. Let S be the set of conditional independence relations given by ***ReduceMarkov****(G,≺). A probability distribution P is said to satisfy the S-Markov property for G with respect to ≺, if*

$$(S\text{-MP},\prec) \qquad P \text{ satisfies all the conditional independence relations in } S. \tag{4.60}$$

Theorem 5 *Let G be an ADMG and ≺ be a consistent ordering on the vertices of G. Let S be the set of conditional independence relations given by* ***ReduceMarkov****(G,≺). If a probability distribution P satisfies the composition axiom, then*

$$(\text{GMP}) \Longleftrightarrow (S\text{-MP},\prec). \tag{4.61}$$

Proof: (GMP) $\Longrightarrow$ (S-MP,$\prec$) since every conditional independence relation in (S-MP,$\prec$) corresponds to a valid d-separation. We show (S-MP,$\prec$) $\Longrightarrow$ (GMP). Without any loss of generality, let $\prec$: $V_1 \prec \ldots \prec V_n$. The proof is by induction on the sequence of ordered vertices. Suppose that (S-MP,$\prec$) $\Longrightarrow$ (GMP) holds for $V_1, \ldots V_{i-1}$. Let $S_{i-1} = I_1 \cup \ldots \cup I_{i-1}$. Then, by the induction hypothesis, S_{i-1} contains all the conditional independence relations invoked by (LMP,$\prec$) for $V_1, \ldots V_{i-1}$. If V_i is not c-ordered, $I_i = I(\{V_i\}, \text{mb}(V_i, A), A \setminus (\text{mb}(V_i, A) \cup \{V_i\}))$ for all maximal ancestral sets A such that $V_i \in A \subseteq \text{pre}_{G,\prec}(V_i)$, $A \notin \text{rd}_{G,\prec}(V_i)$. The conditional independence relations invoked by (LMP,$\prec$) with respect to V_i and any $A \in \text{rd}_{G,\prec}(V_i)$ can be derived from other conditional independence relations by Lemma 2. Thus, $S_i = S_{i-1} \cup I_i$ contains all the conditional independence relations invoked by (LMP,$\prec$) for $V_1, \ldots V_i$, which implies (GMP). If V_i is c-ordered, applying the arguments in the proof of (GMP) $\Longleftrightarrow$ (PMP,$\prec_c$), we have

$$I(\{V_i\}, \text{pa}_G(V_i), \text{pre}_{G,\prec}(V_i) \setminus (\{V_i\} \cup \text{pa}_G(V_i) \cup \text{sp}_G(V_i))). \tag{4.62}$$

By the induction hypothesis and the definition of a c-ordered vertex, we have for all $V_j \in \text{dis}_G(V_i) \cap \text{pre}_{G,\prec}(V_i)$

$$I(\{V_j\}, \text{pa}_G(V_j), \text{pre}_{G,\prec}(V_j) \setminus (\{V_j\} \cup \text{pa}_G(V_j) \cup \text{sp}_G(V_j))). \tag{4.63}$$

By the arguments in the proof of (GMP) $\Longleftrightarrow$ (RLMP,$\prec_c$), we have for all maximal ancestral sets A such that $V_i \in A \subseteq \text{pre}_{G,\prec}(V_i)$

$$I(\{V_i\}, \text{mb}(V_i, A), A \setminus (\text{mb}(V_i, A) \cup \{V_i\})). \tag{4.64}$$

www.manaraa.com

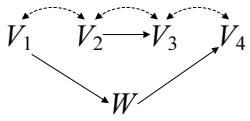


Figure 4.7 The c-component $\{V_1, V_2, V_3, V_4\}$ has the root set $\{V_1, V_2\}$

Therefore, $S_i = S_{i-1} \cup I_i$ derives all the conditional independence relations invoked by (GMP). ■

As we have seen earlier, the number of zero partial correlations critically depends on the number of c-ordered vertices in a given ordering. This motivates us to find the ordering with the most c-ordered vertices. An obvious way of finding this ordering is to explore the space of all the consistent orderings. However, this exhaustive search may become infeasible as the number of vertices grows. We propose a greedy algorithm to get an ordering that has a large number of c-ordered vertices. The basic idea is to first find a large c-component in which many vertices can be c-ordered and place the vertices consecutively in the ordering, then repeating this until we cannot find a set of vertices that can be c-ordered. To describe the algorithm, we define the following notion, which identifies the largest subset of a c-component that can be c-ordered.

Definition 13 (Root Set) *The root set of a c-component C, denoted* $\mathrm{rt}(C)$ *is defined to be the set* $\{V_i \in C \mid$ *there is no* $V_j \in C$ *such that a directed path* $V_j \to \ldots \to V_i$ *exists in* $G\}$.

For example, the c-component $\{V_1, V_2, V_3, V_4\}$ in Figure 4.7 has the root set $\{V_1, V_2\}$. V_3 and V_4 are not in the root set since there are paths $V_2 \to V_3$ and $V_1 \to W \to V_4$. The root set has the following properties.

Proposition 3 *Let $\prec$ be a consistent ordering on the vertices of an ADMG G and C be a c-component of G. If the vertices in* $\mathrm{rt}(C)$ *are consecutive in $\prec$, then all the vertices in* $\mathrm{rt}(C)$ *are c-ordered in $\prec$.*

Proposition 4 *Let $\prec$ be a consistent ordering on the vertices of an ADMG G and C be a c-component of G. If a vertex X in C is c-ordered in $\prec$, then $X \in \mathrm{rt}(C)$.*

Proposition 3 and 4 imply that the root set of a c-component is the largest subset of the c-component that can be c-ordered in a consistent ordering. If G does not have directed mixed cycles, $\mathrm{rt}(C) = C$ for

www.manaraa.com

procedure GetOrdering

INPUT: An ADMG G
OUTPUT: A consistent ordering $\prec$ on V
Step 1:
$G' \leftarrow G$
while (there is a c-component C of G' such that $|\mathrm{rt}(C)| > 1$) **do**
  $M \leftarrow \emptyset$
  for each c-component C of G' **do**
    if $|\mathrm{rt}(C)| > |M|$ **then**
      $M \leftarrow \mathrm{rt}(C)$
    end if
  end for
  Add a vertex V_M to $G'_{V' \setminus M}$
  Draw an edge $V_M \leftarrow X$ (respectively $V_M \rightarrow X$, $V_M \leftrightarrow X$) if there is
  $Y \leftarrow X$ (respectively $Y \rightarrow X$, $Y \leftrightarrow X$) in G' such that $Y \in M, X \in V'$
  Let G' be the resulting graph
end while
Step 2:
Let $\prec'$ be any consistent ordering on V'. Construct a consistent ordering $\prec$ from $\prec'$ by replacing each $V_S \in V' \setminus V$ with the vertices in S (the ordering of the vertices in S is arbitrary)

Figure 4.8 A greedy algorithm to generate a good consistent ordering on the vertices of an ADMG G

every c-component C.

The procedure **GetOrdering** in Figure 4.8 is our proposed greedy algorithm that generates a good consistent ordering for G. In Step 1, it searches for the largest root set M and then merges all the vertices in M to one vertex V_M modifying edges accordingly. Then, it repeats the same operation for the modified graph until there is no root set that contains more than one vertex. Since the vertices in a root set are merged at each iteration, the modified graph is acyclic as otherwise there would be a directed path between two vertices in the root set, which contradicts the condition of a root set. After Step 1, we can easily obtain a consistent ordering for the original graph from the modified graph.

4.3.2 An Example

In this section, we show the application of the procedures **ReduceMarkov** and **GetOrdering** by considering the ADMG G in Figure 4.4 (a). First, we apply **GetOrdering** to get a consistent ordering on the vertices V of G. In Step 1, we first look for the largest root set. The c-component $\{V_6, V_7, V_8\}$ has

www.manaraa.com

the largest root set $\{V_6, V_7, V_8\}$. Then, the vertices in $\{V_6, V_7, V_8\}$ is merged into a vertex V_{678}. Figure 4.4 (b) shows the modified graph G' after the first iteration of the while loop. In the next iteration, we find that every c-component has the root set of size 1. Note that for $C = \{V_5, V_9\}$, $\mathrm{rt}(C) = \{V_5, V_9\}$ in G but $\mathrm{rt}(C) = \{V_5\}$ in G'. Thus, Step 1 ends. In Step 2, from G' in Figure 4.4 (b), we can obtain an ordering $\prec'$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_{678} \prec V_9$. This is converted to a consistent ordering $\prec$: $V_1 \prec V_2 \prec V_3 \prec V_4 \prec V_5 \prec V_6 \prec V_7 \prec V_8 \prec V_9$ for G.

With the ordering $\prec$, we now apply **ReduceMarkov** to obtain a set of conditional independence relations that can derive those invoked by the global Markov property. It is easy to see that the vertices $V_1, \ldots, V_8$ are c-ordered in $\prec$. Thus, the following conditional independence relations corresponding to the pairwise Markov property are added to the set S (initially empty).

$$
\begin{aligned}
&I(\{V_2\}, \emptyset, \{V_1\}), && I(\{V_3\}, \emptyset, \{V_2\}),\\
&I(\{V_4\}, \emptyset, \{V_3, V_1\}), && I(\{V_5\}, \emptyset, \{V_4, V_3, V_2, V_1\}),\\
&I(\{V_6\}, \emptyset, \{V_5, V_4, V_2\}), && I(\{V_6\}, \{V_3\}, \{V_1\}),\\
&I(\{V_7\}, \emptyset, \{V_5, V_4, V_2\}), && I(\{V_7\}, \{V_3\}, \{V_1\}),\\
&I(\{V_8\}, \emptyset, \{V_6, V_3, V_1\}), && I(\{V_8\}, \{V_4\}, \{V_2\}).
\end{aligned} \tag{4.65}
$$

V_9 is not c-ordered in $\prec$ since V_5 is not adjacent in $\prec$. Thus, we use the ordered local Markov property (LMP,$\prec$) for V_9. The maximal ancestral sets that we need to consider are

$$A_1 = \mathrm{an}_G(\{V_6, V_8, V_9\}) = \{V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9\} \text{ and} \tag{4.66}$$

$$A_2 = \mathrm{an}_G(\{V_4, V_6, V_9\}) = \{V_1, V_2, V_3, V_4, V_6, V_7, V_9\}. \tag{4.67}$$

The corresponding conditional independence relations are

$$I(\{V_9\}, \{V_7, V_5\}, \{V_8, V_6, V_4, V_3, V_2, V_1\}), \tag{4.68}$$

$$I(\{V_9\}, \{V_7\}, \{V_6, V_4, V_3, V_2, V_1\}). \tag{4.69}$$

However, it turns out that $A_2 \in \mathrm{rd}_{G,\prec}(V_9)$ and (4.69) is not added to S. Let's check the condition of

www.manaraa.com

Lemma 2. The global Markov property for $G_{\text{pre}_{G,\prec}(V_8)}$ is satisfied by (4.65). Also,

$$\text{dis}_{G_{A_1}}(V_9) = \{V_5, V_9\} \tag{4.70}$$

$$\text{dis}_{G_{A_2}}(V_9) = \{V_9\} \tag{4.71}$$

$$A_2 \cap \text{dis}_{G_{A_1}}(V_9) = \{V_9\} = \text{dis}_{G_{A_2}}(V_9) \tag{4.72}$$

$$\text{pa}_G(\text{dis}_{G_{A_1}}(V_9) \setminus \text{dis}_{G_{A_2}}(V_9)) = \emptyset \subseteq \{V_7\} = \text{mb}(V_9, A_2). \tag{4.73}$$

Therefore, the condition of Lemma 2 is satisfied and it follows that (4.69) is redundant. To see how much we reduced the testing requirements, the conditional independence relations invoked by (LMP,$\prec$) are shown below.

$$\begin{aligned}
&I(\{V_2\}, \emptyset, \{V_1\}), && I(\{V_3\}, \{V_1\}, \{V_2\}),\\
&I(\{V_4\}, \{V_2\}, \{V_3, V_1\}), && I(\{V_5\}, \emptyset, \{V_4, V_3, V_2, V_1\}),\\
&I(\{V_6\}, \{V_3\}, \{V_5, V_4, V_2, V_1\}), && I(\{V_7\}, \{V_3\}, \{V_5, V_4, V_2, V_1\}),\\
&I(\{V_7\}, \{V_6, V_3\}, \{V_5, V_4, V_2, V_1\}), && I(\{V_8\}, \{V_5, V_4\}, \{V_6, V_3, V_2, V_1\}),\\
&I(\{V_8\}, \{V_7, V_5, V_4, V_3\}, \{V_2, V_1\}), && I(\{V_8\}, \{V_7, V_6, V_5, V_4, V_3\}, \{V_2, V_1\}),\\
&I(\{V_9\}, \{V_7\}, \{V_6, V_4, V_3, V_2, V_1\}), && I(\{V_9\}, \{V_7, V_5\}, \{V_8, V_6, V_4, V_3, V_2, V_1\}).
\end{aligned} \tag{4.74}$$

S invokes 26 zero partial correlations while (LMP,$\prec$) invokes 39. Also, S involves much smaller conditioning sets. We have at most one vertex in each conditioning set in (4.65) and two vertices in (4.68) while 23 zero partial correlations in (4.74) involve more than 2 vertices in the conditioning set.

The ADMG G in this example turns out to be a MAG. As we discussed in Section 4.2.4.1, we have two options: either we use the constraints in (4.65) and (4.68) or the constraints given by the pairwise Markov property for MAGs. In this example, both sets of constraints involve the same number of zero partial correlations. However, the pairwise Markov property for MAGs involves much larger conditioning sets. For example, the pairwise Markov property for MAGs gives the following conditional independence relation for the pair V_6 and V_8: $I(\{V_8\}, \{V_5, V_4, V_3, V_2, V_1\}, \{V_6\})$. Our method uses an empty set as the conditioning set for the pair. Hence, in this example, we are better off using the constraints in (4.65) and (4.68).

www.manaraa.com

4.3.3 Comparison of (LMP,≺) and (S-MP,≺)

From (4.59), it is clear that (S-MP,≺) invokes less conditional independence relations than (LMP,≺) if there are c-ordered vertices in ≺. But how much more economical is (S-MP,≺) than (LMP,≺) and for what type of graphs is the reduction large?

For simplicity, we will compare the number of conditional independence relations rather than zero partial correlations and ignore the reduction done by Lemma 2. For now assume

$$S = \bigcup_{X:X \text{ is c-ordered in } \prec} I(\{X\}, \mathrm{pa}_G(X), \mathrm{pre}_{G,\prec}(X) \setminus (\{X\} \cup \mathrm{pa}_G(X) \cup \mathrm{sp}_G(X))) \bigcup \bigcup_{X:X \text{ is not c-ordered in } \prec} \Big(\bigcup_{\substack{\text{all maximal sets } A: \\ X \in A \subseteq \mathrm{pre}_{G,\prec}(X)}} I(\{X\}, \mathrm{mb}(X,A), A \setminus (\mathrm{mb}(X,A) \cup \{X\})) \Big).$$

Let $\mathrm{M}(X, \prec)$ be the number of different Markov blankets of a vertex X, that is, $\mathrm{M}(X, \prec) = \left|\{\mathrm{dis}_{G_A}(X) \mid A \text{ is an ancestral set such that } X \in A \subseteq \mathrm{pre}_{G,\prec}(X)\}\right|$, and $\mathrm{C}(\prec)$ be the set of vertices that are c-ordered in ≺. Then, (LMP,≺) lists $\sum_{X \in V} \mathrm{M}(X, \prec)$ conditional independence relations and (S-MP,≺) lists $|\mathrm{C}(\prec)| + \sum_{X \notin \mathrm{C}(\prec)} \mathrm{M}(X, \prec)$ conditional independence relations. Hence, the difference in the number of conditional independence relations between (LMP,≺) and (S-MP,≺) is

$$\sum_{X \in \mathrm{C}(\prec)} \big(\mathrm{M}(X, \prec) - 1\big).$$

This difference is large when $|\mathrm{C}(\prec)|$ or $\mathrm{M}(X, \prec)$ for each X is large.

The size of $\mathrm{C}(\prec)$ depends on the number of directed mixed cycles. From Definition 11, it follows that $\mathrm{C}(\prec)$ is large if there are a small number of directed mixed cycles. Note that a directed mixed cycle such as that in Figure 4.3 induces the violation of the first condition in Definition 11 and a directed mixed cycle of the form $\alpha \overset{\leftrightarrow}{\rightarrow} \beta$ induces the violation of the second condition in Definition 11.

$\mathrm{M}(X, \prec)$ depends on the structure of $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$. We will reformulate $\mathrm{M}(X, \prec)$ to show the properties that affect $\mathrm{M}(X, \prec)$. Let $G_{\leftrightarrow,\mathrm{dis}}(X, \prec) = (V', E')$ where $V' = \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$ and $E' = \{V_i \leftrightarrow V_j \mid V_i \leftrightarrow V_j \text{ in } G_{V'}\}$. For example, for an ADMG G in Figure 4.7 and an ordering $V_1 \prec V_2 \prec V_3 \prec V_4$, $G_{\leftrightarrow,\mathrm{dis}}(V_3, \prec)$ is $V_1 \leftrightarrow V_2 \leftrightarrow V_3$. Let $G_{\leftrightarrow,\mathrm{dis}}(X, \prec)_S$ be the induced subgraph of $G_{\leftrightarrow,\mathrm{dis}}(X, \prec)$ on a set $S \subseteq \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$. Then, $\mathrm{M}(X, \prec) = \left|\{S \mid S \subseteq \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X) \text{ such that } G_{\leftrightarrow,\mathrm{dis}}(X, \prec)_S \text{ is a } \textit{connected component} \text{ of } G_{\leftrightarrow,\mathrm{dis}}(X, \prec)_{S \cup (\mathrm{an}_G(S) \cap \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X))}\}\right|$, that is, $\mathrm{M}(X, \prec)$ corresponds to

www.manaraa.com

a set of subsets S of $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$ satisfying two conditions: (i) $G_{\leftrightarrow,\mathrm{dis}}(X,\prec)_S$ is connected; and (ii) for all $Y \in \left(\mathrm{an}_G(S) \cap \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)\right) \setminus S$, there is no path from Y to any vertices in S. The condition (i) implies that $\mathrm{M}(X,\prec)$ will be large if the vertices in $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$ are connected by many bi-directed edges. The condition (ii) implies that $\mathrm{M}(X,\prec)$ will be large if there are few directed mixed cycles. Note that for ADMGs without directed mixed cycles, (ii) trivially holds since $\left(\mathrm{an}_G(S) \cap \mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)\right) \setminus S = \emptyset$. For example, consider a subset of vertices $\{V_1,\ldots,V_k\}$ in an ADMG with edges $V_i \leftrightarrow V_k, i = 1,\ldots,k-1$, which has no directed mixed cycles. Then, for an ordering $V_1 \prec \ldots \prec V_k$, $\mathrm{M}(V_k,\prec) = 2^{k-1}$. Also, consider a subset of vertices $\{V_1,\ldots,V_k\}$ in an ADMG with edges $V_1 \overset{\leftrightarrow}{\rightarrow} V_2 \overset{\leftrightarrow}{\rightarrow} \cdots \overset{\leftrightarrow}{\rightarrow} V_k$, which has $k-1$ directed mixed cycles. Then, $\mathrm{M}(V_k,\prec) = 1$. Hence, it is clear that $\mathrm{M}(X,\prec)$ is large if

1. the set $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$ is large,
2. there are many bi-directed edges connecting vertices in $\mathrm{dis}_G(X) \cap \mathrm{pre}_{G,\prec}(X)$, and
3. there are few directed mixed cycles.

Thus, (LMP,$\prec$) will invoke a large number of conditional independence relations for an ADMG with few directed mixed cycles and large c-components with many bi-directed edges. However, for such an ADMG, $\sum_{X\in\mathrm{C}(\prec)}\left(\mathrm{M}(X,\prec)-1\right)$, the reduction made by (S-MP,$\prec$), is also large. An extreme case is an ADMG that has no directed mixed cycles and each c-component of which is a clique joined by bi-directed edges. An example of such an ADMG is given in Figure 4.9. For this ADMG and an ordering $W \prec V \prec X \prec Y \prec Z$, (LMP,$\prec$) invokes $\mathrm{M}(W,\prec) + \mathrm{M}(V,\prec) + \mathrm{M}(X,\prec) + \mathrm{M}(Y,\prec) + \mathrm{M}(Z,\prec) = 1 + 1 + 1 + 2 + 4 = 9$ conditional independence relations while (S-MP,$\prec$) invokes $|\mathrm{C}(\prec)| = n = 5$ conditional independence relations. If we enlarge the clique joined by bi-directed edges such that it contains k vertices, then (LMP,$\prec$) invokes $2 + \sum_{i=0}^{k-1} 2^i = 1 + 2^k$ conditional independence relations while (S-MP,$\prec$) invokes $k+2$.

In general, although (S-MP,$\prec$) greatly reduces (LMP,$\prec$), it may still invoke an exponential number of conditional independence relations if there exist directed mixed cycles.

www.manaraa.com

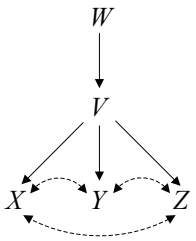


Figure 4.9 An example ADMG for which using (S-MP,$\prec$) is most beneficial. There is no directed mixed cycle and each c-component is a clique joined by bi-directed edges.

www.manaraa.com

CHAPTER 5. POLYNOMIAL CONSTRAINTS IN CAUSAL BAYESIAN NETWORKS

In this chapter, we seek the constraints imposed by a causal BN on both nonexperimental and interventional distributions. When all variables are observed, a complete characterization of constraints on interventional distributions imposed by a given causal BN has been given in (Pearl, 2000, pp.23-4). In a causal BN containing hidden variables, a class of equality and inequality constraints on interventional distributions are given in Kang and Tian (2006). In this chapter, we propose to use the implicitization procedure to generate polynomial constraints on interventional distributions induced by a causal BN with hidden variables. The main challenges in applying the implicitization procedure on interventional distributions are:

(i) *Computational complexity*. The generic complexity of implicitization is known to be exponential in the number of variables (number of parameters for this problem). When we consider interventional distributions, the number of variables greatly increases compared to the case of non-experimental distribution, which makes the computation infeasible even for small causal BNs.

(ii) *Understanding structures of constraints*. Finding a syntactic structure of the constraints computed by implicitization also becomes complicated.

To deal with challenge (i), we show three methods to reduce the complexity of the implicitization problem (Section 5.3). We illustrate our methods showing a model in which the generic implicitization procedure is intractable while our methods can solve the problem (Section 5.3.2). We also show an example of new constraints on interventional distributions that are not captured by the types of constraints in Kang and Tian (2006) (Section 5.3.2). To deal with challenge (ii), we present some preliminary results on the algebraic structure of polynomial constraints on interventional distributions implied by

www.manaraa.com

certain classes of causal BNs with hidden variables (Section 5.3.2). We show some preliminary results in causal BNs without hidden variables, which are expected to be useful in understanding syntactic structures of the constraints for BNs with hidden variables (Section 5.2).

We provide a model testing procedure using polynomial constraints and present some experiments validating this procedure (Section 5.4). We also discuss a possibility of using polynomial constraints to differentiate Markov equivalent models (Section 5.4).

5.1 Problem Statement

We define the *implicitization* problem for a set of interventional distributions. We explain what the polynomial constraints computed by the implicitization problem mean algebraically.

Let $\boldsymbol{P}_{intv}$ denote a set of interventional distributions. For example, $\boldsymbol{P}_{intv}$=$\{P(v_1, v_2), P_{V_1=1}(V_1 = 1, v_2)\}$ contains a non-experimental distribution $P(v_1, v_2)$ and an interventional distribution $P_{V_1=1}(V_1 = 1, v_2)$ where the treatment variable V_1 is fixed to 1. We will regard $P(v)$ to be a special interventional distribution where $T = \emptyset$ allowing it to be in $\boldsymbol{P}_{intv}$. Let $\boldsymbol{P}_*$ denote the set of all interventional distributions $\boldsymbol{P}_* = \{P_t(v) | T \subset V, t \in Dm(T), v \in Dm(V), v \text{ is consistent with } t\}$ where $Dm(T)$ represents the domain of T. For example, let $V = \{V_1, V_2\}$ where both variables are binary, then $\boldsymbol{P}_* = \{P(v_1, v_2), P_{V_1=1}(V_1 = 1, v_2), P_{V_1=2}(V_1 = 2, v_2), P_{V_2=1}(v_1, V_2 = 1), P_{V_2=2}(v_1, V_2 = 2)\}$.

We can describe $\boldsymbol{P}_{intv}$ in terms of a polynomial mapping from a set of parameters to the distributions as follows.

First, consider a causal BN G without hidden variables. Let $V_1, \ldots, V_n$ be the vertices of G. We denote the joint space parameter defining $P_t(v)$ for v consistent with t by p^t_v and the model parameter defining $P(v_i|pa_i)$ by $q^i_{v_i pa_i}$. The model parameters are subjected to the linear relations $\sum_{v_i} q^i_{v_i pa_i} = 1$. Thus, we have introduced $(d_i - 1) \prod_{\{j|V_j \in PA_i\}} d_j$ model parameters for the vertex V_i where $d_i = |Dm(V_i)|$. Let $J_{\boldsymbol{P}_{intv}}$ denote the set of joint space parameters associated with $\boldsymbol{P}_{intv}$ and M denote the set of model parameters. For example, consider the causal BN G in Figure 5.1 (a) in which variables are binary. Let $\boldsymbol{P}_{intv}$ be the set of two distributions $\{P(v_1, v_2, v_3), P_{V_1=1}(V_1 = 1, v_2, v3)\}$. Then, $J_{\boldsymbol{P}_{intv}}$=$\{p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, p_{222}, p^{V_1=1}_{111}, p^{V_1=1}_{112}, p^{V_1=1}_{121}, p^{V_1=1}_{122}\}$ and $M = \{q^1_{11}, q^1_{12}, q^2_{11}, q^2_{12}, q^3_1\}$. The mapping re-

www.manaraa.com

lated to (3.5) is

$$\phi : \mathbb{R}^M \to \mathbb{R}^{J_{\boldsymbol{P}_{intv}}},$$
$$p_v^t = \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i} \tag{5.1}$$

where $\mathbb{R}^M$ and $\mathbb{R}^{J_{\boldsymbol{P}_{intv}}}$ denote the real vector space of dimension $|M|$ and $|J_{\boldsymbol{P}_{intv}}|$ respectively. For example, the mapping for the previous example is given by the following relationships:

$$\begin{aligned}
p_{111} &= q^1_{11} q^2_{11} q^3_1, \\
p_{112} &= q^1_{12} q^2_{12} (1 - q^3_1), \\
p_{121} &= q^1_{11} (1 - q^2_{11}) q^3_1, \\
p_{122} &= q^1_{12} (1 - q^2_{12}) (1 - q^3_1), \\
p_{211} &= (1 - q^1_{11}) q^2_{11} q^3_1, \\
p_{212} &= (1 - q^1_{12}) q^2_{12} (1 - q^3_1), \\
p_{221} &= (1 - q^1_{11}) (1 - q^2_{11}) q^3_1, \\
p_{222} &= (1 - q^1_{12}) (1 - q^2_{12}) (1 - q^3_1), \\
p^{V_1=1}_{111} &= q^2_{11} q^3_1, \\
p^{V_1=1}_{112} &= q^2_{12} (1 - q^3_1), \\
p^{V_1=1}_{121} &= (1 - q^2_{11}) q^3_1, \\
p^{V_1=1}_{122} &= (1 - q^2_{12}) (1 - q^3_1).
\end{aligned}$$

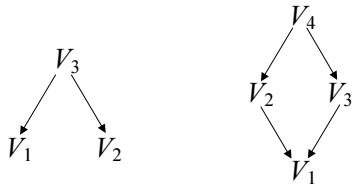
(5.1) induces a ring homomorphism

$$\Phi : \mathbb{R}[J_{\boldsymbol{P}_{intv}}] \to \mathbb{R}[M] \tag{5.2}$$

which takes the unknown p_v^t to $\prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i}$.

Second, consider a causal BN G with hidden variables. Let $\{V_1, \ldots, V_n\}$ and $\{U_1, \ldots, U_{n'}\}$ be sets of observed and hidden variables respectively. We denote the joint space parameters defining $P_t(v)$ for v consistent with t by p_v^t and the model parameters defining $P(v_i|pa_i, u^i)$ and $P(u_j)$ by $q^i_{v_i pa_i u^i}$ and $r^j_{u_j}$ respectively. The joint space parameters and the model parameters form two rings of polynomials

www.manaraa.com



(a) (b)

Figure 5.1 Two causal BNs.

$\mathbb{R}[J_{\boldsymbol{P}_{intv}}]$ and $\mathbb{R}[M]$. The mapping related to (3.7) is

$$\pi : \mathbb{R}^M \to \mathbb{R}^{J_{\boldsymbol{P}_{intv}}},$$

$$p_v^t = \sum_{u_1 \ldots u_{n'}} \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i u^i} \prod_{j=1}^{n'} r^j_{u_j}. \tag{5.3}$$

(5.3) induces a ring homomorphism

$$\Psi : \mathbb{R}[J_{\boldsymbol{P}_{intv}}] \to \mathbb{R}[M]. \tag{5.4}$$

By Tarski-Seidenberg theorem, the image of ϕ (or π) corresponds to a semi-algebraic set, which can be described by a set of polynomial equalities and inequalities. Finding all of these equalities and inequalities is usually infeasible. In this chapter, we choose to find a set of polynomial equalities that define the smallest algebraic set that contains the image of ϕ (or π). These polynomial equalities are a subset of the constraints that describe the image of ϕ (or π) and are equal to the *kernel* of the ring homomorphism Φ (or Ψ). The *kernel* of Φ, denoted by $\ker(\Phi)$ is the ideal consisting of all polynomials f in $\mathbb{R}[J_{\boldsymbol{P}_{intv}}]$ such that $\Phi(f) = 0$. Thus, the vanishing of the polynomial equalities in $\ker(\Phi)$ and $\ker(\Psi)$ is a necessary condition that there exist the model parameters in (5.1) and (5.3) respectively. The process of computing $\ker(\Phi)$ is called *implicitization*.

Our goal is to compute and analyze the kernels for causal BNs with or without hidden variables.

5.2 Causal Bayesian Network with No Hidden Variables

Consider a causal BN G and a set of interventional distributions $\boldsymbol{P}_{intv}$. If checking whether each $P_t(v) \in \boldsymbol{P}_{intv}$ factors as in (3.5) is the only goal, it is not necessary to solve the implicitization problem

www.manaraa.com

since you can use the constraints (3.5) given by the definition or the constraints given in (Pearl, 2000, pp.23-4). However, we study the implicitization problem for a set of interventional distributions associated with a causal BN without hidden variables, since we expect that the structure of the constraints for a causal BN without hidden variables may reveal some syntactic structure of the constraints for a causal BN with hidden variables. For non-experimental distribution, Garcia et al. (2005) showed that the constraints for a BN without hidden variables can help finding the structure of the constraints for a BN with hidden variables.

Since the computation of the constraints for causal BNs without hidden variables is relatively easy, we will focus on the analysis of the computed constraints. In this section, we give a preliminary result on the algebraic structure of the constraints for a set of interventional distributions associated with causal BNs without hidden variables. The problem of characterizing the structure of the constraints for arbitrary set of interventional distributions is still open. We show a few cases in which the constraints can be nicely described by a simple set of polynomials.

5.2.1 One Interventional Distribution

Suppose $\boldsymbol{P}_{intv}$ contains only one interventional distribution $P_t(v)$. For non-experimental distribution $P(v)$, Garcia et al. (2005) showed that

$$\ker(\Phi) = (I_{\text{local}(G)} : \mathbf{p}^{\infty}) + \langle \sum_{v} p_v - 1 \rangle \tag{5.5}$$

where $I_{\text{local}(G)}$ is the ideal associated to the local Markov property on a BN G and $\mathbf{p}$ is the product of all linear forms $p_{+\ldots+v_{r+1}\ldots v_n} = \sum_{v_1,\ldots,v_r} p_{v_1\ldots v_r v_{r+1}\ldots v_n}$ and $I : f^{\infty} = \{g \in \mathbb{R}[J_{\{P(v)\}}] \mid g f^N \in I, \text{ for some } N\}$ denotes the *saturation* of I by f.

The local Markov property on G is the set of independence statements

$$\text{local}(G) = \{V_i \perp\!\!\!\perp \text{ND}(V_i) | \text{PA}(V_i) : i = 1, \ldots, n\} \tag{5.6}$$

where $\text{ND}(V_i)$ denotes the set of nondescendants of V_i in G and $\text{PA}(V_i)$ denotes the set of parents of V_i in G.

For example, consider the causal BN G in Figure 5.1 (a). Assume that all variables are binary. The local Markov property on G has only one element $V_1 \perp\!\!\!\perp V_2 \mid V_3$. The constraints induced by an

www.manaraa.com

independence statement, $A \perp\!\!\!\perp B \mid C$ are given by the vanishing of the polynomials

$$\begin{aligned} &P(A=a,B=b,C=c)P(A=a',B=b',C=c) \\ &- P(A=a',B=b,C=c)P(A=a,B=b',C=c) \end{aligned} \tag{5.7}$$

for all a, a', b, b', c. Thus, the ideal $I_{\text{local}(G)}$ associated with the local Markov property on G is

$$I_{\text{local}(G)} = \langle p_{111}p_{221} - p_{121}p_{211}, p_{112}p_{222} - p_{122}p_{212} \rangle. \tag{5.8}$$

For this particular BN G, it turns out that

$$\begin{aligned} I_{\text{local}(G)} : \mathbf{p}^{\infty} =& I_{\text{local}(G)} : (p_{111} \dots p_{222}p_{+11} \dots p_{+22}p_{++1}p_{++2})^{\infty} \\ =& I_{\text{local}(G)}. \end{aligned} \tag{5.9}$$

From (5.5), it follows that

$$\ker(\Phi) = I_{\text{local}(G)} + \langle \sum_{v} p_v - 1 \rangle. \tag{5.10}$$

In general, however, $\ker(\Phi)$ does not coincide with $I_{\text{local}(G)}$. For example, $I_{\text{local}(G)} : \mathbf{p}^{\infty}$ for the causal BN G in Figure 5.1 (b) includes additional generators other than $I_{\text{local}(G)}$. See Sturmfels (2002); Garcia et al. (2005) for details.

The above result can be applied to an arbitrary interventional distribution $P_t(v)$. We see that the mapping in (5.1) defined for $P_t(v)$ and G is equivalent to the mapping defined for $P(v \setminus t)$ and $G(V \setminus T)$ where $G(C)$ denotes the subgraph of G composed only of the variables in C. Thus, the following holds.

Proposition 5 *Let Φ be a ring homomorphism*

$$\Phi : \mathbb{R}[J_{\{P_t(v)\}}] \rightarrow \mathbb{R}[M] \tag{5.11}$$

induced by (5.1). Then, we have

$$\ker(\Phi) = (I_{local(G(V \setminus T))} : \mathbf{p}^{\infty}) + \langle \sum_{v \setminus t} p_v^t - 1 \rangle \tag{5.12}$$

where $\mathbf{p}$ is the product of all linear forms $p_{+\dots+v_{i_{r+1}}\dots v_{i_k}}$ when $V \setminus T = \{V_{i_1}, \dots, V_{i_k}\}, V_{i_1} > \dots > V_{i_k}$.

www.manaraa.com

5.2.2 All Interventional Distributions

Consider the set of all interventional distributions $\boldsymbol{P}_*$. For any joint space parameter p_v^t, we have

$$p_v^t = \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i} = \prod_{\{i|V_i \notin T\}} p_{v_i}^{v \setminus v_i}. \tag{5.13}$$

Thus, every joint space parameter can be written as the product of some other joint space parameters. Then,

$$\ker(\Phi) = \langle p_v^t - \prod_{\{i|V_i \notin T\}} p_{v_i}^{v \setminus v_i} : \forall v, t \rangle. \tag{5.14}$$

5.2.3 Two Interventional Distributions

Consider the case in which $\boldsymbol{P}_{intv}$ has two distributions. We show some cases in which $\ker(\Phi)$ can be described by a simple set of polynomials.

Consider the causal BN G in Figure 5.1 (a) where all variables are binary. Suppose $\boldsymbol{P}_{intv} = \{P(v), P_{V_1=1}(v)\}$. We have the following relation between $p^{V_1=1}_{1v_2v_3}$ and p_v. For any v_2 and v_3,

$$p^{V_1=1}_{1v_2v_3} = \sum_{v_1} p_{v_1v_2v_3}. \tag{5.15}$$

Let Φ denote a ring homomorphism

$$\Phi : \mathbb{R}[J_{\{P(v), P_{V_1=1}(v_2,v_3)\}}] \rightarrow \mathbb{R}[M]. \tag{5.16}$$

Since the joint space parameter $p^{V_1=1}_{1v_2v_3}$ for any v_2 and v_3 is a polynomial function of some of joint space parameters p_v, we have

$$\ker(\Phi) = \ker(\Phi') + \langle p^{V_1=1}_{1v_2v_3} - \sum_{v_1} p_{v_1v_2v_3} : \forall v_2, v_3 \rangle \tag{5.17}$$

where Φ' denotes the ring homomorphism

$$\Phi' : \mathbb{R}[J_{\{P(v)\}}] \rightarrow \mathbb{R}[M]. \tag{5.18}$$

From (5.10), it follows that

$$\ker(\Phi) = I_{\text{local}(G)} + \langle \sum_v p_v - 1 \rangle + \langle p^{V_1=1}_{1v_2v_3} - \sum_{v_1} p_{v_1v_2v_3} : \forall v_2, v_3 \rangle. \tag{5.19}$$

Note that the equation in (5.15) holds because the set $\{V_2, V_3\}$ contains its own ancestors in G. We have the following proposition.

www.manaraa.com

Proposition 6 *Suppose* $\boldsymbol{P}_{intv} = \{P(v), P_t(v)\}$. *Let* Φ *and* Φ' *be ring homomorphisms*

$$\Phi : \mathbb{R}[J_{\{P(v),P_t(v)\}}] \rightarrow \mathbb{R}[M],\ \Phi' : \mathbb{R}[J_{\{P(v)\}}] \rightarrow \mathbb{R}[M]. \tag{5.20}$$

If $V \setminus T$ *contains its own ancestors in* G, *we have*

$$\ker(\Phi) = \ker(\Phi') + \langle p_v^t - \sum_t p_v : \forall(v \setminus t)\rangle. \tag{5.21}$$

The relationship between two distributions in the above proposition is the result of Lemma 5 in Section 5.3.

Now consider the causal BN G in Figure 5.1 (a) and suppose that $\boldsymbol{P}_{intv} = \{P(v), P_{V_3=1}(v)\}$. In this case, $P_{V_3=1}(v)$ cannot be represented as a polynomial function of $P(v)$. However, we can describe the generators of $\ker(\Phi)$ as follows. Given an instantiation of all the variables v and an instantiation of treatment variables t, let $V_{cons} = \{V_i \in V \setminus T \mid v_i pa_i$ in v is consistent with $t\}$ and $\text{cons}(v, t)$ denote the instantiation of V obtained by replacing the inconsistent variables in v with the values of t. For example, for G in Figure 5.1 (a), if $v = (V_1 = 1, V_2 = 1, V_3 = 1)$ and $t = (V_2 = 2)$, then $V_{cons} = \{V_1, V_3\}$ and $\text{cons}(v, t) = (V_1 = 1, V_2 = 2, V_3 = 1)$. We have the following lemma.

Lemma 3 *Suppose* $\boldsymbol{P}_{intv} = \{P(v), P_t(v)\}$. *Let* Φ, Φ' *and* Φ'' *be ring homomorphisms*

$$\Phi : \mathbb{R}[J_{\{P(v),P_t(v)\}}] \rightarrow \mathbb{R}[M],\ \Phi' : \mathbb{R}[J_{\{P(v)\}}] \rightarrow \mathbb{R}[M],\ \Phi'' : \mathbb{R}[J_{\{P_t(v)\}}] \rightarrow \mathbb{R}[M]. \tag{5.22}$$

If for any two vertices V_i *and* V_j *in* $V \setminus T$, V_i *is neither* V_j*'s ancestor nor its descendent, then*

(i) there exist two disjoint subsets $W_1 = \{A_1, \ldots, A_i\}$ *and* $W_2 = \{C_1, \ldots, C_k\}$ *of* T *such that*

$$A_1 > \ldots > A_i > B_1 > \ldots > B_j > C_1 > \ldots > C_k \tag{5.23}$$

is a consistent topological ordering of variables in G *where* $V \setminus T = \{B_1, \ldots, B_j\}$ *and*

(ii)

$$\ker(\Phi) = \ker(\Phi') + \ker(\Phi'') + \langle f(v, t) \sum_{w_1, v_{cons}} p_v - \sum_{w_1} p_v : \forall v\rangle \tag{5.24}$$

where

$$f(v, t) = \prod_{\{i \mid V_i \in V_{cons}\}} \sum_{v_{cons} \setminus v_i} p^t_{cons(v,t)}. \tag{5.25}$$

www.manaraa.com

Proof: We define the ideal I associated with Φ.

$$I = \langle p_v - \prod_i q^i_{v_i pa_i} : \forall v \rangle + \langle p^t_v - \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i} : \forall (v \setminus t) \rangle. \tag{5.26}$$

The elimination ideal $I \cap \mathbb{R}[J_{\{P(v),P_t(v)\}}]$ is equivalent to $\ker(\Phi)$. The idea is that we can represent I as the sum of three ideal I_1, I_2 and I_3 such that the model parameters in I_1 and those in I_2 are disjoint and no model parameter appears in I_3 and thus

$$\begin{aligned} \ker(\Phi) =& I \cap \mathbb{R}[J_{\{P(v),P_t(v)\}}] \\ =& I_1 \cap \mathbb{R}[J_{\{P(v)\}}] + I_2 \cap \mathbb{R}[J_{\{P_t(v)\}}] + I_3 \\ =& \ker(\Phi') + \ker(\Phi'') + I_3. \end{aligned} \tag{5.27}$$

Let $I_1 = \langle p_v - \prod_i q^i_{v_i pa_i} : \forall v \rangle$ and $I_2 = \langle p^t_v - \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i} : \forall (v \setminus t) \rangle$. We will replace each generator in I_1 with two other polynomials and add one polynomial to I_3 which is initially empty as follows.

For any polynomial $p_v - \prod_i q^i_{v_i pa_i}$, we have

$$p_v - \prod_i q^i_{v_i pa_i} \tag{5.28}$$

$$= p_v - \Big(\prod_{\{i|V_i \in W_1\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in V \setminus T\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i} \Big)$$

$$= p_v - \Big(\prod_{\{i|V_i \in W_1\}} q^i_{v_i pa_i} \Big) \Big(\sum_{w_1} p_v \Big) \tag{5.29}$$

since

$$\sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V \setminus T\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i} \Big)$$

is in I. Also,

$$\begin{aligned} &\sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V \setminus T\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i} \Big) \\ &= \sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V_{cons}\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in (V \setminus T) \setminus V_{cons}\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i} \Big) \end{aligned}$$

From the property that any two vertices V_i and V_j in $V \setminus T$, V_i is neither V_j's ancestor nor its parent, it follows that the polynomial

$$\sum_{w_1, V_{cons}} p_v - \Big(\prod_{\{i|V_i \in (V \setminus T) \setminus V_{cons}\}} q^i_{v_i pa_i} \Big) \Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i} \Big) \tag{5.30}$$

www.manaraa.com

is in I. Thus,

$$\sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V \setminus T\}} q^i_{v_i pa_i}\Big)\Big(\prod_{\{i|V_i \in W_2\}} q^i_{v_i pa_i}\Big) = \sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V_{cons}\}} q^i_{v_i pa_i}\Big)\Big(\sum_{w_1, v_{cons}} p_v\Big)$$
$$= \sum_{w_1} p_v - \Big(\prod_{\{i|V_i \in V_{cons}\}} \sum_{v_{cons} \setminus v_i} p^t_{cons(v,t)}\Big)\Big(\sum_{w_1, v_{cons}} p_v\Big). \quad (5.31)$$

We replace the polynomial (5.28) with the polynomials (5.29) and (5.30) and add the polynomial (5.31) to I_3. After processing every polynomial in I_1, we have three ideal I_1, I_2 and I_3 with the desired property.

■

We can use Lemma 3 to compute ker(Φ) for the causal BN G in Figure 5.1 (a) and $\boldsymbol{P}_{intv} = \{P(v), P_{V_3=1}(v)\}$ since V_1 is neither V_2's ancestor nor its descendent. It turns out that

$$\ker(\Phi) = \ker(\Phi') + \ker(\Phi'') + \langle p^{V_3=1}_{v_1 v_2 1} \sum_{v_1, v_2} p_{v_1 v_2 1} - p_{v_1 v_2 1} : \forall v_1, v_2 \rangle$$
$$= I_{\text{local}(G)} + \langle \sum_v p_v - 1 \rangle + I_{\text{local}(G(\{V_1, V_2\})} + \langle \sum_{v_1, v_2} p^{V_3=1}_v - 1 \rangle$$
$$+ \langle p^{V_3=1}_{v_1 v_2 1} \sum_{v_1, v_2} p_{v_1 v_2 1} - p_{v_1 v_2 1} : \forall v_1, v_2 \rangle. \quad (5.32)$$

5.3 Causal Bayesian Network with Hidden Variables

Solving the implicitization problem for a causal BN with hidden variables has a high computational demand. The implicitization problem can be solved by computing a certain Groebner basis and it is known that computing a Groebner basis has the generic complexity $m^{O(1)}g^{O(N)}$ where m is the number of equations, g is the degree of the polynomials and N is the number of variables. In our implicitization problems, N is the sum of the number of joint space parameters and model parameters. Consider the implicitization for non-experimental distribution. The number of joint space parameters for non-experimental distribution is $d_1 \dots d_n$. Solving the implicitization problem becomes intractable as the number of vertices in the causal BN and the domains of variables increase. Now consider the cases in which we have a set of interventional distributions. The number of joint space parameters for $\boldsymbol{P}_*$ is $d_1 \dots d_n(d_1 \dots d_n - 1)$. This greatly increases the complexity of the already hard problem. In this section, we show three methods to reduce the complexity of our implicitization problem.

www.manaraa.com

5.3.1 Two-step Method

Garcia et al. (2005) proposed a two-step method to compute ker(Ψ) for a BN with hidden variables and non-experimental distribution. It is known that this method usually works faster than direct implicitization. We apply it to our problem in which we have a set of interventional distributions.

Suppose we have a causal BN G with n observed variables $V_1, \ldots, V_n$ and n' unobserved variables $U_1, \ldots, U_{n'}$ and a set of interventional distributions $\boldsymbol{P}_{intv}$ for G. Let Ψ be the ring homomorphism defined in (5.4). We denote $\boldsymbol{P}^U_{intv}$ be the set of joint distributions assuming that all $U_1, \ldots, U_{n'}$ are observed

$$\boldsymbol{P}^U_{intv} = \{P_t(vu) | P_t(v) \in \boldsymbol{P}_{intv}\}. \tag{5.33}$$

Let Φ denote the ring homomorphism

$$\Phi : \mathbb{R}[J_{\boldsymbol{P}^U_{intv}}] \rightarrow \mathbb{R}[M] \tag{5.34}$$

induced by the mapping

$$p^t_{vu} = \prod_{\{i|V_i \notin T\}} q^i_{v_i pa_i u^i} \prod_{j=1}^{n'} r^j_{u_j}. \tag{5.35}$$

For the non-experimental distribution $P(v)$, Garcia et al. (2005) showed that

$$\ker(\Psi) = \ker(\Phi) \cap \mathbb{R}[J_{\{P(v)\}}]. \tag{5.36}$$

It can be naturally extended to the case of arbitrary $\boldsymbol{P}_{intv}$. We have

$$\ker(\Psi) = \ker(\Phi) \cap \mathbb{R}[J_{\boldsymbol{P}_{intv}}]. \tag{5.37}$$

Following Garcia et al. (2005), ker(Ψ) can be computed in two steps. First, we compute ker(Φ) corresponding to the case where all variables are assumed to be observed. Then we compute the subset of ker(Φ) that corresponds to the polynomial constraints on observable distributions. We have implemented our method using a computer algebra system, Singular (Greuel et al., 2005).

5.3.2 Reducing the Implicitization Problem Using Known Constraints

We can reduce the complexity of the implicitization problem by using some known constraints among interventional distributions. Given the set of joint space parameters $J_{\boldsymbol{P}_{intv}}$, suppose that we have

www.manaraa.com

some known constraints among $J_{\boldsymbol{P}_{intv}}$ stating that a joint space parameter p_v^t can be represented as a polynomial function of some other joint space parameters in $J_{\boldsymbol{P}_{intv}} \setminus p_v^t$. Then, the relation reduces the implicitization problem as follows. Let f be a polynomial function such that

$$p_v^t = f(J_{\boldsymbol{P}_{intv}} \setminus p_v^t) \tag{5.38}$$

and let Ψ and Ψ' be two ring homomorphisms

$$\Psi : \mathbb{R}[J_{\boldsymbol{P}_{intv}}] \to \mathbb{R}[M],\ \Psi' : \mathbb{R}[J_{\boldsymbol{P}_{intv}} \setminus p_v^t] \to \mathbb{R}[M]. \tag{5.39}$$

Then, we have

$$\ker(\Psi) = \ker(\Psi') + \langle p_v^t - f(J_{\boldsymbol{P}_{intv}} \setminus p_v^t) \rangle. \tag{5.40}$$

This suggests that the more we find such relations among parameters, the more we can reduce the implicitization problem. The following two lemmas provide a class of such relations.

A *c-component* is a maximal set of vertices such that any two vertices in the set are connected by a path on which every edge is of the form $\dashleftarrow U \dashrightarrow$ where U is a hidden variable. A set $A \subseteq V$ is called an *ancestral set* if it contains its own observed ancestors.

Lemma 4 Tian and Pearl (2002b) *Let $T \subseteq V$ and assume that $V \setminus T$ is partitioned into c-components $H_1, \ldots, H_l$ in the subgraph $G(V \setminus T)$. Then we have*

$$P_t(v) = \prod_i P_{v \setminus h_i}(v). \tag{5.41}$$

Lemma 5 Tian and Pearl (2002b) *Let $C \subseteq T \subseteq V$. If $V \setminus T$ is an ancestral set in $G(V \setminus C)$, then*

$$P_t(v) = \sum_{t \setminus c} P_c(v). \tag{5.42}$$

We give a procedure in Figure 5.2 that lists a set of polynomial relations among $\boldsymbol{P}_{intv}$ based on these two lemmas. Given a set of joint space parameters $J_{\boldsymbol{P}_{intv}}$, it outputs a subset $J'_{\boldsymbol{P}_{intv}}$ of $J_{\boldsymbol{P}_{intv}}$ which contains the joint space parameters that cannot be represented as a polynomial function of other joint space parameters, and the ideal I generated by all the relations found by Lemma 4 and Lemma 5. In Step 1, we look for the parameters that can be represented as the product of other parameters using Lemma 4. In Step 2, we find the parameters that can be represented as the sum of other parameters using Lemma 5. We have the following proposition.

www.manaraa.com

procedure PolyRelations$(G, J_{\boldsymbol{P}_{intv}})$

INPUT: a causal BN G, joint space parameters $J_{\boldsymbol{P}_{intv}}$ associated with a set of interventional distributions $\boldsymbol{P}_{intv}$

OUTPUT: a subset $J'_{\boldsymbol{P}_{intv}} \subseteq J_{\boldsymbol{P}_{intv}}$ of joint space parameters and the ideal I containing polynomial relations among the joint space parameters

Initialization:

$I := \emptyset$

$J'_{\boldsymbol{P}_{intv}} := J_{\boldsymbol{P}_{intv}}$

Step 1:

for each $p_v^t \in J'_{\boldsymbol{P}_{intv}}$ **do**

Let $H_1, \ldots, H_l$ be the c-components in the subgraph $G(V \setminus T)$.

$I := I + \langle p_v^t - \prod_i p_v^{v \setminus h_i} \rangle$

$J'_{\boldsymbol{P}_{intv}} := J'_{\boldsymbol{P}_{intv}} \setminus p_v^t$

end for

Step 2:

for each $p_v^t \in J'_{\boldsymbol{P}_{intv}}$ **do**

if there is a joint space parameter p_v^c that satisfies

(i) $C \subseteq T \subseteq V$

(ii) $V \setminus T$ is an ancestral set in $G(V \setminus C)$

then

$I := I + \langle p_v^t - \sum_{t \setminus c} p_v^c \rangle$

$J'_{\boldsymbol{P}_{intv}} := J'_{\boldsymbol{P}_{intv}} \setminus p_v^t$

end if

end for

Figure 5.2 A procedure for listing polynomial relations among interventional distributions

Proposition 7 *Given a set of interventional distributions $\boldsymbol{P}_{intv}$, a causal BN G with hidden variables and a ring homomorphism Ψ defined in (5.4), let $J'_{\boldsymbol{P}_{intv}}$ and I be the results computed by* **PolyRelations**. *Then,*

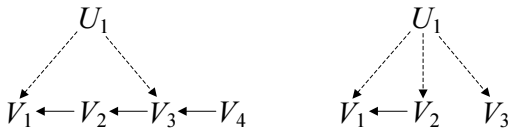
$$\ker(\Psi) = \ker(\Psi') + I \tag{5.43}$$

where Ψ' is a ring homomorphism

$$\Psi' : \mathbb{R}[J'_{\boldsymbol{P}_{intv}}] \to \mathbb{R}[M]. \tag{5.44}$$

To illustrate the procedure, consider a causal BN G with four observed variables V_1, V_2, V_3, V_4 and one hidden variable U_1 in Figure 5.3 (a). We will compute $\ker(\Psi)$ for the set of all interventional

www.manaraa.com



(a) (b)

Figure 5.3 Two causal BNs with one hidden variable

distributions $\boldsymbol{P}_*$ using **PolyRelations**. In Step 1, we find that most of joint space parameters can be represented as the product of other parameters. For example, we have

$$p_v^{v_1} = p_v^{v_1 v_3 v_4} p_v^{v_1 v_2 v_4} p_v^{v_1 v_2 v_3} \tag{5.45}$$

since $V \setminus V_1 = \{V_2, V_3, V_4\}$ is partitioned into three c-components $\{V_2\}$, $\{V_3\}$ and $\{V_4\}$. Also,

$$p_v^{v_2} = p_v^{v_2 v_4} p_v^{v_1 v_2 v_3} \tag{5.46}$$

since $V \setminus V_2 = \{V_1, V_3, V_4\}$ is partitioned into two c-components $\{V_1, V_3\}$ and $\{V_4\}$. The only joint space parameters that do not decompose in Step 1 are

$$p_v^{v_2 v_4}, p_v^{v_1 v_3 v_4}, p_v^{v_1 v_2 v_3}, p_v^{v_2 v_3 v_4} \text{ and } p_v^{v_1 v_2 v_4}. \tag{5.47}$$

Thus, after Step 1 we have

$$J'_{\boldsymbol{P}_{intv}} = J_{\{P_{v_2 v_4}(v), P_{v_1 v_3 v_4}(v), P_{v_1 v_2 v_3}(v), P_{v_2 v_3 v_4}(v), P_{v_1 v_2 v_4}(v): \forall v_1, v_2, v_3, v_4\}}. \tag{5.48}$$

In Step 2, we find that

$$p_v^{v_2 v_3 v_4} = \sum_{v_3} p_v^{v_2 v_4} \text{ and } p_v^{v_1 v_2 v_4} = \sum_{v_1} p_v^{v_2 v_4} \tag{5.49}$$

since $V \setminus \{V_2, V_3, V_4\} = \{V_1\}$ and $V \setminus \{V_1, V_2, V_4\} = \{V_3\}$ are ancestral sets in $G(V \setminus \{V_2, V_4\}) = G(\{V_1, V_3\})$. After Step 2, we have

$$J'_{\boldsymbol{P}_{intv}} = J_{\{P_{v_2 v_4}(v), P_{v_1 v_3 v_4}(v), P_{v_1 v_2 v_3}(v): \forall v_1, v_2, v_3, v_4\}} \tag{5.50}$$

www.manaraa.com

and I is generated by all the relations found in Step 1 and 2. Finally, we have

$$\ker(\Psi) = \ker(\Psi') + I \tag{5.51}$$

where Ψ' is the ring homomorphism

$$\Psi' : \mathbb{R}[J'_{\boldsymbol{P}_{intv}}] \to \mathbb{R}[M]. \tag{5.52}$$

Moreover, we find that $\ker(\Psi')$ can be represented as $\ker(\Psi_1) + \ker(\Psi_2) + \ker(\Psi_3)$ where

$$\Psi_1 : \mathbb{R}[J_{\{P_{v_2 v_4}(v):\forall v_2, v_4\}}] \to \mathbb{R}[M],\ \Psi_2 : \mathbb{R}[J_{\{P_{v_1 v_3 v_4}(v):\forall v_1, v_3, v_4\}}] \to \mathbb{R}[M],\ \Psi_3 : \mathbb{R}[J_{P_{\{v_1 v_2 v_3}(v):\forall v_1, v_2, v_3\}}] \to \mathbb{R}[M] \tag{5.53}$$

since the mappings inducing Ψ_1, Ψ_2 and Ψ_3 do not share model parameters. This gives

$$\ker(\Psi) = \ker(\Psi_1) + \ker(\Psi_2) + \ker(\Psi_3) + I. \tag{5.54}$$

Compared to the original implicitization problem of computing $\ker(\Psi)$ involving 240 joint space parameters which is intractable, we now have three small implicitization problems. Computing $\ker(\Psi_1)$ involves 16 joint space parameters and each of the computation of $\ker(\Psi_2)$ and $\ker(\Psi_3)$ involves 16 joint space parameters. The reduced problem can be solved easily.

Note that $J'_{\boldsymbol{P}_{intv}}$ computed by **PolyRelations** in the above example contains only the joint space parameters related to c-components in G. This holds generally for G in which the subgraph $G(C)$ for each c-component C of G has no edges.

Proposition 8 *Let $C_1, \ldots, C_l$ be c-components of a causal BN G. If every subgraph $G(C_i)$ has no edges, then*

$$\ker(\Psi) = \ker(\Psi_1) + \ldots + \ker(\Psi_l) + I \tag{5.55}$$

where

$$\Psi_i : \mathbb{R}[J_{\{P_{v \setminus c_i}(v):\forall v \setminus c_i\}}] \to \mathbb{R}[M] \tag{5.56}$$

and I is the ideal computed by the procedure **PolyRelations**.

www.manaraa.com

The implicitization problem for a large causal BN G is computationally feasible if G has the structure described in Proposition 8 and the size of each c-component in G is small. Our method becomes infeasible as the size of each c-component grows.

In general, there may be some constraints that are not included in the constraints for each c-component and cannot be found by Lemma 4 and 5. For example, for the causal BN G in Figure 5.3 (b), we find the following constraint by the method in Section 5.3.1 using the Singular system:

$$\begin{aligned} & p_{222}p_{122}^{V_2=2}p_{211}^{V_2=1} + p_{222}p_{122}^{V_2=2}p_{212}^{V_2=1} + p_{212}p_{122}^{V_2=2}p_{221}^{V_2=2} + p_{122}p_{212}^{V_2=1}p_{221}^{V_2=2} + p_{222}p_{212}^{V_2=1}p_{221}^{V_2=2} \\ & - p_{122}^{V_2=2}p_{212}^{V_2=1}p_{221}^{V_2=2} + p_{212}p_{122}^{V_2=2}p_{222}^{V_2=2} - p_{122}p_{211}^{V_2=1}p_{222}^{V_2=2} + p_{222}p_{212}^{V_2=1}p_{222}^{V_2=2} - p_{122}^{V_2=2}p_{212}^{V_2=1}p_{222}^{V_2=2} \\ & + p_{212}p_{221}^{V_2=2}p_{222}^{V_2=2} - p_{212}^{V_2=1}p_{221}^{V_2=2}p_{222}^{V_2=2} + p_{212}p_{222}^{V_2=2}p_{222}^{V_2=2} - p_{212}^{V_2=1}p_{222}^{V_2=2}p_{222}^{V_2=2} - p_{222}p_{212}^{V_2=1} \\ & - p_{212}p_{222}^{V_2=2} + p_{212}^{V_2=1}p_{222}^{V_2=2} \end{aligned} \tag{5.57}$$

which is in $\ker(\Psi)$ but cannot be induced by Lemma 4 and 5.

5.3.3 Constraints in Subgraphs

When the sizes of the c-components of a causal BN are large, it may not be feasible to compute the polynomial constraints by the methods described thus far. Instead, suppose that we wish to test a part (subgraph) of a causal BN assuming that all the conditional independence relations captured by the causal BN are correct. Our goal is to compute constraints (by implicitization) for this subgraph or another subgraph that includes the subgraph with as small number of additional vertices as possible. This can be achieved by finding a subgraph in which the local Markov property (every variable be independent of all its nondescendants conditional on its parents) is satisfied. More formally, given a causal BN G and a subset $S \subseteq V \cup U$, we seek to find the smallest set S^* such that $S \subseteq S^* \subseteq V \cup U$ and for every $X \in S^*$, $X \perp\!\!\!\perp \mathrm{ND}_{S^*}(X) | \mathrm{PA}_{S^*}(X)$ where $\mathrm{ND}_A(X)$ is the set of nondescendants of X in $G(A)$ and $\mathrm{PA}_A(X)$ is the set of parents of X in $G(A)$. It is easy to see that the local Markov property is satisfied for $G(\mathrm{AN}(S))$ where $\mathrm{AN}(S)$ is the union of S and the set of ancestors of the vertices in S. However, there can exist a smaller such set S^* than $\mathrm{AN}(S)$. By these conditional independence relations, we have the

www.manaraa.com

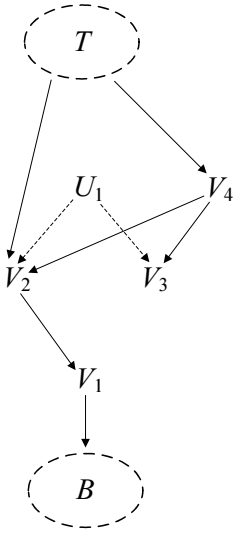


Figure 5.4 Testing a subgraph that includes the vertices V_1, V_2 and V_3

following factorization:

$$P(s^* \cap v) = \sum_{s^* \cap u} \prod_{\{i|V_i \in S^* \cap V\}} P(v_i|\text{pa}_{S^*}(V_i)) \prod_{\{j|U_j \in S^* \cap U\}} P(u_j). \tag{5.58}$$

Given this factorization, the truncated factorizations for interventional distributions are straightforward. These factorizations define the implicitization problem for the subgraph $G(S^*)$, which involves fewer joint space parameters and model parameters than those in the implicitization problem for G. Then, we can test the subgraph $G(S^*)$ using the polynomial constraints computed by the methods described in the previous sections.

The conditional independence relations in a causal BN are specified by the d-separation criterion as defined in the following Pearl (1988). If X, Y and Z are three disjoint subsets of vertices in a DAG, then Z is said to *d-separate* X from Y if along every path between a vertex in X and a vertex in Y there is vertex w satisfying one of the following two conditions: (i) w has converging arrows and none of w or its descendants are in Z, of (ii) w does not have converging arrows and w is in Z. If a path satisfies this condition, it is said to be *blocked*; otherwise, it is said to be *activated* by Z.

Suppose that we wish to test a subgraph including the vertices V_1, V_2 and V_3 of a causal BN G in Figure 5.4 where T and B are subgraphs consisting of a large number of vertices. Consider a subgraph

www.manaraa.com

$G(\{V_1, V_2, V_3, U_1\})$. If the local Markov property were satisfied, then we would have the factorization $P(v_1, v_2, v_3) = \sum_{u_1} P(v_1|v_2)P(v_2|u_1)P(v_3|u_1)P(u_1)$, which would define a new implicitization problem with new sets of joint space parameters and model parameters. However, $V_2 \perp\!\!\!\perp V_3|U_1$ and $V_3 \perp\!\!\!\perp \{V_1, V_2\}|U_1$ do not hold in the entire graph G and the factorization does not follow. This is because there is an activated path between V_2 and V_3: $V_2 \leftarrow V_4 \rightarrow V_3$ (also, there may be another activated path via vertices in T) in G. Hence, we look for some other parents of V_2 (together with U_1) that d-separate V_2 from V_3. We find that $\{U_1, V_4\}$ d-separates V_2 from V_3. In $G(\{V_1, V_2, V_3, V_4, U_1\})$, the local Markov property is satisfied: $V_1 \perp\!\!\!\perp \{V_3, V_4, U_1\}|V_2, V_2 \perp\!\!\!\perp V_3|\{V_4, U_1\}, V_3 \perp\!\!\!\perp \{V_1, V_2\}|\{V_4, U_1\}, V_4 \perp\!\!\!\perp U_1$. Thus, we have the factorization

$$P(v_1, v_2, v_3, v_4) = \sum_{u_1} P(v_1|v_2)P(v_2|v_4u_1)P(v_3|v_4u_1)P(v_4)P(u_1) \tag{5.59}$$

which defines an implicitization problem for the subgraph $G(\{V_1, V_2, V_3, V_4, U_1\})$.

The next lemma provides the basis for finding the smallest subgraph (containing a given subgraph) in which the local Markov property is satisfied.

Lemma 6 *Suppose that $X \perp\!\!\!\perp ND_S(X)|PA_S(X)$ does not hold and $T_1 \subseteq PA_{V\cup U}(X)$ is a minimal set such that $X \perp\!\!\!\perp ND_{S\cup T_1}(X)|PA_{S\cup T_1}(X)$, that is, there is no $T_1' \subseteq T$ such that $X \perp\!\!\!\perp ND_{S\cup T_1'}(X)|PA_{S\cup T_1'}(X)$. Also, suppose that $T_2 \subseteq PA_{V\cup U}(X)$ is a minimal set such that $X \perp\!\!\!\perp ND_{S\cup T_2}(X)|PA_{S\cup T_2}(X)$. Then, $T_1 = T_2$.*

In other words, there is a unique minimal set $T \subseteq \mathrm{PA}_{V\cup U}(X)$ such that $X \perp\!\!\!\perp \mathrm{ND}_{S\cup T}(X)|\mathrm{PA}_{S\cup T}(X)$.
Proof: Suppose for a contradiction that $T_1 \neq T_2$. Then, by the minimality assumption, $T_1 \setminus T_2 \neq \emptyset$ and $T_2 \setminus T_1 \neq \emptyset$. Let α be a vertex in $T_1 \setminus T_2$. Then, there is an activated path (by $\mathrm{PA}_{S\cup(T_1\setminus\alpha)}(X)$) $X \leftarrow \alpha \cdots n$ for some $n \in \mathrm{ND}_S(X)$ (otherwise, T_1 would not be minimal). There are two cases to consider.

(i) For every vertex t in the path $X \leftarrow \alpha \cdots n$, $t \notin T_2 \setminus T_1$. Let $\beta_1, \ldots, \beta_i \in T_1 \setminus T_2$ such that $X \leftarrow \alpha \cdots \beta_1 \cdots \beta_i \cdots n$. Then, the path $X \leftarrow \beta_i \cdots n$ is activated by $\mathrm{PA}_{S\cup T_2}(X)$ since $\beta_i \notin T_2$. This contradicts that $\mathrm{PA}_{S\cup T_2}(X)$ d-separates X from n.

(ii) There is a vertex t in the path $X \leftarrow \alpha \cdots n$ such that $t \in T_2 \setminus T_1$. Let $r_1, \ldots, r_j \in T_2 \setminus T_1$ such that $X \leftarrow \alpha \cdots r_1 \cdots r_j \cdots n$. Then, the path $X \leftarrow r_j \cdots n$ is activated by $\mathrm{PA}_{S\cup T_1}(X)$ since $r_j \notin T_1$. This contradicts that $\mathrm{PA}_{S\cup T_1}(X)$ d-separates X from n.

www.manaraa.com

```
procedure MarkovSubgraph(G,S)
INPUT: a causal BN G, a list S of vertices in G
OUTPUT: an updated list S
while true do
  start_size := S.size()
  i := 1
  while i <= S.size() do
    T := PA(S[i]) \ S
    for each A ∈ PA(S[i]) \ S do
      if S[i] ⫫ ND_S(S[i])|PA(S[i]) \ A then
        T := T \ A
      end if
    end for
    S := S ∪ T
    i := i + 1
  end while
  Break if S.size() == start_size
end while
```

Figure 5.5 A procedure for finding a subgraph in which the local Markov property is satisfied

Hence, it follows that $T_1 = T_2$. ■

We give a procedure called **MarkovSubgraph** in Figure 5.5 that extends a given subgraph so that the local Markov property is satisfied in the extended subgraph. Given a subgraph $G(S)$, **MarkovSubgraph** examines the local Markov property for each vertex $S[i]$ in S. If the condition is not satisfied, a minimal set T of parents of $S[i]$ that d-separates (together with $\mathrm{PA}_S(S[i])$) $S[i]$ from $\mathrm{ND}_S(S[i])$ are added to S. Finding this minimal set can be done by eliminating from $\mathrm{PA}(S[i]) \setminus S$ the vertices that are not necessary for the d-separation. By Lemma 6, for any S' such that $S \subseteq S'$ and the local Markov property is satisfied in $G(S')$, we have that $T \subseteq S'$ (otherwise, such T cannot be unique). Thus, the output S of **MarkovSubgraph** is the smallest set such that the local Markov property is satisfied and (5.58) follows.

www.manaraa.com

5.4 Model Testing Using Polynomial Constraints

In this section, we consider the problem of testing a causal BN G given a data set D using the equality constraints induced by G. D may be either observational or experimental data. To simplify the discussion, we will focus on a single data set D since it will be obvious that the same idea can be applied to a set of experimental data sets. To apply these constraints to finite data in practice, we need to design test statistics for non-independence constraints. However, these non-independence constraints are in general too complex to obtain theoretical test statistics. We use a *bootstrap* method (Efron, 1979) to avoid this difficulty. In particular, we use a parametric bootstrap method, in which a parametric model is fit to the data, by maximum likelihood, and samples are drawn from this parametric model. Then, the estimate of a constraint is calculated from these samples.

In general, we may use any single equality constraint f induced by G as follows. First, we compute the bootstrap distribution of f. Then, we select an appropriate critical region. If the estimate of f on D is in the critical region, we reject G. Our goal is to have small Type I and Type II errors. To this end, we propose to use a set of equality constraints simultaneously by adding the absolute values of the constraints and using it as a single constraint. For example, the causal BN G_1 in Figure 5.6 induces the following constraints: For any v_1 and v_2,

$$\begin{aligned} g(v_1, v_2) = &\sum_{v_3} P(V_1 = v_1 | V_2 = v_2, V_3 = v_3, V_4 = 1) P(V_3 = v_3 | V_4 = 1) \\ &- \sum_{v_3} P(V_1 = v_1 | V_2 = v_2, V_3 = v_3, V_4 = 2) P(V_3 = v_3 | V_4 = 2) \\ =&0 \end{aligned} \tag{5.60}$$

assuming that V_4 is binary. We can combine these equality constraints to form a single equality constraint:

$$f = \sum_{v_1, v_2} \left| g(v_1, v_2) \right| = 0. \tag{5.61}$$

Note that the combined constraint is satisfied if and only if all the original constraints are satisfied. Moreover, a test using the combined constraint is likely to give a smaller Type II error than using the original constraints separately. Suppose that D has been generated by some other causal BN G'. It

www.manaraa.com

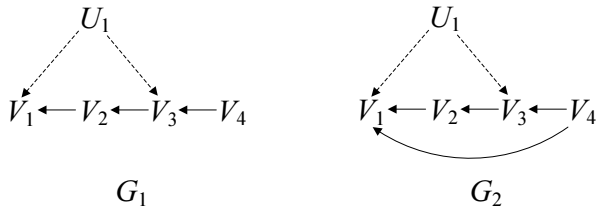


Figure 5.6 Two causal BNs that are Markov equivalent

Testing a causal BN G given a data set D

1. Compute an equality constraint f from G.
2. Learn the parameters of G using the EM algorithm.
3. Sample data sets $D_1^*, D_2^*, \ldots, D_k^*$ from G with the learned parameters.
4. Compute the estimate $f_{D_i^*}$ of f on D_i^* for $i = 1, \ldots, k$.
5. Determine a critical region: Decide a value $\alpha = \text{argmin}_\beta \left| p - \left| \{ f_{D_i^*} | f_{D_i^*} > \beta \} \right| / k \right|$ where p is a significance level.
6. If the estimate f_D of f on D is greater than α, then reject G.

Figure 5.7 A model testing procedure for a causal BN using a polynomial constraint

may be difficult to reject G' based on each constraint, but the combined constraint may give us enough confidence to reject G'. Figure 5.7 describes our proposed model testing procedure.

It is easy to generalize this procedure to a set of experimental data sets $\boldsymbol{D}_{intv}$. We simply replace a data set D_i^* with a set of data sets $\boldsymbol{D}_{intv}^*$ in the procedure. Then, sampling data sets and computing the estimate of the constraint are straightforward.

We now demonstrate our model testing procedure using data sets generated by causal BNs in Figure 5.7. We wish to test a causal BN G_1 against the alternative causal BN G_2 given a data set. We used the constraint f in (5.61) for the test. We generated 150 data sets from G_1 and another 150 data sets from G_2 and measured the Type I and Type II errors of our testing procedure. For Step 3 of the testing

www.manaraa.com

Table 5.1 The Type I and Type II errors in testing G_1 against G_2

	$p = 0.05$		$p = 0.01$	
N	Type I	Type II	Type I	Type II
1000	0.0400	0.0733	0.0200	0.1733
2000	0.0667	0.0667	0.0067	0.1200

procedure, we sampled 100 data sets. We repeated this for two different sizes $N = 1000, 2000$ of each data set. Table 5.1 shows the Type I and Type II errors for two significance levels $p = 0.05, 0.01$.

Note that G_1 and G_2 are Markov equivalent: They induce the same set of conditional independence relations. It is known to be difficult to differentiate Markov equivalent models using only observational data. Our testing procedure provides a way to differentiate Markov equivalent models G_1 and G_2. Now suppose that we wish to select one model from G_1 and G_2 given a data set D. Some scoring functions can be used to evaluate each model. We experimented with the *minimum description length (MDL)* scoring function for this purpose. We select G_1 if the MDL score of G_1 is smaller than that of G_2 and select G_2 otherwise. We measured the error rate of the selection method on 150 data sets from G_1 and another 150 data sets from G_2. Our testing procedure in this section gives an alternative way to select a model from G_1 and G_2: We select G_1 if G_1 is not rejected by our test using the constraint f and select G_2 otherwise. The error rate of our selection method is simply the average of the Type I and Type II errors in Table 5.1. Table 5.2 compares the error rates of the two methods. In this experiment, our method based on a polynomial constraint clearly outperformed the MDL-based method, which was not better than random.

This model selection method based on a polynomial constraint is not easily generalized to selecting a model from more than two models. A difficulty is that there may be multiple constraints, each of which is induced by a distinct model. How to use these multiple constraints to select a single model from a set of candidate models needs further study. However, if you do not need to select a single model, our testing procedure can be used to reduce the size of the set of candidate models.

www.manaraa.com

Table 5.2 Comparison of the error rates of two model selection methods

N	Constraint f		MDL
	$p = 0.05$	$p = 0.01$	
1000	0.0567	0.0967	0.5500
2000	0.0667	0.0634	0.5767

www.manaraa.com

CHAPTER 6. INEQUALITY CONSTRAINTS IN CAUSAL BAYESIAN NETWORKS

We present a class of inequality constraints on the set of distributions induced by local interventions on variables governed by a causal Bayesian network, in which some of the variables remain unmeasured. We derive bounds on causal effects that are not directly measured in randomized experiments. We derive instrumental inequality type of constraints on nonexperimental distributions. The results have applications in testing causal models with observational or experimental data.

6.1 Constraints on Interventional Distributions

Let $\boldsymbol{P}_*$ denote the set of all interventional distributions induced by a given semi-Markovian model,

$$\boldsymbol{P}_* = \{P_t(v) | T \subseteq V, t \in Dm(T), v \in Dm(V)\} \tag{6.1}$$

where $Dm(T)$ represents the domain of T. What are the constraints imposed by the model on the interventional distributions in $\boldsymbol{P}_*$? The structure of the causal graph G will play an important role in finding these constraints. A *c-component* is a maximal set of vertices such that any two vertices in the set are connected by a path on which every edge is of the form $\dashleftarrow U \dashrightarrow$ where U is a hidden variable. The set of variables V is then partitioned into a set of c-components. For example, the causal graph G in Figure 6.1 consists of two c-components $\{X, Y, Z\}$ and $\{W_1, W_2\}$.

Let $G(H)$ denote the subgraph of G composed only of the variables in H and the hidden variables that are ancestors of H. In general, equality constraints on the set of interventional distributions can be found using the following three lemmas.

Lemma 7 Tian and Pearl (2002b) *Let $H \subseteq V$, and assume that H is partitioned into c-components $H_1, \ldots, H_l$ in the subgraph $G(H)$. Then we have*

www.manaraa.com

(i) $P_{v\setminus h}(v)$ *decomposes as*

$$P_{v\setminus h}(v) = \prod_i P_{v\setminus h_i}(v). \tag{6.2}$$

(ii) Let k *be the number of variables in* H*, and let a topological order of the variables in* H *be* $V_{h_1} < \ldots < V_{h_k}$ *in* $G(H)$*. Let* $H^{(i)} = \{V_{h_1}, \ldots, V_{h_i}\}$ *be the set of variables in* H *ordered before* V_{h_i} *(including* V_{h_i}*),* $i = 1, \ldots, k$*, and* $H^{(0)} = \emptyset$*. Then each* $P_{v\setminus h_j}(v)$*,* $j = 1, \ldots, l$*, is computable from* $P_{v\setminus h}(v)$ *and is given by*

$$P_{v\setminus h_j}(v) = \prod_{\{i|V_{h_i}\in H_j\}} \frac{P_{v\setminus h^{(i)}}(v)}{P_{v\setminus h^{(i-1)}}(v)}, \tag{6.3}$$

where each $P_{v\setminus h^{(i)}}(v)$*,* $i = 0, 1, \ldots, k$*, is given by*

$$P_{v\setminus h^{(i)}}(v) = \sum_{h\setminus h^{(i)}} P_{v\setminus h}(v). \tag{6.4}$$

A special case of Lemma 7 is when $H = V$, and we have the following Lemma.

Lemma 8 Tian and Pearl (2002b) *Assuming that* V *is partitioned into c-components* $S_1, \ldots, S_k$*, we have*

(i) $P(v) = \prod_i P_{v\setminus s_i}(v)$.

(ii) Let a topological order over V *be* $V_1 < \ldots < V_n$*, and let* $V^{(i)} = \{V_1, \ldots, V_i\}, i = 1, \ldots, n$*, and* $V^{(0)} = \emptyset$*. Then each* $P_{v\setminus s_j}(v)$*,* $j = 1, \ldots, k$*, is computable from* $P(v)$ *and is given by*

$$P_{v\setminus s_j}(v) = \prod_{\{i|V_i\in S_j\}} P(v_i|v^{(i-1)}). \tag{6.5}$$

The next lemma provides a condition under which we can compute $P_{v\setminus w}(w)$ from $P_{v\setminus c}(c)$ where W is a subset of C, by simply summing $P_{v\setminus c}(c)$ over other variables $C \setminus W$.

Lemma 9 Tian and Pearl (2002b) *Let* $W \subseteq C \subseteq V$*, and* $W' = C \setminus W$*. If* W *contains its own observed ancestors in* $G(C)$*, then*

$$\sum_{w'} P_{v\setminus c}(v) = P_{v\setminus w}(v). \tag{6.6}$$

The set of equality constraints implied by these three lemmas can be systematically listed by slightly modifying the procedure in Tian and Pearl (2002b) for listing equality constraints on nonexperimental

www.manaraa.com

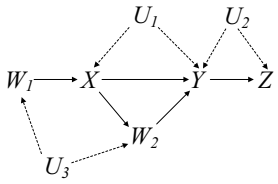


Figure 6.1 U_1,U_2 and U_3 are hidden variables.

distributions. We will not show the details of the procedure here since the focus of this chapter is on inequality constraints.

For example, the model in Figure 2.1 imposes the following equality constraints.

$$P_z(xy) = P(xy|z) \tag{6.7}$$

$$P_{yz}(x) = P(x|z) \tag{6.8}$$

$$P_{xz}(y) = P_x(y) \tag{6.9}$$

The model in Figure 6.1 imposes the following equality constraints.

$$P_{w_1w_2}(xyz) = P(z|w_1xw_2y)P(y|w_1xw_2)P(x|w_1) \tag{6.10}$$

$$P_{w_1w_2z}(xy) = P(y|w_1xw_2)P(x|w_1) \tag{6.11}$$

$$P_{w_1w_2y}(xz) = P_{w_1y}(xz) \tag{6.12}$$

$$P_{w_1w_2x}(yz) = P_{w_2x}(yz) \tag{6.13}$$

$$P_{w_1w_2yz}(x) = P(x|w_1) \tag{6.14}$$

$$P_{w_1w_2xz}(y) = P_{w_2x}(y) \tag{6.15}$$

$$P_{w_1w_2xy}(z) = P_y(z) \tag{6.16}$$

$$P_{xyz}(w_1w_2) = P(w_2|w_1x)P(w_1) \tag{6.17}$$

$$P_{xyzw_2}(w_1) = P(w_1) \tag{6.18}$$

$$P_{xyzw_1}(w_2) = \sum_{w_1} P(w_2|w_1x)P(w_1) \tag{6.19}$$

www.manaraa.com

6.1.1 Inequality Constraints

In this chapter, we are concerned with inequality constraints imposed by a model. The $\boldsymbol{P}_*$ set induced from a semi-Markovian model must satisfy the following inequality constraints.

Lemma 10 *For any $S_1 \subseteq V$ and any superset $S'_1 \subseteq V$ of S_1, we have*

$$\sum_{S_2 \subseteq S'_1 \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \quad \forall v \in Dm(V) \tag{6.20}$$

where $|S_2|$ represents the number of variables in S_2.

Proof: We use the following equation.

$$\prod_{i=1}^{k}(1-a_i) = 1 - \sum_i a_i + \sum_{i,j} a_i a_j - \ldots + (-1)^k a_1 \ldots a_k. \tag{6.21}$$

Take $a_j = P(v_j|pa_j, u^j)$, we have that

$$\sum_u \prod_{\{i|V_i \in S_1\}} P(v_i|pa_i, u^i) \prod_{\{j|V_j \in S'_1 \setminus S_1\}} (1 - P(v_j|pa_j, u^j))P(u) = \sum_{S_2 \subseteq S'_1 \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0 \tag{6.22}$$

since for all $V_i \in V$

$$0 \leq P(v_i|pa_i, u^i) \leq 1. \tag{6.23}$$

■

For a fixed S'_1 set, there are $2^{|S'_1|}$ number of Eq. (6.20) type of inequalities. For different S'_1 sets, some of those inequalities may imply others as shown in the following proposition.

Proposition 9 *If $S'_1 \subset S''_1$, then the set of $2^{|S''_1|}$ inequalities, $\forall S_1 \subseteq S''_1$,*

$$\sum_{S_2 \subseteq S''_1 \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \quad \forall v \in Dm(V) \tag{6.24}$$

imply the set of $2^{|S'_1|}$ inequalities, $\forall S_1 \subseteq S'_1$,

$$\sum_{S_2 \subseteq S'_1 \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \quad \forall v \in Dm(V) \tag{6.25}$$

Assume that the set of variables V in the model is partitioned into c-components $T_1, \ldots, T_k$. Due to the equality constraints given in Lemma 7, instead of listing $2^{|V|}$ Eq. (6.20) type of inequalities, we only need to give $2^{|T_i|}$ Eq. (6.20) type of inequalities for each c-component T_i.

www.manaraa.com

Proposition 10 *Let the set of variables V in a semi-Markovian model be partitioned into c-components* $T_1, \ldots, T_k$. *The* $\mathbf{P}_*$ *set must satisfy the following inequality constraints: for* $i = 1, \ldots, k$, $\forall S_1 \subseteq T_i$,

$$\sum_{S_2 \subseteq T_i \setminus S_1} (-1)^{|S_2|} P_{v \setminus (s_1 \cup s_2)}(v) \geq 0, \quad \forall v \in Dm(V) \tag{6.26}$$

For example, Proposition 10 gives the following inequality constraints for the model in Figure 2.1,

$$1 - P_{yz}(x) - P_{xz}(y) + P_z(xy) \geq 0 \tag{6.27}$$

$$P_{yz}(x) - P_z(xy) \geq 0 \tag{6.28}$$

$$P_{xz}(y) - P_z(xy) \geq 0 \tag{6.29}$$

$$P_z(xy) \geq 0, \tag{6.30}$$

in which (6.30) is trivial, and (6.28) becomes trivial because of equality constraints (6.7) and (6.8).

For the model in Figure 6.1, Proposition 10 gives the following inequality constraints for the c-component $\{X, Y, Z\}$,

$$1 - P_{w_1 w_2 yz}(x) - P_{w_1 w_2 xz}(y) - P_{w_1 w_2 xy}(z) + P_{w_1 w_2 z}(xy) + P_{w_1 w_2 y}(xz) + P_{w_1 w_2 x}(yz) - P_{w_1 w_2}(xyz) \geq 0 \tag{6.31}$$

$$P_{w_1 w_2 yz}(x) - P_{w_1 w_2 z}(xy) - P_{w_1 w_2 y}(xz) + P_{w_1 w_2}(xyz) \geq 0 \tag{6.32}$$

$$P_{w_1 w_2 xz}(y) - P_{w_1 w_2 z}(xy) - P_{w_1 w_2 x}(yz) + P_{w_1 w_2}(xyz) \geq 0 \tag{6.33}$$

$$P_{w_1 w_2 xy}(z) - P_{w_1 w_2 y}(xz) - P_{w_1 w_2 x}(yz) + P_{w_1 w_2}(xyz) \geq 0 \tag{6.34}$$

$$P_{w_1 w_2 z}(xy) - P_{w_1 w_2}(xyz) \geq 0 \tag{6.35}$$

$$P_{w_1 w_2 y}(xz) - P_{w_1 w_2}(xyz) \geq 0 \tag{6.36}$$

$$P_{w_1 w_2 x}(yz) - P_{w_1 w_2}(xyz) \geq 0 \tag{6.37}$$

$$P_{w_1 w_2}(xyz) \geq 0, \tag{6.38}$$

some of which are implied by the set of equality constraints (6.10)-(6.19). It can be shown that all inequality constraints for c-component $\{W_1, W_2\}$ are implied by equality constraints.

Note that in general, the inequality constraints given in this section are not the complete set of constraints that are implied by a given model. For example, for the model given in Figure 2.1, the sharp

www.manaraa.com

bounds on $P_x(y)$ given in Balke and Pearl (1994) for X, Y, and Z being binary variables are not implied by (6.27)-(6.30).

6.2 Inequality Constraints On a Subset of Interventional Distributions

Proposition 10 gives a set of inequality constraints on the set of interventional distributions in $\boldsymbol{P}_*$. In practical situations, we may be interested in constraints involving only a certain subset of interventional distributions. For example, (i) We have done some experiments, and obtained $P_s(v)$ for some sets S. We want to know whether these data are compatible with the given model. For this purpose, we would like inequality constraints involving only those known interventional distributions; (ii) A special case of (i) is that we only have the nonexperimental distribution $P(v)$. We want inequality constraints on $P(v)$ imposed by the model; (iii) In practice, certain experiments may be difficult or expensive to perform. Still, we want some information about a particular causal effect, given some known interventional distributions and nonexperimental distribution. We may provide bounds on concerned causal effect that can be derived from those inequality constraints (if this causal effect is not computable from given quantity through equality constraints).

To restrict the set of inequality constraints given in Proposition 10 to constraints involving only certain subset of interventional distributions, in principle, we can treat each $P_s(v)$ for an instantiation of $v \in Dm(V)$ as a variable, and solve the inequalities to eliminate unwanted variables using methods like Fourier-Motzkin elimination or quantifier elimination. However, this is typically only practical for small number of binary variables due to high computational complexity. In this chapter, we show some inequality constraints involving only interventional distributions of interests that can be derived from those in Proposition 10. In general, these constraints may not include all the possible constraints that could be derived from Proposition 10 in principle.

Instead of directly solving the inequality constraints given in Proposition 10, we consider the inequality in Eq. (6.20) for every $S_1' \subseteq T_i$. We keep every inequality that involves only the interventional distributions of interests. Those inequalities that contain unwanted interventional distributions may lead to some new inequalities. For example, in the model in Figure 6.1, consider the following inequality

www.manaraa.com

that follows from (6.20) with $S_1 = \{Z\}$ and $S_1' = \{Y, Z\}$,

$$P_{w_1 w_2 xy}(z) - P_{w_1 w_2 x}(yz) \geq 0. \tag{6.39}$$

Suppose we want constraints on $P_{w_1 w_2 x}(yz)$ and get rid of unknown quantity $P_{w_1 w_2 xy}(z)$. First we have equality constraints (6.13) and (6.16), and Eq. (6.39) becomes

$$P_{w_2 x}(yz) \leq P_y(z) \tag{6.40}$$

$P_{w_2 x}(yz)$ is a function of W_2 and X but $P_y(z)$ is not, which leads to

$$\max_{w_2, x} P_{w_2 x}(yz) \leq P_y(z) \tag{6.41}$$

$$\sum_z \max_{w_2, x} P_{w_2 x}(yz) \leq 1 \tag{6.42}$$

Eq. (6.42) is a nontrivial inequality constraint on $P_{w_1 w_2 x}(yz) = P_{w_2 x}(yz)$, which can also be represented as

$$P_{w_2 x}(yz_0) + P_{w_2' x'}(yz_1) \leq 1 \tag{6.43}$$

for any $w_2 \in Dm(W_2)$, $x \in Dm(X)$, $w_2' \in Dm(W_2)$ and $x' \in Dm(X)$ when Z is binary ($Dm(Z) = \{z_0, z_1\}$).

From the above considerations, we give a procedure in Figure 6.2 that lists the inequality constraints on the interventional distributions of interest. The procedure has a complexity of $3^{2|T_i|}$. Note that A will always contain the nonexperimental distribution and all interventional distributions that can be computed from $P(v)$ (via equality constraints).

In Step 1, we list the inequalities that do not involve unwanted quantities (i.e., interventional distributions not included in A). Note that we remove some redundant inequalities based on the following lemma.

Lemma 11 *Let $Sup(S_1)$ denote the set of supersets of S_1 such that $S_1' \in Sup(S_1)$ if and only if every interventional distribution in $e_{S_1, S_1'} = \sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v \setminus (s_1 \cup s_2)}(v) \geq 0$ is in A. For a set of sets W, let $Max(W) = \{S | S \in W$, there is no $S' \in W$ such that $S \subset S'\}$ denote the set of maximal sets in W. Then, the set of inequalities*

$$\forall S_1 \subseteq T_i, \forall S_1' \in Max(Sup(S_1)),$$
$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v \setminus (s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V) \tag{6.44}$$

www.manaraa.com

procedure FindIneqs(G,A)
INPUT: a causal graph G, interventional distributions of interest A, equality constraints implied by G
OUTPUT: inequalities of interests, IE_{T_i} for each c-component T_i, $i = 1, \ldots, k$
Step 1:
For each c-component T_i, $i = 1, \ldots, k$
  For each $S_1 \subseteq T_i$ (small to large)
    For each $S_1' \subseteq T_i$ such that $S_1 \subseteq S_1'$(small to large)
      Study the inequality
      $e_{S_1,S_1'} = \sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0$
      If every interventional distribution in $e_{S_1,S_1'}$ is in A
        $IE_{T_i} = IE_{T_i} \cup \{e_{S_1,S_1'} \geq 0\}$;
        Remove any $e_{S_1,R}$ in IE_{T_i} such that $R \subset S_1'$;
Step 2:
For each c-component T_i, $i = 1, \ldots, k$
  For each $S_1 \subseteq T_i$ (small to large)
    For each $S_1' \subseteq T_i$ such that $S_1 \subseteq S_1'$(small to large)
      Study the inequality
      $e_{S_1,S_1'} = \sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0$
      If some interventional distribution in $e_{S_1,S_1'}$ is
      not in A
        $IE_{T_i} = IE_{T_i} \cup \{e_{S_1,S_1'} \geq 0$ reformulated
        in the form of (6.54)$\}$;

Figure 6.2 A Procedure for Listing Inequality Constraints On a Subset of Interventional Distributions

imply the inequalities

$$\forall S_1 \subseteq T_i, \forall S_1' \in Sup(S_1)$$
$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V). \tag{6.45}$$

Proof: We will show that if the inequalities in (6.44) hold, then for any $n \leq |V|$ we have

$$\forall S_1 \subseteq T_i, \forall S_1' \in Max^n(Sup(S_1)),$$
$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V) \tag{6.46}$$

where $Max^n(S) = Max(S \setminus \{R | R \in S, |R| > n\})$. (6.45) will follow from (6.46) if we let n be the size of the set S_1' in (6.45). Assuming (6.44), we prove (6.46) by induction on n.

Base: $n = |V|$. (6.46) is equivalent to (6.44).

www.manaraa.com

Hypothesis: Assume that

$$\forall S_1 \subseteq T_i, \forall S_1' \in Max^n(Sup(S_1)),$$
$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V). \tag{6.47}$$

Induction step: We show that

$$\forall S_1 \subseteq T_i, \forall S_1' \in Max^{n-1}(Sup(S_1)),$$
$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V). \tag{6.48}$$

If $|S_1'| < n-1$, then S_1' is in $Max^n(Sup(S_1))$. Thus, (6.48) follows from (6.47). If $|S_1'| = n-1$, then one of the followings should hold.

Case 1: S_1' is in $Max^n(Sup(S_1))$.

Case 2: There exists a variable α such that $S_1' \cup \{\alpha\}$ is in $Max^n(Sup(S_1))$.

In Case 1, (6.48) follows from (6.47). In Case 2, we have

$$\sum_{S_2 \subseteq (S_1' \cup \{\alpha\}) \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup s_2)}(v) \geq 0, \forall v \in Dm(V) \tag{6.49}$$

and

$$\sum_{S_2 \subseteq S_1' \setminus S_1} (-1)^{|S_2|} P_{v\setminus(s_1 \cup \{\alpha\} \cup s_2)}(v) \geq 0, \forall v \in Dm(V). \tag{6.50}$$

(6.50) follows from (6.47) since $S_1' \cup \{\alpha\}$ is in $Max^n(Sup(S_1 \cup \{\alpha\}))$. Summing (6.49) and (6.50) gives (6.48). ■

In Step 2, we deal with the inequalities that contain unwanted quantities as follows. We rewrite the inequality in Eq. (6.20) as $e_{S_1,S_1'} \geq 0$, with

$$e_{S_1,S_1'} = \sum_{R \in W_1} (-1)^{|R|-|S_1|} P_{v\setminus r}(v) + \sum_{R \in W_2} (-1)^{|R|-|S_1|} P_{v\setminus r}(v) \tag{6.51}$$

where $W_1 = \{S_1 \cup S_2 | S_2 \subseteq S_1' \setminus S_1, P_{v\setminus(s_1 \cup s_2)}(v)$ is in $A\}$ and $W_2 = \{S_1 \cup S_2 | S_2 \subseteq S_1' \setminus S_1, P_{v\setminus(s_1 \cup s_2)}(v)$ is not in $A\}$. We have

$$\sum_{R \in W_1} (-1)^{|R|-|S_1|} P_{v\setminus r}(v) \geq -\sum_{R \in W_2} (-1)^{|R|-|S_1|} P_{v\setminus r}(v). \tag{6.52}$$

www.manaraa.com

Suppose the left-hand side is a function of variables E_1 and the right-hand side is a function of variables E_2. Then,

$$\min_{E_1 \setminus E_2} \sum_{R \in W_1} (-1)^{|R|-|S_1|} P_{v \setminus r}(v) \geq - \sum_{R \in W_2} (-1)^{|R|-|S_1|} P_{v \setminus r}(r). \tag{6.53}$$

Let $Q = \bigcup_{R \in W_2} R$. We obtain,

$$\sum_{Q} \min_{E_1 \setminus E_2} \sum_{R \in W_1} (-1)^{|R|-|S_1|} P_{v \setminus r}(v) \geq - \sum_{R \in W_2} (-1)^{|R|-|S_1|} \prod_{\{i | V_i \in Q \setminus R\}} |Dm(V_i)|. \tag{6.54}$$

Note that if $E_1 \setminus E_2 = \emptyset$, then we do not need $\min_{E_1 \setminus E_2}$.

To illustrate the procedure, suppose we want to get the inequality constraints on the two interventional distributions $P_{w_1 w_2 xy}(z)$ and $P_{w_1 w_2 x}(yz)$ and we are given a tried interventional distribution $P_{w_1 w_2 y}(xz)$.

In Step 1, consider the loop in which $T_i = \{X, Y, Z\}$ and $S_1 = \{\emptyset\}$. The procedure first adds $e_{\emptyset,\{X\}}$ and $e_{\emptyset,\{Z\}}$. When it adds $e_{\emptyset,\{X,Z\}}$, it will remove $e_{\emptyset,\{X\}}$ and $e_{\emptyset,\{Z\}}$ from IE_{T_i} and keep $e_{\emptyset,\{X,Z\}}$ which turns out to be $Max(Sup(\emptyset))$. This repeats for every $S_1 \subseteq T_i$.

In Step 2, consider the loop where $T_i = \{X, Y, Z\}$ and $S_1 = \{Y\}$. The procedure studies $e_{S_1,S_1'}$ for each $S_1' \in \{\{Y\}, \{X, Y\}, \{Y, Z\}, \{X, Y, Z\}\}$. For example, for $S_1' = \{X, Y, Z\}$, we have the inequality (6.33). From (6.10), (6.11), (6.13) and (6.15), we obtain

$$\max_{w_1,z} \Big(P(y|w_1 x w_2) P(x|w_1) + P_{w_2 x}(yz) - P(z|w_1 x w_2 y) P(y|w_1 x w_2) P(x|w_1) \Big) \leq P_{w_2 x}(y). \tag{6.55}$$

Summing both sides over Y gives

$$\sum_{y} \max_{w_1,z} \Big(P(y|w_1 x w_2) P(x|w_1) + P_{w_2 x}(yz) - P(z|w_1 x w_2 y) P(y|w_1 x w_2) P(x|w_1) \Big) \leq 1. \tag{6.56}$$

6.2.1 Bounds on Causal Effects

Suppose that our goal is to bound a particular interventional distribution. For this case, A in the procedure **FindIneqs** consists of the particular interventional distribution that we want to bound, the nonexperimental distribution $P(v)$, and all interventional distributions that are computable from $P(v)$.

For example, consider the graph in Figure 6.1. Suppose that we want to bound the interventional distribution $P_{w_1 w_2 xy}(z)$ and that the interventional distribution $P_{w_1 w_2 y}(xz)$ is available from experiments.

www.manaraa.com

FindIneqs will list the following bounds for $P_{w_1 w_2 xy}(z)$ in Step 1.

$$1 - P(x|w_1) - P_{w_1 w_2 xy}(z) + P_{w_1 w_2 y}(xz) \geq 0 \tag{6.57}$$

$$P_{w_1 w_2 xy}(z) - P_{w_1 w_2 y}(xz) \geq 0 \tag{6.58}$$

which provides a lower and upper bound for $P_{w_1 w_2 xy}(z)$ respectively.

6.2.2 Inequality Constraints on Nonexperimental Distribution

Now assume that we want to find inequality constraints on nonexperimental distribution. For this case, A in the procedure **FindIneqs** consists of the nonexperimental distribution $P(v)$ and all interventional distributions that are computable from $P(v)$.

The inequality constraints produced by **FindIneqs** in this case include the instrumental inequality type of constraints. Consider the graph in Figure 2.1. For the c-component $\{X, Y\}$, **FindIneqs** will produce the inequality (6.29). From (6.7) and (6.9), we have

$$\max_z P(xy|z) \leq P_x(y) \tag{6.59}$$

and summing both sides over Y gives

$$\sum_y \max_z P(xy|z) \leq 1. \tag{6.60}$$

Since this must hold for all X, we obtain the instrumental inequality (2.1).

To illustrate more general instrumental inequality type of constraints, consider the graph in Figure 6.1. For $S_1 = \{Y, Z\}$ and $S_1' = \{X, Y, Z\}$, **FindIneqs** produces the inequality (6.37). From (6.10) and (6.13), we have

$$\max_{w_1} P(z|w_1 x w_2 y) P(y|w_1 x w_2) P(x|w_1) \leq P_{w_2 x}(yz). \tag{6.61}$$

Summing both sides over Y and Z gives

$$\sum_{yz} \max_{w_1} P(z|w_1 x w_2 y) P(y|w_1 x w_2) P(x|w_1) \leq 1. \tag{6.62}$$

www.manaraa.com

CHAPTER 7. CONCLUSION

This chapter provides a broad summary of our work and proposes several potential directions of future work.

7.1 Markov Properties for Linear Causal Models with Correlated Errors

We present local Markov properties for ADMGs representing linear SEMs with correlated errors. The results have applications in testing linear SEMs against the data by testing for zero partial correlations implied by the model. For general linear SEMs with correlated errors, we provide a procedure that lists a subset of zero partial correlations that will imply all other zero partial correlations implied by the model. In particular, for a class of models whose corresponding path diagrams contain no directed mixed cycles, this subset invokes one zero partial correlation for each pair of variables.

In general, our procedure may invoke an exponential number of zero partial correlations if the path diagram G satisfies all of the following properties: (i) G has large c-components; (ii) the vertices in each c-component are heavily connected by bi-directed edges; and (iii) G has directed mixed cycles. If one of these properties is not satisfied, then the number of zero partial correlations derived by our method is typically not exponential.

For the class of MAGs, which is a strict superclass of ADMGs without directed mixed cycles, one might use the pairwise Markov property for MAGs given in Richardson and Spirtes (2002) instead of our results in Section 4.3. However, when the two approaches give a similar number of constraints, it may be better to use our approach since it may use smaller conditioning sets as shown in the example in Section 4.3.2.

The potential advantages of testing linear SEMs based on vanishing partial correlations over the classical test method based on maximum likelihood estimation of the covariance matrix have been

www.manaraa.com

discussed in Pearl (1998); Shipley (2000); McDonald (2002); Shipley (2003). The results presented in this paper provide a theoretical foundation for the practical applications of this test method in linear SEMs with correlated errors. How to implement this test method in practice still needs further study as it requires multiple testing of hypotheses about zero partial correlations (Shipley, 2000; Drton and Perlman, 2007). We also note that, in linear SEMs *without* correlated errors, all the constraints on the covariance matrix are implied by vanishing partial correlations. This also holds in linear SEMs *with* correlated errors that are represented by ADMGs *without* directed mixed cycles. However, it is possible that linear SEMs *with* correlated errors represented by ADMGs *with* directed mixed cycles may imply constraints on the covariance matrix that are not implied by zero partial correlations.

Although the intended application is in linear SEMs, the local Markov properties presented in the paper are valid for ADMGs associated with any probability distributions that satisfy the composition axiom. For example, any probability distribution that is faithful[1] to some DAG or undirected graph (and the marginals of the distribution) satisfies the composition axiom.

Model debugging for ADMGs using vanishing partial correlations is another area of current research. In this model debugging problem, the goal is to modify a graph based on the pattern of rejected hypotheses. The properties of ADMGs presented in this paper may facilitate the development of a new model debugging method.

7.2 Polynomial Constraints in Causal Bayesian Networks

We obtain polynomial constraints on the interventional distributions induced by a causal BN with hidden variables, via the implicitization procedure. These constraints constitute a necessary test for a causal model to be compatible with given observational and experimental data. We present a model testing procedure using theses polynomial constraints.

Future work will investigate the general characterization of the constraints computed by implicitization for causal BNs without hidden variables, which will be helpful in finding the algebraic structure of the constraints implied by causal BNs with hidden variables which typically have complicated structures.

[1] A probability distribution P is said to be faithful to a graph G if all the conditional independence relations embedded in P are encoded in G (via the global Markov property).

www.manaraa.com

7.3 Inequality Constraints in Causal Bayesian Networks

We present a class of inequality constraints imposed by a given causal BN with hidden variables on the set of interventional distributions that can be induced from the network. We show a method to restrict these inequality constraints on to that only involving interventional distributions of interests. These inequality constraints can be used as necessary test for a causal model to be compatible with given observational and experimental data. Another application permits us to bound the effects of untried interventions from experiments involving auxiliary interventions that are easier or cheaper to implement.

We derive a type of inequality constraints upon the nonexperimental distribution in a complexity of 3^{2m} where m is the number of variables in the largest c-component. These constraints are imposed by the network structure, regardless of the number of states of the (observed or hidden) variables involved. These constraints can be used to test a model or distinguish between models. How to test these inequality constraints in practice and use them for model selection would be interesting future research.

www.manaraa.com

BIBLIOGRAPHY

Andersson, S., Madigan, D., and Perlman, M. (2001). Alternative markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28:33–86.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds, and applications. In de Mantaras, R. L. and Poole, D., editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA.

Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley, New York.

Bonet, B. (2001). Instrumentality tests revisited. In *Proc. 17th Conf. on Uncertainty in Artificial Intelligence*, pages 48–55, Seattle, WA. Morgan Kaufmann.

Chickering, D. and Pearl, J. (1996). A clinician's apprentice for analyzing non-compliance. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume II, pages 1269–1276. MIT Press, Menlo Park, CA.

Cox, D., Little, J., and O'Shea, D. (1996). *Ideals, Varieties and Algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer-Verlag, New York.

Cox, D. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218.

Desjardins, B. (1999). *On the theoretical limits to reliable causal inference*. PhD dissertation, University of Pittsburgh.

Drton, M. and Perlman, M. (2007). Multiple testing and error control in gaussian graphical model selection. *Statistical Science*, 22(3):430–449.

www.manaraa.com

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*, 17:333–353.

Garcia, L., Stillman, M., and Sturmfels, B. (2005). Algebraic geometry of bayesian networks. *Journal of Symbolic Computation*, 39(3–4):331–355.

Garcia, L. D. (2004). Algebraic statistics in model selection. In *Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 177–18, Arlington, Virginia. AUAI Press.

Geiger, D. and Meek, C. (1998). Graphical models and exponential families. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 156–165, San Francisco, CA. Morgan Kaufmann Publishers.

Geiger, D. and Meek, C. (1999). Quantifier elimination for statistical problems. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 226–235, San Francisco, CA. Morgan Kaufmann Publishers.

Goldberger, A. (1972). Structural equation models in the social sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001.

Greuel, G.-M., Pfister, G., and Schönemann, H. (2005). SINGULAR 3.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern. `http://www.singular.uni-kl.de`.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

www.manaraa.com

Kang, C. and Tian, J. (2006). Inequality constraints in causal models with hidden variables. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 233–240, Arlington, Virginia. AUAI Press.

Kang, C. and Tian, J. (2007). Polynomial constraints in causal Bayesian networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pages 200–208, Arlington, Virginia. AUAI Press.

Kauermann, G. (1996). On a dualization of graphical gaussian models. *Scandinavian Journal of Statistics*, 23:105–116.

Koster, J. (1999). On the validity of the Markov interpretation of path diagrams of gaussian structural equations systems with correlated errors. *Scandinavian Journal of Statistics*, 26:413–431.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S., Dawid, A., Larsen, B., and Leimer, H. (1990). Independence properties of directed Markov fields. *Networks*, 20:491–505.

Lauritzen, S. and Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57.

McDonald, R. (2002). What can we learn from the path equations?: Identifiability, constraints, equivalence. *Psychometrika*, 67(2):225–249.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligence Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In Besnard, P. and Hanks, S., editors, *Uncertainty in Artificial Intelligence 11*, pages 435–443. Morgan Kaufmann.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Socioligical Methods and Research*, 27:226–284.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY.

www.manaraa.com

Pearl, J., Geiger, D., and Verma, T. (1990). The logic of influence diagrams. In Oliver, R. and Smith, J., editors, *Influence Diagrams, Belief Nets and Decision Analysis*, pages 67–87. John Wiley and Sons, Inc., New York, NY.

Pearl, J. and Meshkat, P. (1999). Testing regression models with fewer regressors. In *Proceedings of AI-STAT*, pages 255–259.

Riccomagno, E. and Smith, J. (2003). Non-graphical causality: a generalization of the concept of a total cause. Technical Report No. 394, Department of Statistics, University of Warwick.

Riccomagno, E. and Smith, J. (2004). Identifying a cause in models which are not simple bayesian networks. In *Proceedings of IPMU*, pages 1315–1322, Perugia.

Richardson, T. (2003). Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157.

Richardson, T. and Spirtes, P. (2002). Ancestral graph markov models. *Annals of Statistics*, 30(4):962–1030.

Robins, J. M. and Wasserman, L. A. (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–97)*, pages 409–420, San Francisco, CA. Morgan Kaufmann Publishers.

Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, 7:206–218.

Shipley, B. (2003). Testing recursive path models with correlated errors using d-separation. *Structural Equation Modeling*, 10:214–221.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA.

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Socioligical Methods and Research*, 27:182–225.

www.manaraa.com

Sturmfels, B. (2002). Solving systems of polynomial equations. In *CBMS Lectures Series*. American Mathematical Society.

Tian, J., Kang, C., and Pearl, J. (2006). A characterization of interventional distributions in semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 1239–1244. AAAI Press.

Tian, J. and Pearl, J. (2002a). A new characterization of the experimental implications of causal bayesian networks. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. http://www.cs.iastate.edu/ ~jtian/r298_aaai02.pdf.

Tian, J. and Pearl, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In et al., P. B., editor, *Uncertainty in Artificial Intelligence 6*, pages 220–227. Elsevier Science, Cambridge, MA.

Wermuth, N. and Cox, D. (2001). Graphical models: overview. *International Encyclopedia of the Social and Behavioral Sciences*, 9:6379–6386.

Wermuth, N. and Cox, D. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society B*, 66:687–717.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.*, 5:161–215.

www.manaraa.com

ACKNOWLEDGMENTS

This thesis would not have been possible without the support of many people at Iowa State University. First, I would like to thank my advisor, Jin Tian, for his guidance and support. He taught me to be an independent thinker and learner. I also thank the committee members and Manabu Kuroki for their encouragement and help. I must also thank my fellow graduate students in the Computer Science Department for the help and feedback they provided over the years.

www.manaraa.com