

Studies in Causal Reasoning and Learning

Jin Tian
August 2002

Technical Report R-309
Cognitive Systems Laboratory
Department of Computer Science
University of California
Los Angeles, CA 90095-1596, USA

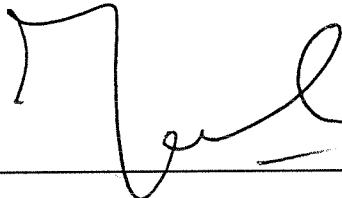
This report reproduces a dissertation submitted to UCLA in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science. This work was supported in part by AFOSR grant #F49620-01-1-0055, NSF grant #IIS-0097082, California State MICRO grants #99-096 and #00-077, and MURI grant #N00014-00-1-0617.

© Copyright by
Jin Tian
2002

The dissertation of Jin Tian is approved.

Adnan Darwiche

Adnan Darwiche



Jan de Leeuw

D. Jan de Leeuw

Jan de Leeuw

Judea Pearl

Judea Pearl, Committee Chair

University of California, Los Angeles

2002

To my parents

TABLE OF CONTENTS

1 Causal Models	1
1.1 Introduction	1
1.2 Causal Models and Interventions	1
1.3 Causal Models with Hidden Variables	3
1.4 Contributions	5
1.5 Overview	6
2 A Characterization of Causal Models	7
2.1 Introduction	7
2.2 Interventional Distributions in Markovian Models	7
2.3 Interventional Distributions in Semi-Markovian Models	14
2.4 Applications in the Identification of Causal Effects	16
2.5 Conclusion	17
3 Causal Discovery from Changes	19
3.1 Introduction	19
3.2 Mechanism Changes	21
3.3 Indistinguishability of Causal Graphs	22
3.4 Learning Causation by Detecting Changes	25
3.4.1 Partitioning the variables	26
3.4.2 Extracting causal information	27
3.4.3 Limitation of detecting marginal changes	30
3.4.4 Unknown focal variables	31
3.4.5 TSs absent of influentiality	32
3.4.6 Combining static and dynamic information	32
3.4.7 Experimental results	33
3.5 Causal Discovery by the Bayesian Approach	37
3.5.1 The Bayesian approach	37
3.5.2 Derivation of Bayesian score	37
3.5.3 Likelihood equivalence	40

3.5.4	Incorporating experimental data	41
3.5.5	Combining various types of dynamic data	42
3.5.6	Experimental results	46
3.6	Conclusion	47
4	Testable Implications of Causal Models	48
4.1	Introduction	48
4.2	Functional Constraints	49
4.3	C-components	52
4.4	Finding Constraints	56
4.4.1	Examples	56
4.4.2	Identifying constraints systematically	58
4.5	Projection to Semi-Markovian Models	60
4.6	Conclusion	62
5	Identification of Causal Effects	63
5.1	Introduction	63
5.2	Identification of $P_x(v)$	65
5.2.1	The easiest case	65
5.2.2	A more interesting case	65
5.2.3	The general case	67
5.3	Identification of $P_x(s)$	78
5.3.1	A criterion for identifying $P_x(s)$	78
5.3.2	An example	81
5.3.3	Lemmas	83
5.3.4	Computing $P_x(s)$	84
5.3.5	Useful graphical criteria	87
5.3.6	An example	89
5.3.7	Galles&Pearl's graphical criterion vs. <i>do</i> -calculus	92
5.4	Identification of $P_t(s)$	93
5.4.1	Computing $P_t(s)$	93
5.4.2	Useful graphical criteria	94
5.4.3	Examples	96

5.4.4	Identification of direct effects $P_{pa_y}(y)$	102
5.5	Identification of $P_t(s c)$	103
5.6	Beyond Semi-Markovian Models	105
5.7	Conclusion	105
6	Identification of Causal Effects in Linear Models	108
6.1	Linear Models	108
6.2	Causal Effects	109
6.3	Identifying Causal Effects	110
6.4	Identifying Causal Effects Systematically	113
6.5	Conclusion	117
7	Probabilities of Causation: Bounds and Identification	118
7.1	Introduction	118
7.2	Structural Model Semantics	120
7.3	Probabilities of Causation: Definitions	125
7.4	Bounds and Conditions of Identification	128
7.4.1	Linear programming formulation	129
7.4.2	Bounds with no assumptions	130
7.4.3	Bounds under exogeneity (no confounding)	131
7.4.4	Identifiability under monotonicity	134
7.4.5	Identifiability under monotonicity and exogeneity	137
7.4.6	Summary of results	139
7.5	Example 1: Legal Responsibility	140
7.6	Example 2: Personal Decision Making	143
7.7	Conclusion	144
References		148

LIST OF FIGURES

1.1 A typical causal graph.	2
1.2 A semi-Markovian model.	5
3.1 (a)The <i>Cancer</i> network. (a)-(d) are independence equivalent. (e)-(g) are B -transition equivalent. A mechanism change on A determines a unique causal graph (h).	24
3.2 (a) A causal graph; (b) The order graph for the TS (P, P_X, P_Y) ; (c) The marked order graph.	29
3.3 (a) A causal graph; (b) The order graph for the TS (P, P_X, P_Y) without knowing the focal variables; (c) The marked order graph.	31
3.4 Type I and Type II errors of χ^2 statistics.	35
4.1 The network (a) imposes functional constraints; the network (b) encodes the same set of independence statements as (a) but does not impose functional constraints.	49
4.2 The graph is partitioned into c-components $\{V_1, V_3\}$ and $\{V_2, V_4\}$	53
4.3 Subgraphs for finding constraints.	57
4.4 A model imposing functional constraints.	59
5.1 Theorem 14 is applicable for identifying $P_x(z_1, z_2, z_3, y)$	67
5.2 The problem of identifying $P_x(z_1, z_2, y)$	67
5.3 An example for applying Lemma 7.	70
5.4 A graph used in proving Theorem 16.	73
5.5 $P_x(y, z)$ is not identifiable but $P_x(y)$ is.	78
5.6 A graph used in proving Proposition 1.	79
5.7 Graphs used in proving Proposition 1.	80
5.8 Subgraphs of G used in computing $P_x(y)$	81
5.9 A function determining if $Q[C]$ is computable from $Q[T]$	86
5.10 An algorithm for computing $P_x(s)$	87
5.11 Subgraphs of G used in computing $P_x(y)$	90
5.12 An algorithm for computing $P_t(s)$	94
5.13 By Lemma 16, $P_{x_1x_2}(y)$ is identifiable if $P_{x_1}(y)$ is identifiable in G_S	97
5.14 By Lemma 16, $P_{x_1x_2}(y)$ is identifiable if $P_{x_2}(y)$ is identifiable.	98

5.15	By Lemma 14, $P_{x_1x_2}(v)$ is identifiable if both $P_{x_1}(v)$ and $P_{x_2}(v)$ are identifiable.	99
5.16	The problem of identifying $P_{x_1x_2}(y)$ (from [KM99]).	99
5.17	The problem of identifying $P_{x_1x_2}(y)$ (from [KM99]).	100
5.18	Subgraphs used in identifying $P_{x_1x_2}(w, y)$ in G	101
5.19	A graph in which the direct effect on Y is unidentified.	103
5.20	An algorithm for computing $P_t(s c)$	106
6.1	A linear model.	109
6.2	A function finding the minimum set $S_{min} \supseteq S$ such that $Q[S_{min}]$ is identifiable from $Q[T]$	114
6.3	Subgraphs for identifying path coefficients in G	115

LIST OF TABLES

3.1 Errors in order graphs. k : the number of focal variables. m : the number of buckets. E_o : percentage error of order claims. E_p : percentage error of NDP claims. E_e : percentage error of no-edge claims. u : number of unknown claims.	36
3.2 The posteriors of edges in the <i>Cancer</i> network.	45
7.1 PN (the probability of necessary causation) as a function of assumptions and available data. ERR stands of the excess-risk-ratio $1 - P(y x')/P(y x)$ and CERR is given in Eq. (7.49). The non-entries (—) represent vacuous bounds, that is, $0 \leq PN \leq 1$	139
7.2 Frequency data (hypothetical) obtained in experimental and non-experimental studies, comparing deaths (in thousands) among drug users, x , and non-users, x'	141

ACKNOWLEDGMENTS

Above all, I would like to thank my advisor, Judea Pearl. I am fortunate to have a chance to work with a researcher of his caliber. I thank the other members of my committee: Adnan Darwiche, Stott Parker, Jan de Leeuw, and Rina Dechter. Finally I would like to thank my family for their constant support.

VITA

1970	Born, ShuLe, XinJiang Province, P. R. China.
1992	B.S., Physics, Tsinghua University, Beijing.
1997	M.S., Physics, UCLA.
1998–2002	Research Assistant, Computer Science Department, UCLA.

PUBLICATIONS

- Tian, J. (2000). A branch-and-bound algorithm for MDL learning Bayesian networks. Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).
- , and Pearl, J. (2000). Probabilities of causation: bounds and identification. Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).
- , and Pearl, J. (2000). Probabilities of causation: bounds and identification. Annals of Mathematics and Artificial Intelligence, Vol. 28, 287-313.
- , and Pearl, J. (2001). Causal Discovery from Changes. Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).
- , and Pearl, J. (2002). A new characterization of the experimental implications of causal Bayesian networks. Proceedings of the National Conference on Artificial Intelligence (AAAI).
- , and Pearl, J. (2002). A general identification condition for causal effects. Proceedings of the National Conference on Artificial Intelligence (AAAI).
- , and Pearl, J. (2002). On the testable implications of causal models with hidden variables . Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI).

ABSTRACT OF THE DISSERTATION

Studies in Causal Reasoning and Learning

by

Jin Tian

Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2002
Professor Judea Pearl, Chair

Building intelligent systems that can learn about and reason with causes and effects is a fundamental task in artificial intelligence. This dissertation addresses various issues in causal reasoning and learning in the framework of causal Bayesian networks. We offer a complete characterization of the set of distributions that could be induced by local interventions on variables governed by a causal Bayesian network. The characterization provides a symbolic inferential tool for tasks in causal reasoning. We propose a new method of discovering causal structures, based on the detection of local, spontaneous changes in the underlying data-generating model. We show that the use of information about local changes increases our power of causal discovery beyond the limits set by independence equivalence that governs Bayesian networks. In the presence of unmeasured variables, causal models may impose non-independence functional constraints and no general criterion is previously available for finding those constraints. We offer a systematic method of identifying functional constraints, which facilitates the task of testing causal models. Causal effects permit us to predict how systems would respond to actions or policy decisions. We establish new graphical criteria for ensuring the identification of causal effects that generalize and simplify existing criteria in the literature, and we provide computational procedures for systematically identifying causal effects. Assessing the probability of causation,

that is, the likelihood that one event was the cause of another, guides much of what we understand about and how we act in the world. We show how useful information on the probabilities of causation can be extracted from empirical data, and how data from both experimental and nonexperimental studies can be combined to yield information that neither study alone can provide. Our results clarify the basic assumptions that must be made before statistical measures such as the excess-risk-ratio could be used for assessing attributional quantities such as the probability of causation.

CHAPTER 1

Causal Models

1.1 Introduction

A major challenge in artificial intelligence is to build autonomous intelligent systems that can make sense of their environment, so that they can respond to unexpected events or changes in the environment. Traditional probabilistic and statistical approaches assume a static time-invariant environment, and they can not predict what happens if the environment changes or some external actions occur. Such predictions are not discernible from probabilistic information; they rest on causal relationships. We human beings communicate about the world in the language of causation, and we would like to build intelligent systems that understand causal talk. We must build intelligent systems that can learn about and reason with causes and effects. The two challenges that we will face are:

1. How should an intelligent agent *acquire* causal information from the environment?
2. How should an intelligent agent *process* available causal information?

This dissertation addresses both of the problems in the framework of *causal Bayesian networks*, also called *causal models*¹, which provide a mathematical language for representing and reasoning about causal relations.

1.2 Causal Models and Interventions

The use of causal models for encoding distributional and causal assumptions is now fairly standard (see, for example, [Pea88, SGS93, HS95, Jor98, GPR99, Lau00, Pea00, Daw02]). The simplest such model, called *Markovian*, consists of a directed acyclic graph (DAG) G , called a *causal graph*, over a set $V = \{V_1, \dots, V_n\}$ of vertices, representing variables of interest, and a set of directed edges, or arrows, that connect these vertices (see Figure 1.1 for an example causal graph).

¹Throughout this dissertation, we will refer to the terms causal model and causal Bayesian network interchangeably.

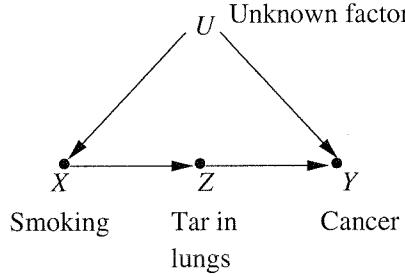


Figure 1.1: A typical causal graph.

The interpretation of a causal graph has two components, probabilistic and causal. The probabilistic interpretation views the arrows as representing probabilistic dependencies among the corresponding variables, and the missing arrows as representing conditional independence assertions: Each variable is independent of all its non-descendants given its direct parents in the graph.² These assumptions amount to asserting that the joint probability function $P(v) = P(v_1, \dots, v_n)$ factorizes according to the product

$$P(v) = \prod_i P(v_i | pa_i) \quad (1.1)$$

where Pa_i denotes the set of parents of variable V_i in the graph.³

The causal interpretation views the arrows as representing causal influences between the corresponding variables. In this interpretation, the factorization of (1.1) 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 chosen 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. 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 (1.1). The simplest such intervention, called *atomic*, involves fixing

²We use family relationships such as “parents,” “children,” “ancestors,” and “descendants,” to describe the obvious graphical relationships. For example, we say that V_i is a parent of V_j if there is an arrow from node V_i to V_j , $V_i \rightarrow V_j$.

³We use uppercase letters to represent variables or sets of variables, and use corresponding lowercase letters to represent their values (instantiations). For example, pa_i represents an instantiation of Pa_i .

a set T of variables to some constants $T = t$, which yields the post-intervention distribution⁴

$$P_t(v) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i) & \text{for all } v \text{ consistent with } T = t. \\ 0 & \text{for all } v \text{ inconsistent with } T = t. \end{cases} \quad (1.2)$$

Eq. (1.2) represents a truncated factorization of (1.1), with factors corresponding to the manipulated variables removed. This truncation follows immediately from (1.1) 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.⁵ If T stands for a set of treatment variables and Y for an outcome variable in $V \setminus T$, then Eq. (1.2) 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. This quantity, often called the *causal effect* of T on Y , is what we normally assess in a controlled experiment with T randomized, in which the distribution of Y is estimated for each level t of T .

We see from Eq. (1.2) that the model needed for predicting the effect of interventions requires the specification of three elements

$$M = \langle V, G, P(v_i|pa_i) \rangle$$

where (i) $V = \{V_1, \dots, V_n\}$ is a set of variables, (ii) G is a directed acyclic graph with nodes corresponding to the elements of V , and (iii) $P(v_i|pa_i)$, $i = 1, \dots, n$, is the conditional probability of variable V_i given its parents in G . Since $P(v_i|pa_i)$ is estimable from nonexperimental data whenever V is observed, we see that, given the causal graph G , all causal effects are estimable from the data as well.

1.3 Causal Models with Hidden Variables

Our ability to estimate $P_t(v)$ from nonexperimental data is severely curtailed when some variables in a Markovian causal model are unobserved. We call unobserved variables *hidden* or *latent* variables. If two or more variables in V are affected by unobserved confounders, the presence of such confounders would not permit the decomposition in (1.1). Letting $V = \{V_1, \dots, V_n\}$ and $U = \{U_1, \dots, U_{n'}\}$ stand for the sets of observed and hidden variables, respec-

⁴[Pea95a, Pea00] used the notation $P(v|set(t))$, $P(v|do(t))$, or $P(v|\hat{t})$ for the post-intervention distribution, while [Lau00] used $P(v||t)$.

⁵Eq. (1.2) was named “Manipulation Theorem” in [SGS93], and is also implicit in Robins’ (1987) G -computation formula.

tively, the observed probability distribution, $P(v)$, becomes a mixture of products:

$$P(v) = \sum_u \prod_{\{i|V_i \in V\}} P(v_i|pa_{v_i}) \prod_{\{i|U_i \in U\}} P(u_i|pa_{u_i}) \quad (1.3)$$

where Pa_{v_i} and Pa_{u_i} stand for the sets of parents of V_i and U_i respectively, and the summation ranges over all the U variables. The post-intervention distribution,⁶ likewise, will be given as a mixture of truncated products

$$P_t(v) = \begin{cases} \sum_u \prod_{\{i|V_i \notin T\}} P(v_i|pa_{v_i}) \prod_i P(u_i|pa_{u_i}) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \quad (1.4)$$

And, the question of identifiability arises, i.e., whether it is possible to express $P_t(v)$ as a function of the observed distribution $P(v)$. Clearly, given a causal model M and any two sets T and S in V , $P_t(s)$ can be determined unambiguously using (1.4). The question of identifiability is whether a given causal effect $P_t(s)$ can be determined uniquely from the distribution $P(v)$ of the observed variables, and is thus independent of the unknown quantities, $P(v_i|pa_{v_i})$ and $P(u_i|pa_{u_i})$, that involve elements of U .

Definition 1 [Causal-Effect Identifiability]

The causal effect of a set of variables T on a disjoint set of variables S is said to be identifiable from a graph G if the quantity $P_t(s)$ can be computed uniquely from any positive probability of the observed variables—that is, if $P_t^{M_1}(s) = P_t^{M_2}(s)$ for every pair of models M_1 and M_2 with $P^{M_1}(v) = P^{M_2}(v) > 0$ and $G(M_1) = G(M_2) = G$.

In other words, given the causal graph G , the quantity $P_t(s)$ can be determined from the observed distribution $P(v)$ alone; the details of M are irrelevant.

If, in a Markovian model with hidden variables, each hidden variable is a root node with exactly two observed children, then the corresponding model is called a *semi-Markovian* model. In a semi-Markovian model, the observed probability distribution $P(v)$ in Eq.(1.3) can be written as

$$P(v) = \sum_u \prod_i P(v_i|pa_i, u^i) \prod_i P(u_i) \quad (1.5)$$

where Pa_i and U^i stand for the sets of the observed and unobserved parents of V_i respectively. The post-intervention distribution is then given by

$$P_t(v) = \begin{cases} \sum_u \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i) \prod_i P(u_i) & v \text{ consistent with } t. \\ 0 & v \text{ inconsistent with } t. \end{cases} \quad (1.6)$$

⁶We only consider interventions on observed variables.

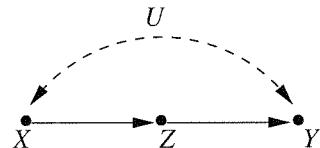


Figure 1.2: A semi-Markovian model.

It is convenient to represent a semi-Markovian model with a causal graph G that does not show the elements of U explicitly but, instead, represents the confounding effects of U variables using bidirected edges. Divergent edges $V_i \leftarrow U_k \rightarrow V_j$ will be represented by a bidirected edge between V_i and V_j . The presence of such bidirected edge in G represents unmeasured factors (or confounders) that may influence two variables in V ; we assume that substantive knowledge permits us to decide if such confounders can be ruled out from the model. For example, Figure 1.1 will be represented by Figure 1.2, assuming that the variable U is a hidden variable.

Causal Bayesian networks provide a strict mathematical language for reasoning with causes and effects. This dissertation addresses various issues in causal reasoning, including learning causal structures from data, testing causal models, assessing the effects of actions, and determining the causes of effects.

1.4 Contributions

The principal contributions of this dissertation are

- The establishment of a necessary and sufficient set of properties for interventional distributions induced by causal Bayesian networks.
- A new method of discovering causal structures, based on the detection of local, spontaneous changes in the underlying data-generating model.
- A procedure for systematically identifying functional constraints induced by causal Bayesian networks with hidden variables.
- A procedure for systematically identifying causal effects, in the presence of unmeasured confounders, from a combination of nonexperimental data and substantive assumptions encoded in the form of a directed acyclic graph.
- The derivation of tight bounds on probabilities of causation, from data obtained in experimental and observational studies, under general assumptions concerning the data-generating process.

1.5 Overview

In Chapter 2, we offer a complete characterization of interventional distributions that could be induced by a causal Bayesian network. We show that the set of interventional distributions must adhere to three norms of coherence, and we demonstrate the use of these norms as inferential tools in tasks of learning and identification. In Chapter 3, we propose a new method of discovering causal structures, based on the detection of local, spontaneous changes in the underlying data-generating model. We analyze the classes of structures that are equivalent relative to a stream of distributions produced by local changes, and devised algorithms that output graphical representations of these equivalence classes. We investigate both the Bayesian approach and an approach that infers structures by detecting distributional changes. Chapter 4 develops a systematic procedure of identifying functional constraints induced by causal Bayesian networks with hidden variables. The procedure facilitates the task of testing causal models as well as inferring such models from data. Chapter 5 concerns the assessment of the causal effects in nonparametric models. The chapter establishes new criteria for deciding whether the assumptions encoded in a causal graph are sufficient for assessing the strength of causal effects and, if the answer is positive, computational procedures are provided for expressing causal effects in terms of the underlying joint distribution. Chapter 6 shows how to use the results in Chapter 5 to identify causal effects in linear models. Chapter 7 deals with the problem of estimating the probability of causation, that is, the probability that one event was the cause of another in a given scenario, for example, the probability that event E would not have occurred if it were not for event C , given that C and E did in fact occur. Starting from structural-semantical definitions of the probabilities of necessary or sufficient causation (or both), we show how to bound these quantities from data obtained in experimental and observational studies, under general assumptions concerning the data-generating process. The results delineate more precisely the basic assumptions that must be made before statistical measures such as the excess-risk-ratio could be used for assessing attributional quantities such as the probability of causation.

CHAPTER 2

A Characterization of Causal Models

2.1 Introduction

In this chapter, we seek a characterization for the set of interventional distributions, $P_t(v)$, that could be induced by some causal Bayesian network. Whereas [Pea00, pp.23-4] has given such characterization relative to a given network, we assume that the underlying network, if such exists, is unknown. Given a collection of arbitrary interventional distributions, we ask whether the collection is compatible with the predictions of some underlying causal Bayesian network. Section 2.2 identifies three properties (of the collection) that are both necessary and sufficient for the existence of such an underlying network. Section 2.3 identifies necessary properties of distributions induced by semi-Markovian models, causal Bayesian networks in which some of the variables are unmeasured. Section 2.4 shows how the properties uncovered in Sections 2.2 and 2.3 can be used as symbolic inferential tools for predicting the effects of actions from nonexperimental data in the presence of unmeasured variables. The Conclusion section outlines the use of these properties in learning tasks which aim at uncovering the structure of the network.

2.2 Interventional Distributions in Markovian Models

Let \mathcal{P}_* be a set of arbitrary interventional distributions

$$\mathcal{P}_* = \{P_t(v) | T \subseteq V, t \in Dm(T)\} \quad (2.1)$$

where $Dm(T)$ represents the domain of T . For example, assume that V consists of two binary variables X and Y with the domain of X being $\{x_0, x_1\}$ and the domain of Y being $\{y_0, y_1\}$, then \mathcal{P}_* contains distributions $P(x, y)$, $P_{x_0}(x, y)$, $P_{x_1}(x, y)$, $P_{y_0}(x, y)$, $P_{y_1}(x, y)$, $P_{x_0, y_0}(x, y), \dots$, where each $P_t(x, y)$ is an arbitrary probability distribution over X, Y . For this set of distributions to be induced by some underlying causal Bayesian network such that each $P_t(x, y)$ corresponds to the distribution of X, Y under the intervention $do(T = t)$ to the causal Bayesian network, they have to satisfy some norms of coherence. For example, it must

be true that $P_{x_0}(x_0) = 1$. For another example, if the causal graph is $X \longrightarrow Y$ then $P_{y_0}(x_0) = P(x_0)$, and if the causal graph is $X \longleftarrow Y$ then $P_{x_0}(y_0) = P(y_0)$, therefore, it must be true that either $P_{y_0}(x_0) = P(x_0)$ or $P_{x_0}(y_0) = P(y_0)$. We would like to know what properties a \mathbf{P}_* set must satisfy such that it is compatible with some underlying causal Bayesian network. In this section, we show that a \mathbf{P}_* set induced from a Markovian causal model is fully characterized by three properties: effectiveness, Markov, and recursiveness.

Property 1 (Effectiveness) *For any set of variables T ,*

$$P_t(t) = 1. \quad (2.2)$$

Effectiveness states that, if we force a set of variables T to have the value t , then the probability of T taking that value t is one.

For any set of variables S disjoint with T , an immediate corollary of effectiveness reads:

$$P_{t,s}(t) = 1, \quad (2.3)$$

which follows from

$$P_{t,s}(t) \geq P_{t,s}(t, s) = 1. \quad (2.4)$$

Equivalently, if $T_1 \subseteq T$, then

$$P_t(t_1) = \begin{cases} 1 & \text{if } t_1 \text{ is consistent with } t. \\ 0 & \text{if } t_1 \text{ is inconsistent with } t. \end{cases} \quad (2.5)$$

We further have that, for $T_1 \subseteq T$ and S disjoint of T ,

$$P_t(s, t_1) = \begin{cases} P_t(s) & \text{if } t_1 \text{ is consistent with } t. \\ 0 & \text{if } t_1 \text{ is inconsistent with } t. \end{cases} \quad (2.6)$$

Property 2 (Markov) *For any two disjoint sets of variables S_1 and S_2 ,*

$$P_{v \setminus (s_1 \cup s_2)}(s_1, s_2) = P_{v \setminus s_1}(s_1)P_{v \setminus s_2}(s_2). \quad (2.7)$$

An equivalent form of the Markov property is: For any set of variables $T \subseteq V$,

$$P_t(v \setminus t) = \prod_{\{i | V_i \in V \setminus T\}} P_{v \setminus \{v_i\}}(v_i). \quad (2.8)$$

Eq. (2.8) can be obtained by repeatedly applying Eq. (2.7), and Eq. (2.7) follows from Eq. (2.8) as follows:

$$\begin{aligned}
P_{v \setminus (s_1 \cup s_2)}(s_1, s_2) &= \prod_{V_i \in S_1 \cup S_2} P_{v \setminus \{v_i\}}(v_i) \\
&= \prod_{V_i \in S_1} P_{v \setminus \{v_i\}}(v_i) \prod_{V_i \in S_2} P_{v \setminus \{v_i\}}(v_i) \\
&= P_{v \setminus s_1}(s_1) P_{v \setminus s_2}(s_2).
\end{aligned} \tag{2.9}$$

Definition 2 For two single variables X and Y , define “ X affects Y ”, denoted by $X \rightsquigarrow Y$, as $\exists W \subset V, w, x, y$, such that $P_{x,w}(y) \neq P_w(y)$. That is, X affects Y if, under some setting w , intervening on X changes the distribution of Y .

Property 3 (Recursiveness) For any set of variables $\{X_0, \dots, X_k\} \subseteq V$,

$$(X_0 \rightsquigarrow X_1) \wedge \dots \wedge (X_{k-1} \rightsquigarrow X_k) \Rightarrow \neg(X_k \rightsquigarrow X_0). \tag{2.10}$$

Property 3 is a stochastic version of the (deterministic) recursiveness axiom given in [Hal98]. It comes from restricting the causal models under study to those having acyclic causal graphs. For $k = 1$, for example, we have $X \rightsquigarrow Y \Rightarrow \neg(Y \rightsquigarrow X)$, saying that for any two variables X and Y , either X does not affect Y or Y does not affect X . [Hal98] pointed out that, recursiveness can be viewed as a collection of axioms, one for each k , and that the case of $k = 1$ alone is not enough to characterize a recursive model.

Theorem 1 (Soundness) Effectiveness, Markov, and recursiveness hold in all Markovian models.

Proof: All three properties follow from the factorization of Eq. (1.2).

Effectiveness From Eq. (1.2), we have

$$P_t(T = t') = 0 \quad \text{for } t' \neq t, \tag{2.11}$$

and since

$$\sum_{t' \in Dm(T)} P_t(t') = 1, \tag{2.12}$$

we obtain the effectiveness property of Eq. (2.2).

Markov From Eq. (1.2), we have

$$P_t(v \setminus t) = P_t(t, v \setminus t) = \prod_{V_i \in V \setminus T} P(v_i | pa_i). \quad (2.13)$$

Letting $T = V \setminus \{V_i\}$ in Eq. (2.13) yields

$$P_{v \setminus \{v_i\}}(v_i) = P(v_i | pa_i). \quad (2.14)$$

Substituting Eq. (2.14) back into Eq. (2.13), we get the Markov property (2.8), which is equivalent to (2.7).

Recursiveness Assume that a total order over V that is consistent with the causal graph is $V_1 < \dots < V_n$, such that V_i is a nondescendant of V_j if $V_i < V_j$. Consider a variable V_j and a set of variables $S \subseteq V$ which does not contain V_j . Let $B_j = \{V_i | V_i < V_j, V_i \in V \setminus S\}$ be the set of variables not in S and ordered before V_j , and let $A_j = \{V_i | V_j < V_i, V_i \in V \setminus S\}$ be the set of variables not in S and ordered after V_j . First we show that

$$P_{v_j, s}(b_j) = P_s(b_j). \quad (2.15)$$

We have

$$\begin{aligned} P_{v_j, s}(b_j) &= \sum_{a_j} P_{v_j, s}(a_j, b_j) \\ &= \sum_{a_j} P_{v_j, s, a_j}(b_j) P_{v_j, s, b_j}(a_j), \quad (\text{by Eq. (2.7)}) \end{aligned} \quad (2.16)$$

where $P_{v_j, s, a_j}(b_j) = \prod_{\{i | V_i \in B_j\}} P(v_i | pa_i)$ is a function of b_j and its parents. Since all variables in A_j are ordered after the variables in B_j , $P_{v_j, s, a_j}(b_j)$ is not a function of a_j . Hence Eq. (2.16) becomes

$$\begin{aligned} P_{v_j, s}(b_j) &= P_{v_j, s, a_j}(b_j) \sum_{a_j} P_{v_j, s, b_j}(a_j) \\ &= P_{v_j, s, a_j}(b_j) \end{aligned} \quad (2.17)$$

Similarly,

$$\begin{aligned} P_s(b_j) &= \sum_{v_j, a_j} P_s(v_j, a_j, b_j) \\ &= \sum_{v_j, a_j} P_{v_j, s, a_j}(b_j) P_{s, b_j}(v_j, a_j) \\ &= P_{v_j, s, a_j}(b_j) \sum_{v_j, a_j} P_{s, b_j}(v_j, a_j) \\ &= P_{v_j, s, a_j}(b_j) \end{aligned} \quad (2.18)$$

Eq. (2.15) follows from (2.17) and (2.18).

From Eq. (2.15), we have that, for any two variables $V_i < V_j$ and any set of variables S ,

$$P_{v_j,s}(v_i) = P_s(v_i), \quad (2.19)$$

which states that if X is ordered before Y then Y does not affect X , based on our definition of “ X affects Y ”. Therefore, we have that if X affects Y then X is ordered before Y , or

$$X \rightsquigarrow Y \Rightarrow X < Y. \quad (2.20)$$

Recursive property (2.10) then follows from (2.20) because the relation “ $<$ ” is a total order.

□

To facilitate the proof of the completeness theorem, we give the following lemma.

Lemma 1 [Pea88, p.124] *Given a DAG over V , if a set of functions $f_i(v_i, pa_i)$ satisfy*

$$\sum_{v_i \in Dm(V_i)} f_i(v_i, pa_i) = 1, \text{ and } 0 \leq f_i(v_i, pa_i) \leq 1, \quad (2.21)$$

and $P(v)$ can be decomposed as

$$P(v) = \prod_i f_i(v_i, pa_i), \quad (2.22)$$

then we have

$$f_i(v_i, pa_i) = P(v_i | pa_i), \quad i = 1, \dots, n. \quad (2.23)$$

Theorem 2 (Completeness) *If a P_* set satisfies effectiveness, Markov, and recursiveness, then there exists a Markovian model with a unique causal graph that can generate this P_* set.*

Proof: Define a relation “ \prec ” as: $X \prec Y$ if $X \rightsquigarrow Y$. Then the transitive closure of \prec , \prec^* , is a partial order over the set of variables V from the recursiveness property as shown in [Hal98]. Let “ $<$ ” be a total order on V consistent with \prec^* . We have that

$$\text{if } X < Y \text{ then } P_{y,s}(x) = P_s(x) \quad (2.24)$$

for any set of variables S . This is because if $P_{y,s}(x) \neq P_s(x)$, then $Y \rightsquigarrow X$, and therefore $Y \prec X$, which contradicts the fact that $X < Y$ is consistent with \prec^* .

Define a set PA_i as a minimal set of variables that satisfies

$$P_{pa_i}(v_i) = P_{v \setminus \{v_i\}}(v_i). \quad (2.25)$$

We have that

$$\text{if } V_i < V_j, \text{ then } V_j \notin PA_i. \quad (2.26)$$

Otherwise, assuming $V_j \in PA_i$ and letting $PA'_i = PA_i \setminus \{V_j\}$, from Eqs. (2.24) and (2.25) we have

$$P_{pa'_i}(v_i) = P_{pa'_i, v_j}(v_i) = P_{v \setminus \{v_i\}}(v_i), \quad (2.27)$$

which contradicts the fact that PA_i is minimal. From Eq. (2.26), drawing an arrow from each member of PA_i toward V_i , the resulting graph G is a DAG.

Substituting Eq. (2.25) into the Markov property (2.8), we obtain, for any set of variables T ,

$$P_t(v \setminus t) = \prod_{\{i | V_i \notin T\}} P_{pa_i}(v_i). \quad (2.28)$$

By Lemma 1, we get

$$P_{pa_i}(v_i) = P(v_i | pa_i). \quad (2.29)$$

From Eqs. (2.28), (2.29), and the effectiveness property (2.6), Eq. (1.2) follows. Therefore, a Markovian model with a causal graph G can generate this \mathbf{P}_* set.

Next, we show that the set PA_i is unique. Assuming that there are two minimal sets PA_i and PA'_i both satisfying Eq. (2.25), we will show that their intersection also satisfies Eq. (2.25). Let $A = PA_i \cap PA'_i$, $B = PA_i \setminus A$, $B' = PA'_i \setminus A$, and $S = V \setminus (PA_i \cup PA'_i \cup \{V_i\})$. From the Markov property Eq. (2.7), we have

$$\begin{aligned} P_a(b, b', s, v_i) &= P_{a, v_i}(b, b', s) P_{v \setminus \{v_i\}}(v_i) \\ &= P_{a, v_i}(b, b', s) P_{a, b}(v_i) \end{aligned} \quad (2.30)$$

Summing both sides of (2.30) over B' and S , we get

$$P_a(b, v_i) = P_{a, v_i}(b) P_{a, b}(v_i). \quad (2.31)$$

Substituting $P_{pa_i}(v_i)$ with $P_{pa'_i}(v_i)$ in (2.31), we get

$$P_a(b, v_i) = P_{a, v_i}(b) P_{a, b'}(v_i). \quad (2.32)$$

Summing both sides of (2.32) over B , we obtain

$$P_a(v_i) = P_{a,b'}(v_i) = P_{pa'_i}(v_i), \quad (2.33)$$

which says that the set $A = PA_i \cap PA'_i$ also satisfies Eq. (2.25). This contradicts the assumption that both PA_i and PA'_i are minimal. Thus PA_i is unique. \square

A Markovian model also satisfies the following properties.

Property 4 *If a set B is composed of nondescendants of a variable V_j , then for any set of variables S ,*

$$P_{v_j,s}(b) = P_s(b). \quad (2.34)$$

Proof: If B is disjoint of S , Eq. (2.34) follows from Eq. (2.15) since $B \subseteq B_j$. If B is not disjoint of S , Eq. (2.34) follows from the Effectiveness property and Eq. (2.15). \square

Property 5 *For any set of variables $S \subseteq V \setminus (PA_i \cup \{V_i\})$,*

$$P_{pa_i,s}(v_i) = P_{pa_i}(v_i). \quad (2.35)$$

Proof: Let $S' = V \setminus (PA_i \cup \{V_i\} \cup S)$.

$$\begin{aligned} P_{pa_i,s}(v_i) &= \sum_{s'} P_{pa_i,s}(s', v_i) \\ &= \sum_{s'} P_{v \setminus \{v_i\}}(v_i) P_{pa_i,s,v_i}(s') \quad (\text{by Eq. (2.7)}) \\ &= P_{pa_i}(v_i) \sum_{s'} P_{pa_i,s,v_i}(s') \quad (\text{by Eq. (2.25)}) \\ &= P_{pa_i}(v_i) \end{aligned} \quad (2.36)$$

\square

Property 6

$$P_{pa_i}(v_i) = P(v_i|pa_i). \quad (2.37)$$

Property 6 has been given in Eq. (2.29).

Property 7 *For any set of variables $S \subseteq V$, and $V_i \notin S$,*

$$P_s(v_i|pa_i) = P(v_i|pa_i), \quad \text{for } pa_i \text{ consistent with } s. \quad (2.38)$$

Proof: Let $S' = V \setminus (PA_i \cup \{V_i\} \cup S)$. Assuming that pa_i is consistent with s , we have

$$\begin{aligned}
P_s(v_i, pa_i) &= \sum_{s'} P_s(v_i, pa_i, s') \\
&= \sum_{s'} P_{v \setminus \{v_i\}}(v_i) P_{s, v_i}(pa_i, s') \quad (\text{by Eq. (2.7)}) \\
&= P(v_i | pa_i) \sum_{s'} P_{s, v_i}(pa_i, s') \quad (\text{by Eq. (2.14)}) \\
&= P(v_i | pa_i) P_{s, v_i}(pa_i) \\
&= P(v_i | pa_i) P_s(pa_i) \quad (\text{by Property 4}) \tag{2.39}
\end{aligned}$$

which leads to Eq. (2.38). \square

2.3 Interventional Distributions in Semi-Markovian Models

When some variables in a Markovian model are unobserved, the probability distribution over the observed variables may no longer be decomposed as in Eq. (1.1). Let $V = \{V_1, \dots, V_n\}$ and $U = \{U_1, \dots, U_{n'}\}$ stand for the sets of observed and unobserved variables respectively. In a semi-Markovian model, as defined in Chapter 1.3, the observed probability distribution and the post-intervention distribution are given by Eqs. (1.5) and (1.6) respectively.

If, in a semi-Markovian model, no U variable is an ancestor of more than one V variable, then $P_t(v)$ in Eq. (1.6) factorizes into a product as in Eq. (1.2), regardless of the parameters $\{P(v_i | pa_i, u^i)\}$ and $\{P(u)\}$. Therefore, for such a model, the causal Markov condition holds relative to G_V (the subgraph of G composed only of V variables), that is, each variable V_i is independent on all its non-descendants given its parents PA_i in G_V . And by convention, the U variables are usually not shown explicitly, and G_V is called the causal graph of the model.

The causal Markov condition is often assumed as an inherent feature of causal models (see e.g. [KSC84, SGS93]). It reflects our two basic causal assumptions: (i) include in the model every variable that is a cause of two or more other variables in the model; and (ii) Reichenbach's (1956) common-cause assumption, also known as "no correlation without causation," stating that, if any two variables are dependent, then one is a cause of the other *or* there is a third variable causing both.

If two or more variables in V are affected by unobserved confounders, the presence of such confounders would not permit the decomposition in Eq. (1.1),

and, in general, $P(v)$ generated by a semi-Markovian model is a mixture of products given in (1.5). However, the conditional distribution $P(v|u)$ factorizes into a product

$$P(v|u) = \prod_i P(v_i|pa_i, u^i), \quad (2.40)$$

and we also have

$$P_t(v|u) = \begin{cases} \prod_{\{i|V_i \notin T\}} P(v_i|pa_i, u^i) & \text{for all } v \text{ consistent with } T = t. \\ 0 & \text{for all } v \text{ inconsistent with } T = t. \end{cases} \quad (2.41)$$

Therefore all Properties 1–7 hold when we condition on u . For example, the Markov property can be written as

$$P_{v \setminus (s_1 \cup s_2)}(s_1, s_2|u) = P_{v \setminus s_1}(s_1|u)P_{v \setminus s_2}(s_2|u). \quad (2.42)$$

Let $\mathbf{P}_*(u)$ denote the set of all conditional interventional distributions

$$\mathbf{P}_*(u) = \{P_t(v|u)|T \subseteq V, t \in Dm(T)\} \quad (2.43)$$

Then $\mathbf{P}_*(u)$ is fully characterized by the three properties effectiveness, Markov, and recursiveness, conditioning on u .

Let \mathbf{P}_* denote the set of all interventional distributions over observed variables V as in (2.1). From the properties of the $\mathbf{P}_*(u)$ set, we can immediately conclude that the \mathbf{P}_* set satisfies the following properties: effectiveness (Property 1), recursiveness (Property 3), Property 4, and Property 5, while Markov (Property 2), Property 6, and Property 7 do not hold. For example, Property 5 can be proved from its conditional version,

$$P_{pa_i,s}(v_i|u) = P_{pa_i}(v_i|u), \quad (2.44)$$

as follows

$$P_{pa_i,s}(v_i) = \sum_u P_{pa_i,s}(v_i|u)P(u) = \sum_u P_{pa_i}(v_i|u)P(u) = P_{pa_i}(v_i). \quad (2.45)$$

Significantly, the \mathbf{P}_* set must satisfy inequalities that are unique to semi-Markovian models, as opposed, for example, to models containing feedback loops. For example, from Eq. (1.6), and using

$$P(v_i|pa_i, u^i) \leq 1, \quad (2.46)$$

we obtain the following property.

Property 8 *For any three sets of variables, T , S , and R , we have*

$$P_{tr}(s) \geq P_t(r, s) + P_r(t, s) - P(t, r, s) \quad (2.47)$$

Additional inequalities, involving four or more subsets, can likewise be derived by this method. However, finding a set of properties that can completely characterize the \mathbf{P}_* set of a semi-Markovian causal model remains an open challenge.

2.4 Applications in the Identification of Causal Effects

Given two disjoint sets T and S , the causal effect $P_t(s)$ is said to be *identifiable* if, given a causal graph, it can be determined uniquely from the distribution $P(v)$ of the observed variables, and is thus independent of the unknown quantities, $P(u)$ and $P(v_i|pa_i, u^i)$, that involve elements of U . Identification means that we can learn the effect of the action $T = t$ (on the variables in S) from sampled data taken prior to actually performing that action. In Markovian models, all causal effects are identifiable and are given in Eq. (1.2). When some confounders are unobserved, the question of identifiability arises. Sufficient graphical conditions for ensuring the identification of $P_t(s)$ in semi-Markovian models were established by several authors [SGS93, Pea93, Pea95a] and are summarized in [Pea00, Chapters 3 and 4]. Since

$$P_t(s) = \sum_u P_t(s|u)P(u), \quad (2.48)$$

and since we have a complete characterization over the set of conditional interventional distributions ($\mathbf{P}_*(u)$), we can use Properties 1–3 (conditioning on u) for identifying causal effects in semi-Markovian models.

The assumptions embodied in the causal graph can be translated into the language of conditional interventional distributions as follows:

For each variable V_i ,

$$P_{v \setminus \{v_i\}}(v_i|u) = P_{pa_i}(v_i|u^i). \quad (2.49)$$

The Markov property (2.8) conditioning on u then becomes

$$P_t(v \setminus t|u) = \prod_{\{i|V_i \in V \setminus T\}} P_{pa_i}(v_i|u^i). \quad (2.50)$$

The significance of Eq. (2.50) rests in simplifying the derivation of elaborate causal effects in semi-Markov models. To illustrate this derivation, consider the model in Figure 1.2, and assume we need to derive the causal effect of X on $\{Z, Y\}$, a task analyzed in [Pea00, pp.86-8] using do-calculus. Applying (2.50) to $P_x(y, z|u)$, (with x replacing t), we obtain:

$$\begin{aligned} P_x(y, z) &= \sum_u P_x(y, z|u)P(u) \\ &= \sum_u P_z(y|u)P_x(z|u)P(u) \\ &= P_x(z)P_z(y) \end{aligned} \quad (2.51)$$

Each of these two factors can be derived by simple means; $P_x(z) = P(z|x)$ because Z has no unobserved parent, and $P_z(y) = \sum_{x'} P(y|x', z)P(x')$ because X blocks all back-door paths from Z to Y (they can also be derived by applying (2.50) to $P(x, y, z|u)$). As a result, we immediately obtain the desired quantity:

$$P_x(y, z) = P(z|x) \sum_{x'} P(y|x', z)P(x'), \quad (2.52)$$

a result that required many steps in do-calculus.

In general, from (2.50), we have

$$P_t(v \setminus t) = \sum_u \prod_{\{i|V_i \in V \setminus T\}} P_{pa_i}(v_i|u^i)P(u). \quad (2.53)$$

Depending on the causal graph, the right hand side of (2.53) may sometimes be decomposed into a product of summations as

$$\begin{aligned} P_t(v \setminus t) &= \prod_j \sum_{n_j} \prod_{V_i \in S_j} P_{pa_i}(v_i|u^i)P(n_j) \\ &= \prod_j P_{v \setminus s_j}(s_j), \end{aligned} \quad (2.54)$$

where N_j 's form a partition of U and S_j 's form a partition of $V \setminus T$. Eq. (2.51) is an example of such a decomposition. Therefore the problem of identifying $P_t(v \setminus t)$ is reduced to identifying some $P_{v \setminus s_j}(s_j)$'s. Based on this decomposition, a method for systematically identifying causal effects is developed in Chapter 5.

2.5 Conclusion

We have shown that all experimental results obtained from an underlying Markovian causal model are fully characterized by three norms of coherence: Effectiveness, Markov, and Recursiveness. We have further demonstrated the use of these norms as inferential tools for identifying causal effects in semi-Markovian models. This permits one to predict the effects of actions and policies, in the presence of unmeasured variables, from data obtained prior to performing those actions and policies.

The key element in our characterization of experimental distributions is the generic formulation of the Markov property (2.7) as a relationship among three experimental distributions, instead of the usual formulation as a relationship between a distribution and a graph (as in (1.1)). The practical implication of this formulation is that violations of the Markov property can be detected without knowledge of the underlying causal graph; comparing distributions from just

three experiments, $P_{v \setminus (s_1 \cup s_2)}(s_1, s_2)$, $P_{v \setminus s_1}(s_1)$, and $P_{v \setminus s_2}(s_2)$, may reveal such violations, and should allow us to conclude, prior to knowing the structure of G , that the underlying data-generation process is non-Markovian. Alternatively, if our confidence in the Markovian nature of the data-generation process is unassailable, such a violation would imply that the three experiments were not conducted on the same population, under the same conditions, or that the experimental interventions involved had side effects and were not properly confined to the specified sets S_1 , S_2 , and $S_1 \cup S_2$.

This feature is useful in efforts designed to infer the structure of G from a combination of observational and experimental data; a single violation of (2.7) suffices to reveal that unmeasured confounders exist between variables in S_1 and those in S_2 . Likewise, a violation of any inequality in (2.47) would imply that the underlying model is not semi-Markovian; this means that feedback loops may operate in data generating process, or that the interventions in the experiments are not “atomic”.

CHAPTER 3

Causal Discovery from Changes

3.1 Introduction

Inferring causal structures from empirical data has become an active research area in recent years. Several graph-based algorithms have been developed for this purpose. Some are based on detecting patterns of conditional independence relationships [PV91, SGS93], and some are based on Bayesian approaches [CH92, Gei95, Coo99]. These discovery methods assume static environment, that is, a time-invariant distribution and a time-invariant data-generating model, and attempt to infer structures that encode dynamic aspects of the environment, for example, how probabilities would change as a result of interventions. This transition, from static to dynamic information, constitutes a major inferential leap, and is severely limited by the inherent indistinguishability (or equivalence) relation that governs Bayesian networks [VP90].

One way of overcoming this basic limitation is to augment the data with partial causal knowledge, if such is available. [SGS93], for example, discussed the use of experimental data to identify causal relationships. [CY99] discussed a Bayesian method of causal discovery from a mixture of observational and experimental data.

We propose a new method of discovering causal relations in data, based on the detection and interpretation of local spontaneous changes in the environment. While previous methods assume that data are generated by a static statistical distribution, our proposal aims at exploiting dynamic changes in that distribution. Such changes are always present in any realistic domain that is embedded in a larger background of dynamically changing conditions. For example, natural disasters, armed conflicts, epidemics, labor disputes, and even mundane decisions by other agents, are unexpected eventualities that are not naturally captured in distribution functions. The occurrence of such eventualities tend to *alter* the distribution under study and yield changes that are markedly different from ordinary statistical fluctuations. Whereas static analysis views these changes as nuisance, and attempts to adjust and compensate for them, we will view them as a valuable source of information about the data-generating process. A controlled experimental study may be thought of as a special case of these environmental

changes, where the external influence involves fixing a designated variable to some predetermined value. In general, however, the external influence may be milder, merely changing the conditional probability of a variable, given its causes. Moreover, in marked contrast to controlled experiments, we may not know in advance the nature of the change, its location, or even whether it took place; these may need to be inferred from the data itself.

The basic idea has its roots in the economic literature. The economist Kevin Hoover (1990) attempted to infer the direction of causal influences among economic variables (e.g., employment and money supply) by observing the changes that sudden modifications in the economy (e.g., tax reform, labor dispute) induced in the statistics of these variables. Hoover assumed that the conditional probability of an effect given its causes remains invariant to changes in the mechanism that generates the cause, while the conditional probability of a cause given the effect would not remain invariant under such changes. This asymmetry may be useful in distinguishing cause and effect.

Today we understand more precisely the conditions under which such asymmetries would prevail and how to interpret such asymmetries in the context of large, multi-variate systems. Whenever we obtain reliable information (e.g., from historical or institutional knowledge) that an abrupt local change has taken place in a specific mechanism that constrains a given family of variables, we can use the observed changes in the marginal and conditional probabilities surrounding those variables to determine the direction of causal influences in the domain. The statistical features that remain invariant under such changes, as well as the causal assumptions underlying this invariance, are encoded in the causal graph at hand, and can be used therefore for testing the validity of a given structure. Likewise, conflicts between observed and predicted changes can be used for automatic restructuring of the topology of the structure at hand.

In this chapter, we will assume that we have data generated from a dynamically changing environment and our task is to recover the actual causal structures. In Section 3.2, we formally present this learning problem. In Section 3.3, we analyze the equivalence classes of causal structures relative to the given data. In Section 3.4, we analyze the patterns of distributional changes induced by data and present recovery methods that infer causal directionality information from those changes. In Section 3.5, we investigate the Bayesian approach for causal discovery. The Bayesian approach [HMC97] gives us a consistent way of combining dynamic datasets to get an overall estimation of causal structures. We show how to derive a Bayesian scoring metric from various types of dynamic data by assigning appropriate priors over probability parameters. The Bayesian scores obtained are extensions of previously derived Bayesian scores [CH92, Gei95]. For mixed observational and experimental data we obtained the same score as given

in [CY99]. We show that dynamic data increase our power of causal discovery beyond the limits set by independence equivalence.

3.2 Mechanism Changes

Let our problem domain be a set of discrete random variables $V = \{V_1, \dots, V_n\}$. In this chapter, we denote a causal model over V by a pair $M = \langle G, \Theta_G \rangle$, where G is the causal graph and Θ_G is a set of probability parameters. We assume that each variable V_i can take values from a finite domain, $Dm(V_i) = \{v_{i1}, \dots, v_{ir_i}\}$, where r_i is the number of states of V_i . Let $\theta_{v_i;pa_i}$, $v_i \in Dm(V_i)$, $pa_i \in Dm(Pa_i)$ denote the multinomial parameter corresponding to the conditional probability $P(v_i|pa_i)$. We will use the following notations: $\vec{\theta}_{pa_i} = \{\theta_{v_i;pa_i} | v_i \in Dm(V_i)\}$, $\Psi_i = \cup_{pa_i \in Dm(Pa_i)} \vec{\theta}_{pa_i}$, $\Theta_G = \cup_{i=1}^n \Psi_i$. A causal model $M = \langle G, \Theta_G \rangle$ generates a probability distribution given in Eq. (1.1), rewritten as

$$P(v) = \prod_i \theta_{v_i;pa_i}. \quad (3.1)$$

A probability distribution $P(V)$ is said to be *compatible* with a causal graph G if $P(V)$ can be generated by some causal model $M = \langle G, \Theta_G \rangle$.

Based on the modularity assumption that each family in the causal graph represents an autonomous physical mechanism and is subjected to change without influencing other mechanisms, we formally define mechanism change as follows.

Definition 3 (Mechanism Change) *A mechanism change to a causal model $M = \langle G, \Theta_G \rangle$ at a variable V_i is a transformation of M that produces a new model, $M_{V_i} = \langle G, \Theta'_G \rangle$, where $\Theta'_G = \Psi'_i \cup (\Theta_G \setminus \Psi_i)$ and Ψ'_i is a set of parameters having values that differ from those in Ψ_i .*

We will assume that the parent set Pa_i does not change in a mechanism change. We see that an intervention that fixes V_i to a particular value is a special case of a mechanism change. Let $P(V)$ be the distribution generated by M , as in Eq. (3.1). Then the distribution generated by M_{V_i} is given by

$$P_{V_i}(v) = \theta'_{v_i;pa_i} \prod_{j \neq i} \theta_{v_j;pa_j}. \quad (3.2)$$

We will call (P, P_{V_i}) a *transition pair* (TP) and V_i the *focal variable* of the transition. Assume that a series of mechanism changes occurred successively to a causal model $M = \langle G, \Theta_G^0 \rangle$, and let $F = (V_{i_1}, \dots, V_{i_k})$ denote the corresponding sequence of focal variables. We use $P_{TS} = (P^0, P^1, \dots, P^k)$ to denote the

sequence of distributions generated by such a series, and call the pair (P_{TS}, F) a *transition sequence (TS)*.

As oracles for cause-and-effect relations, causal models can predict the effects that any external or spontaneous changes have on the distributions. Conversely, by detecting how probability distributions change under various mechanism changes, we obtain information on the structure of the model generating those distributions. We propose to exploit the stream of distributions from mechanism changes to recover underlying causal structures. In this chapter, we make the following assumptions: each mechanism change occurs at one single variable at a time, and we have the distribution (or samples thereof) after each single mechanism change, that is, we know when each mechanism change happens and at which variable. We will then assume that we are given a TS (P_{TS}, F) corresponding to some causal graph G , or, we have a sequence of datasets $\mathbb{D}_{TS} = \{D^0, \dots, D^k\}$, where each D^i is a set of random samples from a distribution P^i , such that each pair (P^{j-1}, P^j) is a TP with focal variable V_{ij} , and our task will be to recover a causal graph (or a set of graphs) that can generate \mathbb{D}_{TS} . First, we study what can be learned from a TS.

3.3 Indistinguishability of Causal Graphs

Our ability to recover causal graphs is limited by the statistical indistinguishability of causal models with given data. In this section, we study the classes of causal structures that are indistinguishable (or “equivalent”) relative to a TS.

The statistical information provided by any causal graph is completely encoded in the independence relationships among the variables. Therefore, two causal graphs are statistically indistinguishable given one static distribution if and only if they are independence equivalent. The graphical conditions for independence equivalence are given by the following theorem.

Theorem 3 (Independence Equivalence) *Two causal graphs are independence equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow [VP90].*

Now assume that we have a TP with focal variable V_i . A causal graph G is said to be *compatible with a transition pair* (P, P_{V_i}) if P can be generated by a causal model $M = \langle G, \Theta_G \rangle$ and P_{V_i} can be generated by a causal model $M_{V_i} = \langle G, \Theta'_G \rangle$ resulted from a mechanism change to M at V_i . Note that a causal graph could be compatible with both P and P_{V_i} but *not* compatible with the TP (P, P_{V_i}) . Among those independence-equivalent graphs compatible with

both P and P_{V_i} , a TP (P, P_{V_i}) can distinguish those that can generate P_{V_i} from P with a *single* mechanism change from those that can not. Two causal graphs G_1 and G_2 are called *transition pair equivalent* with respect to a TP with focal variable V_i , or V_i -*transition equivalent*, if every TP (P, P_{V_i}) compatible with G_1 is also compatible with G_2 . Two causal graphs are statistically indistinguishable given a TP (P, P_{V_i}) if and only if they are V_i -transition equivalent.

Theorem 4 (Transition Pair Equivalence) *Two causal graphs G_1 and G_2 are V_i -transition equivalent if and only if they have the same skeletons, the same sets of v-structures, and the same sets of parents for V_i .*

Proof: Let G_1 be compatible with a TP (P, P_{V_i}) . G_2 must have the same skeletons and the same sets of v-structures as G_1 to be compatible with P (and P_{V_i}) by Theorem 3. We have the following decomposition:

$$P(v) = P(v_i|pa_i^1) \prod_{j \neq i} P(v_j|pa_j^1) = P(v_i|pa_i^2) \prod_{j \neq i} P(v_j|pa_j^2), \quad (3.3)$$

where Pa_i^1 and Pa_i^2 are parents of V_i in G_1 and G_2 respectively. G_1 is compatible with the TP (P, P_{V_i}) , hence can generate P_{V_i} from P by a mechanism change at V_i :

$$P_{V_i}(v) = P_{V_i}(v_i|pa_i^1) \prod_{j \neq i} P(v_j|pa_j^1). \quad (3.4)$$

Plugging the expression for $\prod_{j \neq i} P(v_j|pa_j^1)$ from Eq. (3.3) into Eq. (3.4), we have

$$P_{V_i}(v) = P_{V_i}(v_i|pa_i^1) \frac{P(v_i|pa_i^2)}{P(v_i|pa_i^1)} \prod_{j \neq i} P(v_j|pa_j^2). \quad (3.5)$$

G_2 is also compatible with the transition pair (P, P_{V_i}) if and only if

$$P_{V_i}(v) = P_{V_i}(v_i|pa_i^2) \prod_{j \neq i} P(v_j|pa_j^2). \quad (3.6)$$

Eqs. (3.5) and (3.6) lead to

$$P_{V_i}(v_i|pa_i^1) \frac{P(v_i|pa_i^2)}{P(v_i|pa_i^1)} = P_{V_i}(v_i|pa_i^2), \quad (3.7)$$

which holds for any distribution P and P_{V_i} if and only if G_1 has the same parent set for V_i as G_2 ($Pa_i^1 = Pa_i^2$); if G_1 has a different parent set for V_i with G_2 , Eq. (3.7) will impose some constraints between P and P_{V_i} , and will not hold for arbitrary possible transition pair (P, P_{V_i}) . \square

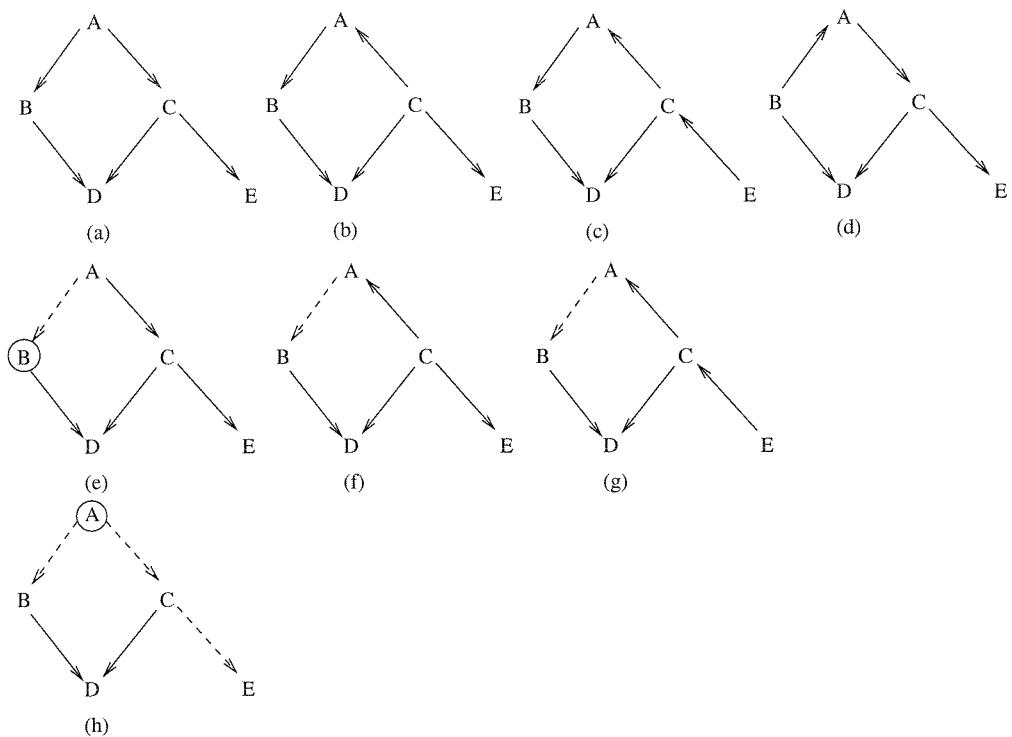


Figure 3.1: (a)The *Cancer* network. (a)-(d) are independence equivalent. (e)-(g) are *B*-transition equivalent. A mechanism change on *A* determines a unique causal graph (h).

A TS is simply a series of TP's. Accordingly, we say that a causal graph is *compatible with a transition sequence* $P_{TS} = (P^0, P^1, \dots, P^k)$, $F = (V_{i_1}, \dots, V_{i_k})$ if it is compatible with each TP (P^{j-1}, P^j) in the sequence. Likewise, two causal graphs G_1 and G_2 are called *transition sequence equivalent* with respect to a TS (P_{TS}, F) , or *F-transition equivalent*, if every TS (P_{TS}, F) compatible with G_1 is also compatible with G_2 . Two causal graphs are statistically indistinguishable given a TS (P_{TS}, F) if and only if they are *F*-transition equivalent.

Theorem 5 (Transition Sequence Equivalence) *Two causal graphs are *F*-transition equivalent if and only if they have the same skeletons, the same sets of v-structures, and the same sets of parents for variables in *F*.*

Theorem 5 says that a TS determines the directions of the edges between the focal variables and their neighbors (among the set of independence-equivalent graphs). See Figure 3.1 for an example of TS equivalence.

Given a TS, the most we can expect to recover is a set of causal graphs that are TS-equivalent, as defined by Theorem 5. We may find this equivalence class by detecting independence relations and distribution changes.

3.4 Learning Causation by Detecting Changes

In this section, we identify the causal information that can be learned by detecting various changes in the probability distributions, in particular, changes in the marginal probability of each variable. The following theorem is obvious.

Theorem 6 *A mechanism change at a variable X to a causal model $M = \langle G, \Theta_G \rangle$ may alter the marginal probabilities of the descendants of X in G and can not alter the marginals of nondescendants of X .*

It is possible of course that, for some peculiar parameter changes, the marginal probabilities of some descendants of X would not change. When recovering causal information from distributional changes, we assume a restriction on a TS called *influentiality*.

Definition 4 (influentiality) *A TP (P, P_X) generated by a causal model $\langle G, \Theta_G \rangle$ is said to be influential if for every descendant Y of X in G , the marginal distribution $P_X(Y)$ is different from $P(Y)$. A TS is influential if every TP in the sequence is influential.*

Assuming influentiality, we can obtain causal information by detecting changes of marginal probabilities.

Given a TP (P, P_X) , and assuming that we can test each variable for marginal distribution change, we can draw the following inferences. If the marginal of a variable Y has changed, we conclude that Y is a descendant of X . If the marginal of a variable Z has not changed, we conclude that Z is a nondescendant of X . We thus conclude that $Z < X < Y$ should be a causal order consistent with the causal graph. Next we discuss how to piece together ordering information of this kind, as obtained from a TS.

3.4.1 Partitioning the variables

Given a TS P_{TS} , $F = (V_{i_1}, \dots, V_{i_k})$, each variable can be characterized by a sequence of 1's and 0's, a tag a_1, \dots, a_k , where a_i reflects whether the marginal of that variable changed ($a_i = 1$) or not ($a_i = 0$) in the i th transition of the sequence. Non-focal variables that are given the same tags cannot be distinguished by the TS (through detecting marginal changes), and no information can therefore be extracted about their relative causal order in the causal graph. We may put all such variables into a bucket labeled with the same tag, denoted by $B_{a_1 \dots a_k}$. Clearly, since we have no information on causal relations among variables within the same bucket, all variables in a bucket stand in the same ordering relation to all variables in another bucket. Focal variables need special treatment since they carry more information, and we will put each focal variable into an individual bucket called a *focal bucket*, denoted by $B_{a_1 \dots a_k}^f$.

We classify variables into buckets with the following algorithm.

Algorithm 1 (Partitioning Variable)

Input: a TS P_{TS} , $F = (V_{i_1}, \dots, V_{i_k})$.

Output: A set of buckets, each associated with a tag $a_1 \dots a_k$, and each containing a set of variables.

Put all variables in a bucket B .

For the i th mechanism change, $i = 1, \dots, k$,

For each bucket $B_{a_1 \dots a_{i-1}}$ including focal buckets

if it contains the i th focal variable, put it in a focal bucket $B_{a_1 \dots a_{i-1} 1}^f$.

put other changing variables in $B_{a_1 \dots a_{i-1} 1}$.

put non-changing variables in $B_{a_1 \dots a_{i-1} 0}$.

We show the partitioning process by an example. Assume that the actual causal graph is the DAG shown in Figure 3.2(a) and that we are given a TS (P, P_X, P_Y) . In the first transition, with X as the focal variable, $P(Y)$ does not change, hence $B_0 = \{Y\}$; $P(X), P(Z), P(W), P(Q)$ do change, hence we form $B_1 = \{Z, W, Q\}$, $B_1^f = \{X\}$. Note that a focal variable is put into an

individual bucket. In the second transition, with Y as the focal variable, $P(Y)$ changes, giving $B_{01}^f = \{Y\}$; $P(Z)$ and $P(W)$ change, giving $B_{11} = \{Z, W\}$; $P(Q)$ and $P(X)$ do not change, giving $B_{10} = \{Q\}$ and $B_{10}^f = \{X\}$. As a result, the variables are partitioned into four buckets: $B_{10}^f = \{X\}$, $B_{01}^f = \{Y\}$, $B_{10} = \{Q\}$, $B_{11} = \{Z, W\}$.

3.4.2 Extracting causal information

We shall now discuss what causal information we can extract from the tags attached to buckets. Consider any two buckets $B_{a_1 \dots a_k}$ and $B_{b_1 \dots b_k}$. If there exists a bit such as $a_i < b_i$ (i.e., $a_i = 0$ and $b_i = 1$), it must be that, in the i th transition, the marginals of variables in $B_{a_1 \dots a_k}$ did not change and the marginals of variables in $B_{b_1 \dots b_k}$ did. Therefore, no variable in $B_{a_1 \dots a_k}$ is a descendant of any variable in $B_{b_1 \dots b_k}$. On the other hand, if there exists another bit such that $a_j > b_j$ ($a_j = 1, b_j = 0$), then no variable in $B_{b_1 \dots b_k}$ is a descendant of any variable in $B_{a_1 \dots a_k}$, which means that there exists no directed path, in particular no edge, between any variable in $B_{a_1 \dots a_k}$ and any variable in $B_{b_1 \dots b_k}$. The equality $a_i = b_i, i = 1, \dots, k$ can only happen if one of the buckets is a focal bucket, in which case the focal variable is an ancestor of all the variables in the other bucket. In summary, the relation between two buckets $B_{a_1 \dots a_k}$ and $B_{b_1 \dots b_k}$ is determined as follows:

R1 $a_i \leq b_i, i = 1, \dots, k$ and $\exists j, a_j < b_j$: variables in $B_{a_1 \dots a_k}$ are nondescendants of variables in $B_{b_1 \dots b_k}$, denoted by $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$.

R2 $a_i \geq b_i, i = 1, \dots, k$ and $\exists j, a_j > b_j$: $B_{b_1 \dots b_k} < B_{a_1 \dots a_k}$.

R3 There exist two bits $i \neq j$ such that $a_i < b_i$ and $a_j > b_j$: there can be no directed path between any variable in $B_{a_1 \dots a_k}$ and any variable in $B_{b_1 \dots b_k}$.

R4 $a_i = b_i, i = 1, \dots, k$, one of the buckets, say $B_{a_1 \dots a_k}^f$, is a focal bucket: all variables in $B_{b_1 \dots b_k}$ must be descendants of the focal variable in $B_{a_1 \dots a_k}^f$, which is a stronger relation than that in R1 and R2 but will still be denoted by $B_{a_1 \dots a_k}^f < B_{b_1 \dots b_k}$.

The focal buckets convey more information. Let $B_{a_1 \dots a_k}$ be a focal bucket containing the focal variable V_{ij} for the j th transition. Then if $b_j = 1$, we have that all variables in $B_{b_1 \dots b_k}$ are descendants of V_{ij} since their marginals changed in the j th transition. This rule is consistent with the above rules R1–R3, hence it is applied only in R4 when R1–R3 cannot determine a relation. However, in practice, due to imperfect statistical tests, there may be conflicts between them. For example, we may determine that there is no edge between $B_{a_1 \dots a_k}$ and $B_{b_1 \dots b_k}$ by R3 and in

the same time $B_{a_1 \dots a_k}$ is a focal bucket for the j th transition and $b_j = 1$. These conflicts signal mistakes in the statistical tests, and whenever there are conflicts, we will declare the relation as “unknown”. We summarize the above discussions with the following algorithm.

Algorithm 2 (Extracting Relation)

Input: two buckets $B_{a_1 \dots a_k}$ and $B_{b_1 \dots b_k}$.

Output: the relation between the two buckets, could be “ $<$ ”, “no-directed-path (NDP)”, or “unknown”.

1. $a_i \leq b_i, i = 1, \dots, k$ and $\exists j, a_j < b_j$: if $B_{b_1 \dots b_k}$ is a focal bucket for the l th transition and $a_l = 1$ then “unknown”, else $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$.
2. $a_i \geq b_i, i = 1, \dots, k$ and $\exists j, a_j > b_j$: if $B_{a_1 \dots a_k}$ is a focal bucket for the l th transition and $b_l = 1$ then “unknown”, else $B_{b_1 \dots b_k} < B_{a_1 \dots a_k}$.
3. There exist two bits $i \neq j$ such that $a_i < b_i$ and $a_j > b_j$: if $B_{b_1 \dots b_k}$ is a focal bucket for the l th transition and $a_l = 1$ or $B_{a_1 \dots a_k}$ is a focal bucket for the l th transition and $b_l = 1$ then “unknown”, else “NDP”.
4. $a_i = b_i, i = 1, \dots, k$: if both buckets are focal buckets then “unknown”, else let the focal bucket be $B_{a_1 \dots a_k}^f$, then $B_{a_1 \dots a_k}^f < B_{b_1 \dots b_k}$.

Consider the binary relation “ $<$ ” on the set of buckets as defined in the Algorithm 2. We have the following theorem.

Theorem 7 *The binary relation “ $<$ ” on the set of buckets is a partial order.*

Proof: The relation is transitive. If $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$ and $B_{b_1 \dots b_k} < B_{c_1 \dots c_k}$, we have $a_i \leq b_i \leq c_i, i = 1, \dots, k$.

1. $\exists j, a_j < c_j$. If $B_{c_1 \dots c_k}$ is not a focal bucket, then we have $B_{a_1 \dots a_k} < B_{c_1 \dots c_k}$. If $B_{c_1 \dots c_k}$ is a focal bucket for the l th transition and $a_l = 1$, then $b_l = 1$ since $a_l \leq b_l \leq c_l$, which contradicts $B_{b_1 \dots b_k} < B_{c_1 \dots c_k}$.
2. $a_i = c_i, i = 1, \dots, k$. Then $a_i = b_i = c_i, i = 1, \dots, k$, and then $B_{a_1 \dots a_k}$ has to be a focal bucket and $B_{b_1 \dots b_k}$ is not one in order to have the relation $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$, which then contradicts $B_{b_1 \dots b_k} < B_{c_1 \dots c_k}$.

The relation is antisymmetric. If $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$ and $B_{b_1 \dots b_k} < B_{a_1 \dots a_k}$, then $a_i = b_i, i = 1, \dots, k$. Since they cannot both be focal buckets, they must be the same bucket. \square

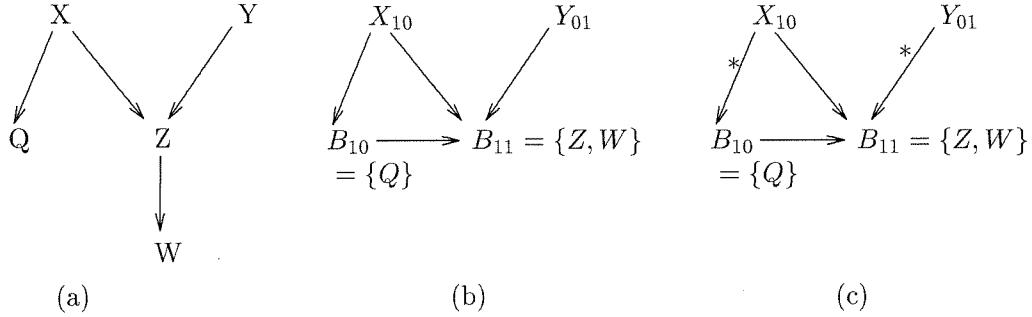


Figure 3.2: (a) A causal graph; (b) The order graph for the TS (P, P_X, P_Y); (c) The marked order graph.

A partially ordered set can be represented by a DAG. We construct a graph with both directed and undirected edges, called an *order graph* (*OG*), as follows: a node represents a bucket; for each pair of buckets B and B' , there is a directed edge $B \rightarrow B'$ if $B < B'$; there is an undirected edge $B—B'$ if the relation between them is “unknown”. If we had a perfect statistical test for distributional changes, an OG would be a DAG. For the causal graph shown in Figure 3.2(a) and the TS (P, P_X, P_Y), the ideal OG is given in Figure 3.2(b).

In an OG, when B is a focal bucket, a directed edge $B \rightarrow B'$ asserts that there exists a directed path from the focal variable contained in B to all the variables in B' . Hence, if there is no other *mixed directed path*, a path that could contain undirected edges but no directed edges in the reverse direction, from B to B' in the OG, there must be an edge from B to at least one variable in B' in the causal graph. We mark this type of edges as $B \xrightarrow{*} B'$, to distinguish them from those that only represent potential edges in the causal graph. This information is useful when the child bucket B' contains only one variable; we then assert that the edge $B \rightarrow B'$ must exist in the causal graph. We will call an OG with marked edges a *marked order graph* (MOG); an example is shown in Figure 3.2(c).

An algorithm for constructing a MOG is given in the following.

Algorithm 3 (Constructing MOG)

Input: an influential TS with known focal variables.

Output: a marked order graph.

1. Put variables into buckets using Algorithm 1.
2. Extracting relations among buckets using Algorithm 2.

3. Let each bucket be a node.
4. For each pair of nodes B and B'
 - If $B < B'$, add an edge $B \rightarrow B'$.
 - If $B' < B$, add an edge $B' \rightarrow B$.
 - If the relation is “unknown”, add an edge $B—B'$.
5. For each focal bucket B^f and each of its child B
 - If there is no other mixed directed path from B^f to B , mark the edge as $B^f \xrightarrow{*} B$.

In summary, the information conveyed by a MOG is as follows:

1. An unmarked edge $B \rightarrow B'$: All variables in B can be ordered before all variables in B' in the causal graph, in other words, there are no directed paths from variables in B' to variables in B . When B is a focal variable, there exists a directed path from B to each variable in B' in the causal graph.
2. A marked edge $B \xrightarrow{*} B'$: There exists a directed path from B to each variable in B' . In the case that both B and B' contain one single variable, the edge $B \rightarrow B'$ must exist in the causal graph.
3. No edge between B and B' : there is no directed path, in particular no edge, between any variable in B and any variable in B' in the causal graph.

3.4.3 Limitation of detecting marginal changes

Can we fully recover a causal graph by detecting marginal distribution changes alone? To fully recover a causal graph, we must construct a MOG in which each bucket contains only one variable and every edge is marked. This may not, in general, be achieved. Considering a causal graph G containing a path $X \rightarrow Z \rightarrow Y$, it is clear that we can never determine if there is an edge $X \rightarrow Y$ in G , since all marginal changes produced by transitions would be the same after adding that edge. What is the best we can get then by detecting marginal changes?

Given a DAG G , if we remove an edge $X \rightarrow Y$ whenever there is a directed path from X to Y , we get the *transitive reduction* of G . The transitive reduction of a DAG G is the graph G' with the fewest edges such that the *transitive closure* of G' is equal to the transitive closure of G . The transitive closure of a DAG G is the graph G'' such that an edge $X \rightarrow Y$ is in G'' iff there is a directed path from X to Y in G . By detecting marginal changes in TS's, the best we can hope to

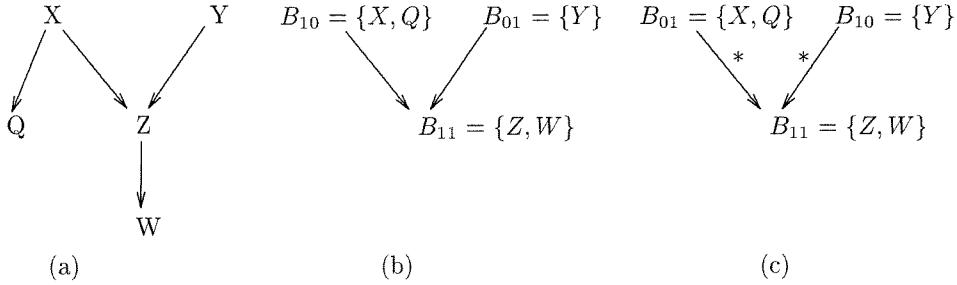


Figure 3.3: (a) A causal graph; (b) The order graph for the TS (P, P_X, P_Y) without knowing the focal variables; (c) The marked order graph.

get is the transitive reduction of the actual causal graph. Since to mark an edge $X \rightarrow Y$, X must be a focal variable, it follows that every node except leaf nodes must be a focal variable in order to mark every edge in the transitive reduction graph. To further make each bucket contain only one variable, every leaf node having the same set of parents as another leaf node must be a focal variable.

In conclusion, by detecting marginal distribution changes, the best we can learn is the transitive reduction of the causal graph, and we can achieve it by a TS in which every variable has had its mechanism changed.

3.4.4 Unknown focal variables

In this section we discuss situations where we know that a mechanism change has occurred at a single variable but we do not know the identity of that variable.

We first note that, without knowing the focal variables, variables can still be partitioned into buckets using Algorithm 1, and the relations between pairs of buckets will be determined by rules R1–R3 of Section 3.4.2. Second, an order graph can be constructed as follows: for each pair of buckets B and B' , there is a directed edge $B \rightarrow B'$ if $B < B'$. For the causal graph of Figure 3.3(a) and the TS (P, P_X, P_Y) , the variables are partitioned into three buckets: $B_{10} = \{X, Q\}$, $B_{01} = \{Y\}$, $B_{11} = \{Z, W\}$, and the OG is shown in Figure 3.3(b).

Finally, we may be able to find to which bucket a focal variable belongs using the following theorem, assuming influentiality and perfect statistical tests. (We still call such a bucket a “focal bucket”, because it behaves as a focal variable with the information at hand.)

Theorem 8 *Let S_j be the set of buckets for which $a_j = 1$ in their tags $a_1 \dots a_k$, then the focal bucket F^j for the j th transition is in S_j and for any other bucket $B \in S_j$, $F^j < B$.*

Proof: Let the focal variable X for the j th transition be tagged as $a_1 \dots a_k$, then $a_j = 1$, since $P(X)$ must change in this transition. All other variables in the set of buckets S_j must be descendants of X since all their marginals changed in the j th transition. Therefore, whenever $P(X)$ changes, their marginals must change too, that is, if $a_i = 1$ then $b_i = 1$ for any variable tagged as $b_1 \dots b_k$ in S_j , which leads to $a_i \leq b_i, i = 1, \dots, k$. Hence for any bucket $B_{b_1 \dots b_k} \in S_j$ not containing X , we have $B_{a_1 \dots a_k} < B_{b_1 \dots b_k}$. \square

In practice, Theorem 8 may fail to identify a focal bucket when (due to imperfect statistical tests) there exists no bucket F^j in S_j satisfying $F^j < B$ for any other bucket $B \in S_j$. In the case that an identified focal bucket contains only one variable, we actually identify a focal variable. For the OG in Figure 3.3(b), the focal buckets for the first and second transitions can be found as $B_{10} = \{X, Q\}$ and $B_{01} = \{Y\}$ respectively, and we actually identify Y as the focal variable of the second transition.

Finally we can get a MOG by marking edges as in Algorithm 3. For our working example, the ideal MOG is shown in Figure 3.3(c).

3.4.5 TSs absent of influentiality

If we allow for the possibility that a mechanism change at X may not alter the marginal probabilities of some of X 's descendants, then detecting no change in $P(Y)$ provides no information on the causal relation between X and Y . The information we may obtain is that detecting a change in $P(Y)$ means that Y is a descendant of the focal variable X . First we partition variables into tagged buckets using Algorithm 1. Then the relationship among buckets is determined as: let B^i be the focal bucket for the i th transition; $B^i < B_{a_1 \dots a_k}$ if $a_i = 1$, where “ $<$ ” represents that all variables in $B_{a_1 \dots a_k}$ are descendants of the focal variable B^i . Finally we compute the transitive closure of $<$ relation, denoted by $<^*$, to get more information. Simultaneous $B <^* B'$ and $B' <^* B$ would mean change detection errors and the relation between B and B' will be declared as unknown. The information conveyed by $B <^* B'$ is that all variables in B' are descendants of the focal variable B in the underlying causal graph.

It is clear that if the identities of the focal variables are not given, we can not get any order information from a TS by detecting marginal changes.

3.4.6 Combining static and dynamic information

So far, we discussed how to extract causal information given a TS by detecting distributional changes. In this section, we briefly describe how to combine this

information with that obtained from independence tests.

Given data from a static stable distribution, we can recover (partially directed) causal graphs using conditional independence tests. Several such algorithms have been developed, including IC algorithm [Pea00, section 2.5] (initially introduced in [PV91]) and PC algorithm [SGS93]. The output of these algorithms is a partially oriented graph representing an independence-equivalence class as defined by Theorem 3.

To recover a causal graph from a TS, we first extract causal information by detecting distribution changes as described in Section 3.4, then run the IC algorithm using the causal information as prior knowledge. Note that since a TS is composed of a series of different distributions, we need to test independence relationships across all distributions.

We may obtain three types of causal information as shown in Section 3.4: causal order among certain variables, no edges between certain variables, and certain directed edges. The last two types (no-edge and determined-edge) can be incorporated directly. Causal order information can be used to restrict the search of candidate conditional sets and thus reduce the complexity of the IC algorithm. Causal order information can also be used to orient more edges: any undirected edge $X—Y$ can be oriented as $X \rightarrow Y$ if X is ahead of Y in the causal order. These methods of incorporating background knowledge have been discussed in [SGS93, Section 5.4.5].

When the identities of all focal variables are known, after incorporating these causal information as background knowledge, the output of the IC Algorithm would be a partially oriented graph representing the TS equivalence class as defined by Theorem 5. This is due to a theorem in [Mee95] which says that the orientation rules in the IC algorithm are complete with respect to any consistent background knowledge. If the identity of a focal variable is not given or identified as in Section 3.4.4, the edge directions between this focal variable and its neighbors may not be fixed, hence the output graph is not maximally oriented, and we have not obtained all the information implied by a TS. Algorithms for identifying focal variables are currently under investigation.

3.4.7 Experimental results

We use χ^2 test to detect distribution changes. Let D^1 and D^2 be two datasets, consisting of N_1 and N_2 cases respectively. Let N_{1x} and N_{2x} be the number of cases in D^1 and D^2 respectively in which a variable X takes the value x . To test the hypothesis that X has the same distribution in the two datasets, we compute

the quantity

$$\chi^2 = N_1 N_2 \sum_x \frac{1}{N_{1x} + N_{2x}} \left(\frac{N_{1x}}{N_1} - \frac{N_{2x}}{N_2} \right)^2, \quad (3.8)$$

which is asymptotically a χ^2 distribution with $r_x - 1$ degree of freedom, where r_x is the number of states of X . Let the significance level be α . If $\chi^2 > \chi_\alpha^2$ then we decide “change”, else we decide “no-change”.

A mechanism change at a variable V_i is simulated as follows. Consider parameters in $\vec{\theta}_{pa_i}$. If $\theta_{v_{i1};pa_i} \leq 0.5$ then let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} + \delta$, else let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} - \delta$, where δ is a parameter for adjusting the change magnitude. The rest of the parameters in $\vec{\theta}_{pa_i}$ are changed in proportional to their original values as: $\theta'_{v_{ij};pa_i} = \alpha \theta_{v_{ij};pa_i}, j = 2, \dots, r_i$, where $\alpha = (1 - \theta'_{v_{i1};pa_i}) / (1 - \theta_{v_{i1};pa_i})$. When we simulate a mechanism change at V_i , we change parameters in $\vec{\theta}_{pa_i}$ as above for each $pa_i \in Dm(Pa_i)$.

In our experiments, we used data generated from a known network, the *Alarm* Bayesian network¹ [BSC89]. Samples used in the experiment were generated from the network using a demo version of Netica API developed by Norsys Software Corporation. We used equal sample sizes for all datasets in a TS, that is, a sample size N represents that N cases were generated for each dataset D^i in $\mathbb{D}_{TS} = \{D^0, \dots, D^k\}$.

3.4.7.1 Errors in detecting changes

There are two types of errors in detecting changes: (i) mistaking “no-change” for a “change”, known as type I error and denoted NC2C, and (ii) mistaking “change” as “no-change”, known as type II error and denoted C2NC. Let G be the causal graph used for generating samples. When a mechanism change occurs at a variable V_i , if our test statistics is perfect, all V_i ’s descendants in G should be identified as “change” and V_i ’s nondescendants as “no-change”. Let Dec_i be the number of descendants of V_i in G and $NDec_i$ be the number of nondescendants of V_i . Let $c2nc_i$ be the number of descendants of V_i identified as “no-change” by the χ^2 test, and let $nc2c_i$ be the number of nondescendants of V_i identified as “change”. $nc2c_i$ and $c2nc_i$ represent the number of type I and type II mistakes made by the χ^2 statistics. In any one run, we simulate a mechanism change at each node $V_i, i = 1, \dots, n$, relative to the *original* network, and compute the C2NC error rate as $\sum_i c2nc_i / \sum_i Dec_i$ and the NC2C error rate as $\sum_i nc2c_i / \sum_i NDec_i$. We computed an average error rate over 5 runs.

¹We used the version downloaded from the web site of Norsys Software Corporation, <http://www.norsys.com>.

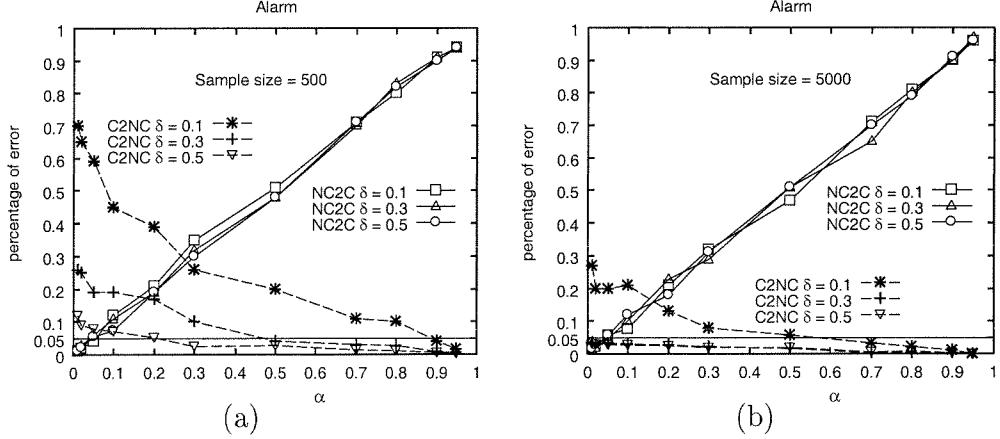


Figure 3.4: Type I and Type II errors of χ^2 statistics.

We varied the change magnitude δ , the sample size, and the significance level α , and the results are shown in Figure 3.4. We see that the NC2C (type I) error rate is nearly the same as the α value for different change magnitudes and sample sizes, as expected. The C2NC (type II) error could be large when the α value is small or the change magnitude is small. This suggests that we should consider using a two-tailed χ^2 test [SBM00] to control the C2NC error, especially when the sample size is not large. In a two-tailed χ^2 test, we use another threshold $\alpha' > \alpha$ such that we decide “no-change” only when $\chi^2 < \chi_{\alpha'}^2$, but we have to decide “unknown” when $\chi_{\alpha'}^2 < \chi^2 < \chi_{\alpha}^2$.

3.4.7.2 Errors in order graphs

In an OG, an edge $B \rightarrow B'$ represents that all variables in B can be causally ordered before the variables in B' . We call this type of information “order claims”. No edge between B and B' represents the absence of directed paths, in particular edges, between variables in B and those in B' ; this information will be called “no-directed-path (NDP) claims” and “no-edge claims” respectively. An edge $B-B'$ only signals mistakes in the statistical tests and will be called “unknown claims”. We performed the following experiments: for certain δ , α , sample size, and focal variables, we generate datasets, construct an OG, count the claims, and check against the true network to compute percentage errors for each type of claims.²

²Claims are counted between pairs of variables not between pairs of buckets. Numbers vary with the focal variables picked, hence we did 100 runs, each time randomly picking a sequence of k variables as focal variables, and computed average numbers.

Table 3.1: Errors in order graphs. k : the number of focal variables. m : the number of buckets. E_o : percentage error of order claims. E_p : percentage error of NDP claims. E_e : percentage error of no-edge claims. u : number of unknown claims.

$N = 500$								
k	δ	α	m	order claim		NDP claim		
				#	E_o	#	E_p	E_e
5	0.1	0.01	8	275	0.13	37	0.3	0.049
5	0.1	0.05	11	355	0.12	88	0.32	0.039
5	0.5	0.01	10	379	0.03	84	0.31	0.027
5	0.5	0.05	12	391	0.036	111	0.3	0.03
10	0.1	0.01	15	354	0.13	137	0.3	0.044
10	0.1	0.05	21	335	0.11	241	0.3	0.044
10	0.5	0.01	18	360	0.02	206	0.3	0.027
10	0.5	0.05	23	323	0.026	274	0.29	0.032
$N = 5000$								
k	δ	α	m	order claim		NDP claim		
				#	E_o	#	E_p	E_e
5	0.1	0.01	10	369	0.044	80	0.3	0.025
5	0.1	0.05	12	393	0.051	109	0.3	0.031
5	0.5	0.01	10	400	0.014	78	0.19	0.015
5	0.5	0.05	12	406	0.026	104	0.26	0.027
10	0.1	0.01	19	364	0.027	207	0.28	0.02
10	0.1	0.05	23	334	0.029	260	0.28	0.033
10	0.5	0.01	19	377	0.0081	191	0.25	0.02
10	0.5	0.05	23	334	0.018	265	0.26	0.03

The results are shown in Table 3.1 for various sample size N , number of focal variables k , mechanism change magnitude δ , and significance level α . From Table 3.1, we see that the NDP claims have a high percentage of error; however, if those claims are interpreted as representing no-edge only, then the error rates are much lower. As expected, the error rates are lower when δ , the change magnitude, is larger, and a TS with more focal variables produces more no-edge claims.

3.5 Causal Discovery by the Bayesian Approach

3.5.1 The Bayesian approach

Assume that we have a set of random samples D generated from a causal model $M = \langle G, \Theta_G \rangle$. In the Bayesian approach, we compute the posterior probability of a causal graph G given the dataset D as:

$$P(G|D, \xi) = \frac{P(D|G, \xi)P(G|\xi)}{P(D|\xi)}, \quad (3.9)$$

where ξ represents our background knowledge. The *marginal likelihood* of the data given G is computed as

$$P(D|G, \xi) = \int P(D|\Theta_G, G, \xi)P(\Theta_G|G, \xi)d\Theta_G. \quad (3.10)$$

The term $P(D|\Theta_G, G, \xi)$ is the probability of the data given a Bayesian network and is computable. We need to provide prior distributions for the probability parameters, $P(\Theta_G|G, \xi)$, and causal graphs, $P(G|\xi)$. The term $P(D|\xi)$ is just a proportional constant.

We can then compute the posterior probability of any hypothesis of interest by averaging over all possible causal models. For example, the posterior probability that X causes Y is computed as

$$P(X \rightarrow Y|D, \xi) = \sum_{X \rightarrow Y \in G} P(G|D, \xi), \quad (3.11)$$

where the summation is over all causal graphs which contain the edge $X \rightarrow Y$. Since the number of possible graphs is exponential in the number of variables n , it is impractical to sum over all graphs unless for very small n . One way to deal with this problem is to use the relative posterior probability $P(D, G|\xi)$ as a *scoring metric* and search for graphs with high scores.

3.5.2 Derivation of Bayesian score

For the case that the dataset D is from a static distribution, closed form expressions for $P(D|G, \xi)$ have been derived [CH92, Gei95]. We will extend previous derivations to incorporate dynamic data.

Assume that we have two data sets, D and D' , generated from a causal graph G but with different parameters, Θ_G and Θ'_G respectively. The marginal likelihood is computed as:

$$P(D, D'|G, \xi) = \int P(D, D'|\Theta_G, \Theta'_G, G, \xi)P(\Theta_G, \Theta'_G|G, \xi)d\Theta_Gd\Theta'_G. \quad (3.12)$$

Assuming that data cases are random samples, and that the data are *complete*, that is, every variable is assigned a value in all data cases, we have

$$\begin{aligned}
P(D, D' | \Theta_G, \Theta'_G, G, \xi) &= P(D | \Theta_G, G, \xi) P(D' | \Theta'_G, G, \xi) \\
&= \prod_{l=1}^N P(C_l | \Theta_G, G, \xi) \prod_{l=1}^{N'} P(C'_l | \Theta'_G, G, \xi) \\
&= \prod_{i=1}^n \prod_{v_i} \prod_{pa_i} \theta_{v_i; pa_i}^{N_{v_i, pa_i}} \theta'^{N'_{v_i, pa_i}},
\end{aligned} \tag{3.13}$$

where N is the number of cases in D , C_l represents a specific case in D , and N_{v_i, pa_i} is the number of cases in D for which V_i takes the value v_i and its parents Pa_i takes the value pa_i . We use \prod_{v_i} as a shorthand for $\prod_{v_i \in Dm(V_i)}$ and \prod_{pa_i} for $\prod_{pa_i \in Dm(Pa_i)}$.

Consider the prior distribution $P(\Theta_G, \Theta'_G | G, \xi)$. Assume that, as a background knowledge, the two datasets D and D' are from a TP (P, P') with known focal variable V_l . Therefore, the two sets of parameters Θ_G and Θ'_G differ only by those parameters in Ψ_l . With this knowledge, we assume the following prior:

$$P(\Theta_G, \Theta'_G | G, V_l, \xi) = P(\Theta_G | G, \xi) P(\Psi'_l | G, \xi) \prod_{i \neq l} \delta(\Psi_i - \Psi'_i), \tag{3.14}$$

where $\delta(x)$ is the Dirac delta function. Eq. (3.14) says that for $i \neq l$, $\Psi'_i = \Psi_i$, and the reader can verify that $P(\Theta_G, \Theta'_G | G, V_l, \xi)$ integrates to 1 and is a valid density function. We have put V_l as a condition to reflect the fact that V_l is known as the focal variable of the TP.

For the parameter priors $P(\Theta_G | G, \xi)$ and $P(\Psi'_l | G, \xi)$, we use the following assumptions given in [Gei95]:

- *Global Parameter Independence*:

$$P(\Theta_G | G, \xi) = \prod_{i=1}^n P(\Psi_i | G, \xi) \tag{3.15}$$

- *Local Parameter Independence*:

$$P(\Psi_i | G, \xi) = \prod_{pa_i} P(\vec{\theta}_{pa_i} | G, \xi), i = 1, \dots, n. \tag{3.16}$$

- *Parameter Modularity*: if V_i has the same parents in two causal graphs G_1 and G_2 , then

$$P(\vec{\theta}_{pa_i} | G_1, \xi) = P(\vec{\theta}_{pa_i} | G_2, \xi), pa_i \in Dm(Pa_i). \tag{3.17}$$

While these assumptions were originally made for learning Bayesian networks, [Hec95] discussed their implications for causal Bayesian networks.

Using Eq.s (3.13)–(3.17), and integrating out $\Theta'_G \setminus \Psi'_l$, Eq. (3.12) is transformed to

$$\begin{aligned} P(\mathbb{D}_{TP} | G, V_l, \xi) &= \prod_{i \neq l} \prod_{pa_i} \int \left(\prod_{v_i} \theta_{v_i;pa_i}^{M_{v_i,pa_i}} \right) P(\vec{\theta}_{pa_i} | \xi) d\vec{\theta}_{pa_i} \\ &\quad \times \prod_{pa_l} \int \left(\prod_{v_l} \theta_{v_l;pa_l}^{N_{v_l,pa_l}} \right) P(\vec{\theta}_{pa_l} | \xi) d\vec{\theta}_{pa_l} \\ &\quad \times \prod_{pa_l} \int \left(\prod_{v_l} \theta'_{v_l;pa_l}^{N'_{v_l,pa_l}} \right) P(\vec{\theta}'_{pa_l} | \xi) d\vec{\theta}'_{pa_l}, \end{aligned} \quad (3.18)$$

where

$$M_{v_i,pa_i} = N_{v_i,pa_i} + N'_{v_i,pa_i}. \quad (3.19)$$

We use the notation $\mathbb{D}_{TP} = \{D, D'\}$ and put V_l as a condition to emphasize that Eq. (3.18) is obtained under the assumption that the datasets D and D' are from a TP with known focal variable V_l . The standard assumption for $P(\vec{\theta}_{pa_i} | \xi)$ is a *Dirichlet distribution*:

$$P(\vec{\theta}_{pa_i} | \xi) = Dir(\vec{\theta}_{pa_i} | \vec{\alpha}_{pa_i}), \quad (3.20)$$

where $\vec{\alpha}_{pa_i} = \{\alpha_{v_i;pa_i} | v_i \in Dm(V_i)\}$ denotes the set of parameters for the Dirichlet distribution. Assuming that the set of parameters $\vec{\theta}'_{pa_l}$ have the same prior distribution as $\vec{\theta}_{pa_l}$ given by Eq. (3.20), we obtain

$$\begin{aligned} P(\mathbb{D}_{TP} | G, V_l, \xi) &= \prod_{i \neq l} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l} + N_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})} \\ &\quad \times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l} + N'_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N'_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})}, \end{aligned} \quad (3.21)$$

where $\Gamma(\cdot)$ is the Gamma function, and

$$\alpha_{pa_i} = \sum_{v_i} \alpha_{v_i;pa_i}, \quad N_{pa_i} = \sum_{v_i} N_{v_i,pa_i}, \quad M_{pa_i} = \sum_{v_i} M_{v_i,pa_i}.$$

3.5.3 Likelihood equivalence

For two independence-equivalent causal graphs G_1 and G_2 , any distribution compatible with G_1 is also compatible with G_2 . Hence, it is reasonable to assume that a dataset D from a static distribution cannot distinguish between independence-equivalent causal graphs, or, $P(D|G_1, \xi) = P(D|G_2, \xi)$. [Gei95] call this assumption *likelihood equivalence*. They show that it constrains the space of prior parameters $\alpha_{v_i;pa_i}$ and call the resulting likelihood-equivalent Bayesian scoring metric the BDe metric. We will use prior parameters that satisfy the likelihood equivalence property, and call the associated metric $P(\mathbb{D}_{TP}, G|V_l, \xi) = P(\mathbb{D}_{TP}|G, V_l, \xi)P(G|\xi)$ the BDe_TP metric.

The BDe_TP metric is *not* likelihood equivalent, and for a good reason. A TP can indeed distinguish independence-equivalent graphs: among those independence-equivalent graphs compatible with both P and P_{V_l} , a TP (P, P_{V_l}) can distinguish those that can generate P_{V_l} from P with a *single* mechanism change from those that can not.

It is natural to extend the likelihood equivalence requirement and define a new property: a marginal likelihood $P(\mathbb{D}|G, \xi)$ is said to be *V_l -transition likelihood equivalent* if for any dataset \mathbb{D} and two V_l -transition equivalent causal graphs G_1 and G_2 , $P(\mathbb{D}|G_1, \xi) = P(\mathbb{D}|G_2, \xi)$.

Theorem 9 *The marginal likelihood $P(\mathbb{D}_{TP}|G, V_l, \xi)$ given by Eq. (3.21) is V_l -transition likelihood equivalent.*

Proof: Eq. (3.21) can be rewritten as

$$\begin{aligned} P(\mathbb{D}_{TP}|G, V_l, \xi) &= \prod_i \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i;pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \left(\prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})\Gamma(\alpha_{pa_l} + M_{pa_l})}{\Gamma(\alpha_{pa_l} + N_{pa_l})\Gamma(\alpha_{pa_l} + N'_{pa_l})} \right. \\ &\quad \left. \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N_{v_l;pa_l})\Gamma(\alpha_{v_l;pa_l} + N'_{v_l;pa_l})}{\Gamma(\alpha_{v_l;pa_l})\Gamma(\alpha_{v_l;pa_l} + M_{v_l;pa_l})} \right). \end{aligned} \quad (3.22)$$

Let G_1 and G_2 be two V_l -transition-equivalent causal graphs. Then G_1 and G_2 are independence equivalent and have the same parent set Pa_l by Theorem 4. The first term in Eq. (3.22) has exactly the same form as the BDe score and takes the same values for two independence-equivalent graphs [Gei95]. The second term obtains the same values for G_1 and G_2 since they have the same Pa_l set. \square

We see that given data from a TP, previously indistinguishable independence-equivalent causal graphs may now be distinguished, and in this sense, two datasets generated from a same causal structure but with different parameters give us more power to learn the structure. This power comes from our assumption (or knowledge) that only a *single* causal mechanism has changed in generating the two datasets. Indeed, if we have no knowledge on how the two sets of parameters Θ_G and Θ'_G differ, we may only assume that they are independent and have the same distributions:

$$P(\Theta_G, \Theta'_G | G, \xi) = P(\Theta_G | G, \xi)P(\Theta'_G | G, \xi), \quad (3.23)$$

which leads to a marginal likelihood given by

$$\begin{aligned} P(D, D' | G, \xi) &= P(D | G, \xi)P(D' | G, \xi) \\ &= \prod_i \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + N_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + N_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_i \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + N'_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + N'_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})}. \end{aligned} \quad (3.24)$$

Eq. (3.24) is a product of two BDe likelihood applied on datasets D and D' respectively, and is still likelihood equivalent. Hence, without knowledge on how they came about, two datasets do not increase our power of discrimination, save for providing more samples.

3.5.4 Incorporating experimental data

Now assume that our knowledge is that the cases in D' are from an experimental study in which the variable V_l is fixed to a value $v_{lj} \in Dm(V_l)$, denoted by $do(V_l = v_{lj})$ or $do(v_{lj})$. Then instead of the Dirichlet distribution, we assign the following prior distribution to the parameter set $\vec{\theta}'_{pa_l}$:

$$P(\vec{\theta}'_{pa_l} | do(v_{lj}), \xi) = \delta(\theta'_{v_{lj};pa_l} - 1) \prod_{v_l \neq v_{lj}} \delta(\theta'_{v_l;pa_l}), \quad (3.25)$$

which asserts that

$$\theta'_{v_l;pa_l} = \begin{cases} 1 & \text{if } v_l = v_{lj} \\ 0 & \text{otherwise} \end{cases}$$

Plugging Eq. (3.25) into Eq. (3.18), we obtain

$$\begin{aligned} P(\mathbb{D}_{TP} | G, do(v_{lj}), \xi) &= \prod_{i \neq l} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l} + N_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})}. \end{aligned} \quad (3.26)$$

Eq. (3.26) has been given in [CY99]. Here we show that it can be derived by providing an informative parameter prior as given by Eqs. (3.14) and (3.25). In the derivation of Eq. (3.26), we have used the following equation

$$\int \left(\prod_{v_l} \theta'_{v_l;pa_l}^{N'_{v_l,pa_l}} \right) \delta(\theta'_{v_l;pa_l} - 1) \prod_{v_l \neq v_{lj}} \delta(\theta'_{v_l;pa_l}) d\vec{\theta}'_{pa_l} = 1, \quad (3.27)$$

which follows from that for $v_l \neq v_{lj}$, $N'_{v_l,pa_l} = 0$.

Theorem 10 *The likelihood $P(\mathbb{D}_{TP}|G, do(v_{lj}), \xi)$ given by Eq. (3.26) is V_l -transition likelihood equivalent.*

Proof: The same proof for Theorem 9. □

3.5.5 Combining various types of dynamic data

So far we have only considered the situations with two datasets. The discussions can be easily extended to the situations with a sequence of datasets, generated from a TS. Let $\mathbb{D} = \{D^0, D^1, \dots, D^k\}$ be a sequence of datasets generated from some causal graph G with parameters $\Theta_G^0, \dots, \Theta_G^k$ respectively, and let $\Xi_G = \cup_{i=0}^k \Theta_G^i$. The marginal likelihood is computed as

$$P(\mathbb{D}|G, \xi) = \int P(\mathbb{D}|\Xi_G, G, \xi) P(\Xi_G|G, \xi) d\Xi_G. \quad (3.28)$$

The term $P(\mathbb{D}|\Xi_G, G, \xi)$ can be computed as in Eq. (3.13). To give an appropriate parameter prior $P(\Xi_G|G, \xi)$, we need to know how these datasets in \mathbb{D} came about. Assume that we have the knowledge that the sequence of datasets, which will now be denoted by \mathbb{D}_{TS} , are from a TS with a sequence of focal variables $F = (V_{i_1}, \dots, V_{i_k})$. Then, we assume the following prior:

$$\begin{aligned} P(\Xi_G|G, F, \xi) &= P(\Theta_G^0|G, \xi) \left(P(\Psi_{i_1}^1|G, \xi) \prod_{i \neq i_1} \delta(\Psi_i^1 - \Psi_i^0) \right) \\ &\quad \left(P(\Psi_{i_2}^2|G, \xi) \prod_{i \neq i_2} \delta(\Psi_i^2 - \Psi_i^1) \right) \\ &\quad \cdots \left(P(\Psi_{i_k}^k|G, \xi) \prod_{i \neq i_k} \delta(\Psi_i^k - \Psi_i^{k-1}) \right), \end{aligned} \quad (3.29)$$

where we have used the notation $\Theta_G^j = \cup_{i=1}^n \Psi_i^j$, $j = 0, \dots, k$ as before. Eq. (3.29) is an extension of Eq. (3.14), and says that the set of parameters Θ_G^j differs with

Θ_G^{j-1} only by the parameters in $\Psi_{i_j}^j$. Let $I = \{i_1, \dots, i_k\}$ be the set of indexes for focal variables. Using the Dirichlet priors, we obtain the following expression for the marginal likelihood (3.28):

$$\begin{aligned} P(\mathbb{D}_{TS}|G, F, \xi) &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_{l=1}^k \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + M_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + M_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})} \\ &\quad \times \prod_{l=1}^k \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + L_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + L_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}, \end{aligned} \quad (3.30)$$

where

$$\begin{aligned} M_{v_i,pa_i}^l &= \sum_{j=0}^{l-1} N_{v_i,pa_i}^j, \quad M_{v_i,pa_i} = M_{v_i,pa_i}^{k+1}, \quad L_{v_i,pa_i}^l = \sum_{j=l}^k N_{v_i,pa_i}^j, \\ M_{pa_i}^l &= \sum_{v_i} M_{v_i,pa_i}^l, \quad M_{pa_i} = \sum_{v_i} M_{v_i,pa_i}, \quad L_{pa_i}^l = \sum_{v_i} L_{v_i,pa_i}^l, \end{aligned}$$

and N_{v_i,pa_i}^j is the number of cases in the dataset D^j for which V_i takes the value v_i and its parents Pa_i takes the value pa_i . Note that $M_{v_i,pa_i} = L_{v_i,pa_i}^l + M_{v_i,pa_i}^l$ is the number of cases in the whole dataset \mathbb{D}_{TS} for which V_i takes the value v_i and its parents Pa_i takes the value pa_i . We will call the Bayesian scoring metric $P(\mathbb{D}_{TS}, G|F, \xi) = P(\mathbb{D}_{TS}|G, F, \xi)P(G|\xi)$ (with parameters $\alpha_{v_i;pa_i}$ satisfying the likelihood equivalence property) the BDe-TS metric.

A marginal likelihood $P(\mathbb{D}|G, \xi)$ is said to satisfy the property of *F-transition likelihood equivalence* if for two *F*-transition equivalent causal graphs G_1 and G_2 , $P(\mathbb{D}|G_1, \xi) = P(\mathbb{D}|G_2, \xi)$.

Theorem 11 *The marginal likelihood $P(\mathbb{D}_{TS}|G, F, \xi)$ given by Eq. (3.30) is *F*-transition likelihood equivalent.*

Proof: Similar to the proof of Theorem 9. □

Assume that a series of mechanism changes occurred to a same causal model $M = \langle G, \Theta_G^0 \rangle$, and let $F = (V_{i_1}, \dots, V_{i_k})$ denote the sequence of focal variables, and $P_{ES} = (P^0, P^1, \dots, P^k)$ the corresponding sequence of distributions, where each pair (P^0, P^j) is a TP with V_{i_j} as the focal variable. We will call the pair (P_{ES}, F) an *experimental sequence (ES)*. An example of an ES is a series of experimental studies performed on a model. Now assume that we have the knowledge

that the sequence of datasets, which will now be denoted by \mathbb{D}_{ES} , are from an ES with the focal variables $F = (V_{i_1}, \dots, V_{i_k})$. We then assume the following prior:

$$\begin{aligned} P(\Xi_G | G, F, \xi) &= P(\Theta_G^0 | G, \xi) \left(P(\Psi_{i_1}^1 | G, \xi) \prod_{i \neq i_1} \delta(\Psi_i^1 - \Psi_i^0) \right) \\ &\quad \left(P(\Psi_{i_2}^2 | G, \xi) \prod_{i \neq i_2} \delta(\Psi_i^2 - \Psi_i^0) \right) \\ &\quad \cdots \left(P(\Psi_{i_k}^k | G, \xi) \prod_{i \neq i_k} \delta(\Psi_i^k - \Psi_i^0) \right). \end{aligned} \quad (3.31)$$

Eq. (3.31) is also an extension of Eq. (3.14), and says that the set of parameters Θ_G^j differs with Θ_G^0 only by the parameters in $\Psi_{i_j}^j$. Using the Dirichlet distribution, the marginal likelihood is given by

$$\begin{aligned} P(\mathbb{D}_{ES} | G, F, \xi) &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_{l=1}^k \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + K_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + K_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})} \\ &\quad \times \prod_{l=1}^k \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + N_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + N_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}, \end{aligned} \quad (3.32)$$

where

$$K_{v_{i_l},pa_{i_l}}^l = M_{v_{i_l},pa_{i_l}} - N_{v_{i_l},pa_{i_l}}^l, \quad K_{pa_{i_l}}^l = \sum_{v_{i_l}} K_{v_{i_l},pa_{i_l}}^l. \quad (3.33)$$

A special case of ES is a series of experimental studies in which each variable in F is fixed to some value respectively. Then we use the prior given in Eq. (3.25) for $P(\Psi_{i_j}^j | G, \xi)$, $j = 1, \dots, k$, and we obtain

$$\begin{aligned} P(\mathbb{D}_{ES} | G, do(F), \xi) &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\quad \times \prod_{l=1}^k \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + K_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + K_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}. \end{aligned} \quad (3.34)$$

Eq. (3.34) has been given in [CY99].

Theorem 12 *The marginal likelihood $P(\mathbb{D}_{ES} | G, F, \xi)$ in (3.32) and $P(\mathbb{D}_{ES} | G, do(F), \xi)$ in (3.34) is F -transition likelihood equivalent.*

Table 3.2: The posteriors of edges in the *Cancer* network.

N	$\delta = 0.1, B$ as the focal variable.									
	$P(A \rightarrow B \mathbb{D})$		$P(A \rightarrow C \mathbb{D})$		$P(B \rightarrow D \mathbb{D})$		$P(C \rightarrow D \mathbb{D})$		$P(C \rightarrow E \mathbb{D})$	
	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe
100	0.138	0.419	0.103	0.0394	0.997	0.87	0.853	0.86	0.552	0.441
200	0.335	0.482	0.354	0.136	1	0.993	0.983	0.993	0.607	0.403
500	0.604	0.686	0.43	0.457	1	0.999	0.996	1	0.713	0.728
1000	0.999	0.733	0.338	0.49	1	1	1	1	0.667	0.74
2000	1	0.75	0.336	0.5	1	1	1	1	0.666	0.75

$\delta = 0.5, B$ as the focal variable.										
N	$P(A \rightarrow B \mathbb{D})$		$P(A \rightarrow C \mathbb{D})$		$P(B \rightarrow D \mathbb{D})$		$P(C \rightarrow D \mathbb{D})$		$P(C \rightarrow E \mathbb{D})$	
	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe
100	0.999	0.238	0.0325	0.0141	1	0.484	0.284	0.293	0.0733	0.239
200	1	0.289	0.212	0.0516	1	0.663	0.83	0.546	0.0476	0.0106
500	1	0.658	0.495	0.651	1	0.992	1	0.989	0.0476	0.00518
1000	1	0.726	0.342	0.547	1	1	1	1	0.645	0.538
2000	1	0.75	0.334	0.5	1	1	1	1	0.666	0.75

$\delta = 0.1, A$ as the focal variable.										
N	$P(A \rightarrow B \mathbb{D})$		$P(A \rightarrow C \mathbb{D})$		$P(B \rightarrow D \mathbb{D})$		$P(C \rightarrow D \mathbb{D})$		$P(C \rightarrow E \mathbb{D})$	
	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe
100	0.832	0.471	0.226	0.106	0.979	0.911	0.958	0.84	0.477	0.441
200	0.827	0.494	0.278	0.0367	0.985	0.978	0.964	0.972	0.389	0.206
500	0.997	0.747	0.961	0.505	1	1	1	1	0.697	0.736
1000	0.995	0.75	0.948	0.5	1	1	1	1	0.961	0.75
2000	1	0.75	0.99	0.5	1	1	1	1	0.986	0.75

$\delta = 0.5, A$ as the focal variable.										
N	$P(A \rightarrow B \mathbb{D})$		$P(A \rightarrow C \mathbb{D})$		$P(B \rightarrow D \mathbb{D})$		$P(C \rightarrow D \mathbb{D})$		$P(C \rightarrow E \mathbb{D})$	
	BDe_TP	BDe								
100	1	0.586	0.832	0.57	0.999	0.916	0.961	0.878	0.0882	0.0171
200	1	0.676	0.992	0.642	1	0.999	1	0.999	0.47	0.113
500	1	0.746	1	0.507	1	1	1	1	0.963	0.739
1000	1	0.744	1	0.513	1	1	1	1	0.932	0.731
2000	1	0.75	1	0.5	1	1	1	1	0.994	0.75

Proof: Similar to the proof of Theorem 9. □

In deriving Eqs (3.30), (3.32), and (3.34), we have assumed that mechanism changes occurred at different variables. The situations in which different mechanism changes happen at a same variable can be easily incorporated. For example, in experimental studies, we may set a variable to different values. For this case, Eq. (3.34) is still applicable while $K_{v_i, p_{a_i}}^l$ as expressed in Eq. (3.33) should exclude all experimental data for which V_{i_l} is set to some fixed value.

In summary, to compute the marginal likelihood for dynamic data, we just need to provide an appropriate prior $P(\Xi_G|G, \xi)$ to reflect our knowledge on how those data came about. We demonstrated this method with several priors given in Eqs. (3.14), (3.29), (3.31) and (3.25).

3.5.6 Experimental results

We tested the BDe_TP score with data generated from a known network, the *Cancer* Bayesian network.³ We assumed a uniform prior distribution over all possible network structures. We used the parameters: $\alpha_{v_i;pa_i} = 1/r_i q_i$, where r_i is the number of states of V_i and q_i is the number of states of Pa_i , which satisfies the likelihood-equivalence requirement [Gei95].

A mechanism change at a variable V_i is simulated as follows. Consider parameters in $\vec{\theta}_{pa_i}$. If $\theta_{v_{i1};pa_i} \leq 0.5$ then let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} + \delta$, else let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} - \delta$, where δ is a parameter for adjusting the change magnitude. The rest of the parameters in $\vec{\theta}_{pa_i}$ are changed in proportional to their original values as: $\theta'_{v_{ij};pa_i} = \alpha \theta_{v_{ij};pa_i}$, $j = 2, \dots, r_i$, where $\alpha = (1 - \theta'_{v_{i1};pa_i})/(1 - \theta_{v_{i1};pa_i})$. When we simulate a mechanism change at V_i , we change parameters in $\vec{\theta}_{pa_i}$ as above for each $pa_i \in Dm(Pa_i)$.

The *Cancer* network is shown in Figure 3.1(a). It has only 5 nodes, hence we can exhaustively go through all 29,281 possible structures to compute the Bayesian average of any hypothesis of interest and to find the graphs with the maximum posterior probabilities. We computed the probability of each edge in the true *Cancer* network as in Eq. (3.11), and compared the results given by the BDe_TP metric (3.21) with that by the BDe metric (3.24). We experimented with δ values of 0.1 and 0.5, and focal variables B and A respectively, and generated a TP dataset $\mathbb{D}_{TP} = \{D^0, D^1\}$ for each case by first generating 2000 cases from the original network as D^0 , then simulating a mechanism change, and finally generating another 2000 cases as D^1 .

The results are shown in Table 3.2 for the first N cases in the dataset (N from D^0 and N from D^1). When using the BDe metric, the *Cancer* network and its independence-equivalent graphs of Figure 3.1(b)-(d) obtain the maximum score when the sample size is large enough, and they obtain a much larger posterior than all other structures. $P(A \rightarrow B|\mathbb{D})$ goes to 0.75 because three of the four graphs of Figure 3.1(a)-(d) have the edge $A \rightarrow B$ and we assumed a uniform distribution over structures. For the same reason, with the BDe metric, $P(A \rightarrow C|\mathbb{D})$ goes to 1/2, $P(B \rightarrow D|\mathbb{D})$ and $P(C \rightarrow D|\mathbb{D})$ goes to 1, and $P(C \rightarrow E|\mathbb{D})$ goes to 3/4. When using the BDe_TP metric and B as the focal variable, the posterior over structures concentrated sharply around the three B -transition equivalent graphs of Figure 3.1(e)-(g) when the sample size is large. Hence with the increasing sample size, $P(A \rightarrow B|\mathbb{D})$ goes to 1, $P(A \rightarrow C|\mathbb{D})$ goes to 1/3, and $P(C \rightarrow E|\mathbb{D})$ goes to 2/3. With A as the focal variable, the BDe_TP score concentrated sharply

³We used the version downloaded from the web site of Norsys Software Corporation, <http://www.norsys.com>.

around the unique *Cancer* network (see Figure 3.1(h)) for large sample size, and the posteriors of all five edges go to 1.

3.6 Conclusion

We proposed a new method of discovering causal structures, based on the detection of local, spontaneous changes in the underlying data-generating model. We analyzed the classes of structures that are equivalent relative to a stream of distributions produced by local changes, and devised algorithms that output graphical representations of these equivalence classes. We derived expressions for the Bayesian score that a causal structure should obtain from streams of data produced by locally changing distributions.

We have demonstrated, using simulated data, that the use of information about local changes may improve the power of discovery up to the theoretical limits set by statistical indistinguishability. The major advantage of the Bayesian treatment of local changes in Section 3.5, vis-a-vis the purely topological approach in Section 3.4, lies in that the Bayesian score is less sensitive to topological errors (e.g., remote descendants of focal variables that do not change). On the other hand, the Bayesian method is more computation intensive; hybrid schemes remain to be investigated.

CHAPTER 4

Testable Implications of Causal Models

4.1 Introduction

It is known that the statistical information encoded in a causal model is completely captured by conditional independence relationships among the variables when all variables are observable [PGV90]. However, when a causal model invokes unobserved variables, or hidden variables, the network structure may impose equality and inequality constraints on the distribution of the observed variables, and those constraints may not be expressed as conditional independencies [SGS93, Pea95b]. [VP90] gave an example of non-independence equality constraints shown in Figure 4.1(a), in which U is unobserved.¹ A simple analysis shows that the quantity $\sum_b P(d|a, b, c)P(b|a)$ is not a function of a , i.e.,

$$\sum_b P(d|a, b, c)P(b|a) = f(c, d). \quad (4.1)$$

This constraint holds even though no restrictions are made on the domains of the variables involved and on the class of distributions involved. This chapter develops a systematic way of finding such functional constraints.

Finding non-independence constraints is useful both for empirically validating causal models and for distinguishing causal models with the same set of conditional independence relationships among the observed variables. For example, the two networks in Figure 4.1(a) and (b) encode the same set of independence statements (A is independent of C given B), but they are empirically distinguishable due to Verma's constraint (4.1). A structure-learning algorithm driven by conditional independence relationships would not be able to distinguish between the two models unless the constraint stated in Eq. (4.1) is tested and incorporated into the model-selection strategy.

Algebraic methods for finding equality and inequality constraints implied by Bayesian networks with hidden variables have been presented in [GM98, GM99]. Those methods assume a priori fixed domains and are limited to small networks

¹We use dashed arrows for edges connected to hidden variables.

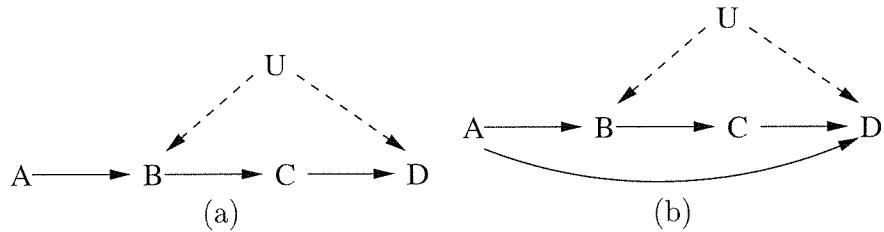


Figure 4.1: The network (a) imposes functional constraints; the network (b) encodes the same set of independence statements as (a) but does not impose functional constraints.

with small number of probabilistic parameters due to high computational demand. This chapter deals with conditional independence constraints and functional constraints, the type of constraints imposed by a network structure alone, regardless the domains of the variables and the class of distributions. The conditional independence constraints can be read via the d-separation criterion [Pea88], but there is no general graphical criterion available for Verma type functional constraints that are not captured by conditional independencies [RW97, Des99]. This chapter shows how the observed distribution factorizes according to the network structure, establishes relationships between this factorization and Verma-type constraints, and presents a procedure that systematically finds these constraints.

The chapter is organized as follows. Section 4.2 shows how functional constraints emerge in the presence of hidden variables. Section 4.3 shows how the observed distribution factorizes according to the network structure and introduces the concept of *c-component*, which plays a key role in identifying constraints. Section 4.4 presents a procedure for systematically identifying constraints. Section 4.5 shows that, for the purpose of finding constraints, instead of dealing with models with arbitrary hidden variables, we can work with a simplified model in which each hidden variable is a root node with two observed children. Section 4.6 concludes the chapter.

4.2 Functional Constraints

Letting $V = \{V_1, \dots, V_n\}$ and $U = \{U_1, \dots, U_{n'}\}$ stand for the sets of observed and hidden variables respectively, the observed probability distribution $P(v)$ is given by Eq. (1.3). Since all the factors of non-ancestors of V can be summed out from Eq. (1.3), letting U' be the set of variables in U that are ancestors of

V , Eq. (1.3) then becomes

$$P(v) = \sum_{u'} \prod_{V_i \in V} P(v_i | pa_{v_i}) \prod_{U_i \in U'} P(u_i | pa_{u_i}). \quad (4.2)$$

Therefore, we can remove from the network G all the hidden variables that are not ancestors of any V variables, and we will assume that each U variable is an ancestor of some V variable.

To illustrate how functional constraints emerge from the factorization of (4.2), we analyze the example in Figure 4.1(a). For any set $S \subseteq V$, let $Q[S](v)$ denote the following function²

$$Q[S](v) = \sum_u \prod_{\{i|V_i \in S\}} P(v_i | pa_{v_i}) \prod_{\{i|U_i \in U\}} P(u_i | pa_{u_i}). \quad (4.3)$$

In particular, we have $Q[V](v) = P(v)$ and, for consistency, we set $Q[\emptyset](v) = 1$, since $\sum_u \prod_{\{i|U_i \in U\}} P(u_i | pa_{u_i}) = 1$. For convenience, we will often write $Q[S](v)$ as $Q[S]$. For Figure 4.1(a), Eq. (4.2) becomes

$$P(a, b, c, d) = P(a)P(c|b)Q[\{B, D\}], \quad (4.4)$$

where

$$Q[\{B, D\}] = \sum_u P(b|a, u)P(d|c, u)P(u). \quad (4.5)$$

From (4.4), we obtain

$$Q[\{B, D\}] = \frac{P(a, b, c, d)}{P(a)P(c|b)} = P(d|a, b, c)P(b|a), \quad (4.6)$$

and from (4.5),

$$Q[\{D\}] = \sum_u P(d|c, u)P(u) \quad (4.7)$$

$$\begin{aligned} &= \sum_b Q[\{B, D\}] \\ &= \sum_b P(d|a, b, c)P(b|a). \end{aligned} \quad (4.8)$$

Eq. (4.7) implies that $Q[\{D\}]$ is a function only of c and d , therefore Eq. (4.8) induces a constraint that the quantity $\sum_b P(d|a, b, c)P(b|a)$ is independent of a .

² $Q[S](v)$ can be interpreted as $Q[S](v) = P_{v \setminus s}(s)$.

Note that the key to obtaining this constraint rests with our ability to express $Q[\{B, D\}]$ and $Q[\{D\}]$ in terms of observed quantities (see (4.6) and (4.8)), namely quantities not involving U . Applying the same analyses to Figure 4.1(b), we have that $Q[\{D\}]$ gives the same expression as in Eq. (4.8), but now $Q[\{D\}] = \sum_u P(d|c, a, u)P(u)$ is also a function of a , and no Verma constraint is induced. In general, for any set $S \subset V$, $Q[S]$ in Eq. (4.3) is a function of values only of a subset of V . Therefore, whenever $Q[S]$ is computable from the observational distribution $P(v)$, it may lead to some constraints — conditional independence relations or Verma-type functional constraints. In the rest of the chapter, we will show how to systematically find computable $Q[S]$, but first, we study what the arguments of $Q[S]$ are.

For any set C , let G_C denote the subgraph of G composed only of variables in C , let $An(C)$ denote the union of C and the set of ancestors of the variables in C , and let $An^u(C) = An(C) \cap U$ denote the set of hidden variables in $An(C)$. In Eq. (4.3), the factors corresponding to the hidden variables that are not ancestors of S in the subgraph $G_{S \cup U}$ can be summed out, and letting $U(S) = An^u(S)_{G_{S \cup U}}$ be the set of hidden variables that are ancestors of S in the graph $G_{S \cup U}$, $Q[S]$ can be written as

$$Q[S] = \sum_{u(S)} \prod_{\{i|V_i \in S\}} P(v_i|pa_{v_i}) \prod_{\{i|U_i \in U(S)\}} P(u_i|pa_{u_i}). \quad (4.9)$$

We see that $Q[S]$ is a function of S , the observed parents of S , and the observed parents of $U(S)$. We will call an observed variable V_i an *effective parent* of an observed variable V_j if V_i is a parent of V_j or if there is a directed path from V_i to V_j in G such that every internal node on the path is a hidden variable. For any set $S \subseteq V$, letting $Pa^+(S)$ denote the union of S and the set of effective parents of the variables in S , then we have that $Q[S]$ is a function of $Pa^+(S)$. Assuming that $Q[S]$ is a function of some set T , when $Q[S](t)$ is computable from $P(v)$, its expression obtained may be a function of values of some set T' larger than T ($T \subset T'$), and this will lead to constraints on the distribution $P(v)$ that the expression obtained for $Q[S]$ is independent of the values $t' \setminus t$, which could be a Verma-type functional constraint or be a set of conditional independence statements.

Next we give a lemma that will facilitate the computation of $Q[S]$ and the proof of other propositions. The lemma provides a condition under which we can compute $Q[W]$ from $Q[C]$, where W is a subset of C , by simply summing $Q[C]$ over the remaining variables (in $C \setminus W$). For any set C , let $An^v(C) = An(C) \cap V$ be the set of observed variables in $An(C)$, and let $De^v(C)$ denote the set of observed variables that are in C or are descendants of any variable in C . A set $A \subseteq V$ is called an *ancestral set* if it contains its own observed ancestors

$(A = An^v(A))$, and a set $A \subseteq V$ is called a *descendent set* if it contains its own observed descendants $(A = De^v(A))$. Letting $G(C) = G_{C \cup U(C)}$ denote the subgraph of G composed only of variables in C and $U(C)$ which corresponds to the quantity $Q[C]$ (see Eq. (4.9)), then we have the following lemma.

Lemma 2 *Let $W \subseteq C \subseteq V$, and $W' = C \setminus W$. If W is an ancestral set in $G(C)$ ($W = An^v(W)_{G(C)}$), or equivalently, if W' is a descendent set in $G(C)$ ($W' = De^v(W')_{G(C)}$), then*

$$\sum_{w'} Q[C] = Q[W]. \quad (4.10)$$

Proof sketch: By Eq. (4.9)

$$\sum_{w'} Q[C] = \sum_{w'} \sum_{u(C)} \prod_{V_i \in C} P(v_i | pa_{v_i}) \prod_{U_i \in U(C)} P(u_i | pa_{u_i}). \quad (4.11)$$

All factors in (4.11) corresponding to the variables (observed or hidden) that are not ancestors of W in $G(C)$ are summed out, and we obtain

$$\sum_{w'} Q[C] = \sum_{An^u(W)_{G(C)}} \prod_{V_i \in W} P(v_i | pa_{v_i}) \prod_{U_i \in An^u(W)_{G(C)}} P(u_i | pa_{u_i}). \quad (4.12)$$

We have $An^u(W)_{G(C)} = An^u(W)_{G_{W \cup U}} = U(W)$ due to that W is an ancestral set. Therefore the left hand side of (4.12) is equal to $Q[W]$ by Eq. (4.9). \square

In the next section, we show how the distribution $P(v)$ decomposes according to the network structure and how the decomposition helps the computation of $Q[S]$.

4.3 C-components

$P(v)$ as a summation of products in (4.2) may sometimes be decomposed into a product of summations. For example, in Figure 4.2, $P(v)$ can be written as

$$\begin{aligned} P(v_1, v_2, v_3, v_4) &= \left(\sum_{u_1} P(v_1 | u_1) P(v_3 | v_2, u_1) P(u_1) \right) \\ &\quad \left(\sum_{u_2, u_3} P(v_2 | u_2, u_3) P(v_4 | v_3, u_2) P(u_2) P(u_3 | v_1) \right) \\ &= Q[\{V_1, V_3\}] Q[\{V_2, V_4\}] \end{aligned} \quad (4.13)$$

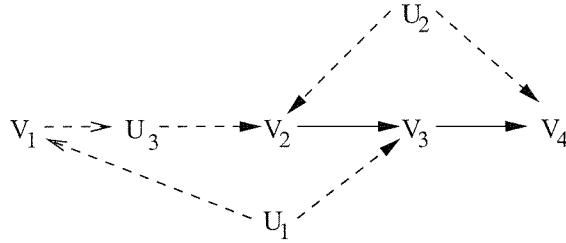


Figure 4.2: The graph is partitioned into c-components $\{V_1, V_3\}$ and $\{V_2, V_4\}$.

The importance of this decomposition lies in that both terms $Q[\{V_1, V_3\}]$ and $Q[\{V_2, V_4\}]$ are computable from $P(v)$ as shown later. First we study graphical conditions under which this kind of decomposition is feasible.

Assume that $P(v)$ in Eq. (4.2) can be decomposed into a product of summations as:

$$P(v) = \prod_{V_i \in S^0} P(v_i | pa_{v_i}) \prod_j \left(\sum_{n_j} \prod_{V_i \in S_j} P(v_i | pa_{v_i}) \prod_{U_i \in N_j} P(u_i | pa_{u_i}) \right) \quad (4.14)$$

where the variables in S^0 have no hidden parents, U is partitioned into N_j 's, and $V \setminus S^0$ is partitioned into S_j 's. U_i and U_j must be in the same set N_k if (i) there is an edge between them ($U_i \rightarrow U_j$ or $U_i \leftarrow U_j$), or (ii) they have a common child ($U_i \rightarrow U_l \leftarrow U_j$ or $U_i \rightarrow V_l \leftarrow U_j$). Repeatedly applying these two rules, we obtain that U_i and U_j are in the same set N_k if there exists a path between U_i and U_j in G such that (i) every internal node of the path is in U , or (ii) every node in V on the path is head-to-head ($\rightarrow V_l \leftarrow$). It is clear that this relation among U_i 's is reflexive, symmetric, and transitive, and therefore it defines a partition of U . We construct S_i as follows: a variable $V_k \in V$ is in S_i if it has a hidden parent that is in N_i . S_i 's form a partition of $V \setminus S^0$ since N_i 's form a partition of U . Let each variable $V_i \in S^0$ form a set by itself $S_i^0 = \{V_i\}$. We have that S_i 's and S_i^0 's form a partition of V . It is clear that if a hidden variable U_k is not in N_j , then it does not appear in the factors of $\prod_{V_i \in S_j} P(v_i | pa_{v_i}) \prod_{U_i \in N_j} P(u_i | pa_{u_i})$, hence the decomposition of $P(v)$ in Eq. (4.14) follows. We will call each S_i or S_i^0 a *c-component* (abbreviating “confounded component”) of V in G or simply c-component of G .

Assuming that V is partitioned into c-components S_1, \dots, S_k , Eq. (4.14) can be rewritten as

$$P(v) = Q[V] = \prod_i Q[S_i], \quad (4.15)$$

which follows from

$$\begin{aligned}
Q[S_j] &= \sum_u \prod_{\{i|V_i \in S_j\}} P(v_i|pa_{v_i}) \prod_{\{i|U_i \in U\}} P(u_i|pa_{u_i}) \\
&= \sum_{n_j} \prod_{V_i \in S_j} P(v_i|pa_{v_i}) \prod_{U_i \in N_j} P(u_i|pa_{u_i}) \sum_{u \setminus n_j} \prod_{U_i \in U \setminus N_j} P(u_i|pa_{u_i}) \\
&= \sum_{n_j} \prod_{V_i \in S_j} P(v_i|pa_{v_i}) \prod_{U_i \in N_j} P(u_i|pa_{u_i}),
\end{aligned} \tag{4.16}$$

where we have used the following formula

$$\sum_w \prod_{\{i|U_i \in W\}} P(u_i|pa_{u_i}) = 1, \text{ for any } W \subseteq U. \tag{4.17}$$

We will call $Q[S_i]$ the *c-factor* corresponding to the c-component S_i . For example, Figure 4.1(a) is partitioned into c-components $\{A\}$, $\{C\}$, and $\{B, D\}$, with corresponding c-factors $Q[\{A\}] = P(a)$, $Q[\{C\}] = P(c|b)$, and $Q[\{B, D\}]$ in (4.5) respectively, and $P(v)$ can be written as a product of c-factors as in Eq. (4.4). In Figure 4.2, V is partitioned into c-components $\{V_1, V_3\}$ and $\{V_2, V_4\}$, and $P(v)$ can be written as a product of c-factors $Q[\{V_1, V_3\}]$ and $Q[\{V_2, V_4\}]$ as in (4.13).

The importance of the c-factors stems from that all c-factors are computable from $P(v)$. We generalize this result to proper subgraphs of G and obtain the following lemma.

Lemma 3 *Let $H \subseteq V$, and assume that H is partitioned into c-components H_1, \dots, H_l in the subgraph $G(H) = G_{H \cup U(H)}$. Then we have*

(i) $Q[H]$ decomposes as

$$Q[H] = \prod_i Q[H_i]. \tag{4.18}$$

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

$$Q[H_j] = \prod_{\{i|V_{h_i} \in H_j\}} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]}, \tag{4.19}$$

where each $Q[H^{(i)}]$, $i = 0, 1, \dots, k$, is given by

$$Q[H^{(i)}] = \sum_{h \setminus h^{(i)}} Q[H]. \tag{4.20}$$

(iii) Each $Q[H^{(i)}]/Q[H^{(i-1)}]$ is a function only of $Pa^+(T_i)$, where T_i is the c-component of the subgraph $G(H^{(i)})$ that contains V_{h_i} .

Proof: (i) The decomposition of $Q[H]$ into Eq. (4.18) follows directly from the definition of c-component (see Eqs. (4.14)–(4.17)).

(ii)&(iii) Eq. (4.20) follows from Lemma 2 since each $H^{(i)}$ is an ancestral set. We prove (ii) and (iii) simultaneously by induction on k .

Base: $k = 1$. There is one c-component $Q[H_1] = Q[H] = Q[H^{(1)}]$ which satisfies Eq. (4.19) because $Q[\emptyset] = 1$, and $Q[H_1]$ is a function of $Pa^+(H_1)$.

Hypothesis: When there are k variables in H , all $Q[H_i]$'s are computable from $Q[H]$ and are given by Eq. (4.19), and (iii) holds for i from 1 to k .

Induction step: When there are $k + 1$ variables in H , assuming that the c-components of $G(H)$ are H_1, \dots, H_m, H' , and that $V_{h_{k+1}} \in H'$, we have

$$Q[H] = Q[H^{(k+1)}] = Q[H'] \prod_i Q[H_i]. \quad (4.21)$$

Summing both sides of (4.21) over $V_{h_{k+1}}$ leads to

$$\sum_{v_{h_{k+1}}} Q[H] = Q[H^{(k)}] = \left(\sum_{v_{h_{k+1}}} Q[H'] \right) \prod_i Q[H_i], \quad (4.22)$$

where we have used Lemma 2. It is clear that each H_i , $i = 1, \dots, m$, is a c-component of the subgraph $G(H^{(k)})$. Then by the induction hypothesis, each $Q[H_i]$, $i = 1, \dots, m$, is computable from $Q[H^{(k)}] = \sum_{v_{h_{k+1}}} Q[H]$ and is given by Eq. (4.19), where each $Q[H^{(i)}]$, $i = 0, 1, \dots, k$, is given by

$$Q[H^{(i)}] = \sum_{h^{(k)} \setminus h^{(i)}} Q[H^{(k)}] = \sum_{h \setminus h^{(i)}} Q[H]. \quad (4.23)$$

From Eq. (4.21), $Q[H']$ is computable as well, and is given by

$$Q[H'] = \frac{Q[H^{(k+1)}]}{\prod_i Q[H_i]} = \prod_{\{i | V_{h_i} \in H'\}} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]}, \quad (4.24)$$

which is clear from (4.19) and the chain decomposition $Q[H^{(k+1)}] = \prod_{i=1}^{k+1} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]}$.

By the induction hypothesis, (iii) holds for i from 1 to k . Next we prove that it holds for $Q[H^{(k+1)}]/Q[H^{(k)}]$. The c-component of G that contains $V_{h_{k+1}}$ is H' . In Eq. (4.24), $Q[H']$ is a function of $Pa^+(H')$, and each term $Q[H^{(i)}]/Q[H^{(i-1)}]$, $V_{h_i} \in H'$ and $V_{h_i} \neq V_{h_{k+1}}$, is a function of $Pa^+(T_i)$, where T_i is a c-component

of the graph $G(H^{(i)})$ that contains V_{h_i} and therefore is a subset of H' . Hence we obtain that $Q[H^{(k+1)}]/Q[H^{(k)}]$ is a function only of $Pa^+(H')$. \square

The proposition (iii) in Lemma 3 may imply a set of constraints to the distribution $P(v)$ whenever $Q[H]$ is computable from $P(v)$.

A special case of Lemma 3 is when $H = V$, and we obtain the following corollary.

Corollary 1 *Assuming that V is partitioned into c-components S_1, \dots, S_k , we have*

$$(i) P(v) = \prod_i Q[S_i].$$

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

$$Q[S_j] = \prod_{\{i | V_i \in S_j\}} P(v_i | v^{(i-1)}) \quad (4.25)$$

(iii) *Each factor $P(v_i | v^{(i-1)})$ can be expressed as*

$$P(v_i | v^{(i-1)}) = P(v_i | pa^+(T_i) \setminus \{v_i\}), \quad (4.26)$$

where T_i is the c-component of $G(V^{(i)})$ that contains V_i .

We see that when hidden variables were invoked, a variable is independent of its non-descendants given its effective parents, the non-descendant variables in its c-component, and the effective parents of the non-descendant variables in its c-component, reminiscence of the property that each variable is independent of its non-descendants given its parents when there is no hidden variables.

4.4 Finding Constraints

With Lemma 2, 3, and Corollary 1, we can systematically find constraints implied by a network structure. First we study a few examples.

4.4.1 Examples

Consider Figure 4.2, which has two c-components $\{V_1, V_3\}$ and $\{V_2, V_4\}$. The only admissible order is $V_1 < V_2 < V_3 < V_4$. Applying Corollary 1, we obtain that the two c-factors are given by

$$Q[\{V_1, V_3\}](v_1, v_2, v_3) = P(v_3 | v_2, v_1)P(v_1), \quad (4.27)$$

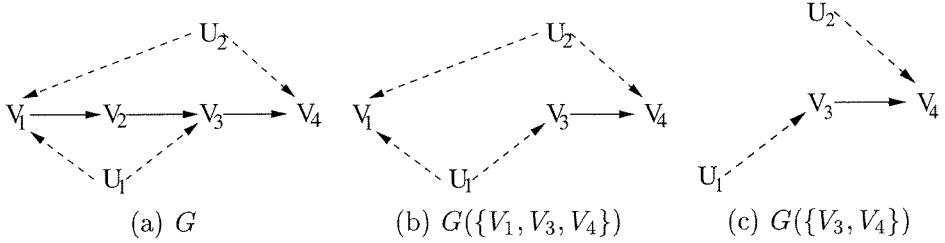


Figure 4.3: Subgraphs for finding constraints.

and

$$Q[\{V_2, V_4\}](v_1, v_2, v_3, v_4) = P(v_4|v_3, v_2, v_1)P(v_2|v_1). \quad (4.28)$$

They do not imply any constraints on the distribution. Summing both sides of (4.28) over V_2 , by Lemma 2, we obtain

$$Q[\{V_4\}](v_3, v_4) = \sum_{v_2} P(v_4|v_3, v_2, v_1)P(v_2|v_1), \quad (4.29)$$

which implies a constraint on the distribution $P(v)$ that the right hand side is independent of v_1 . Computing $Q[\{V_1\}]$, $Q[\{V_2\}]$, and $Q[\{V_3\}]$ does not give any constraints.

Consider Figure 4.3(a), which has two c-components $\{V_2\}$ and $S = \{V_1, V_3, V_4\}$. The only admissible order is $V_1 < V_2 < V_3 < V_4$. Applying Corollary 1, we obtain

$$Q[\{V_2\}](v_1, v_2) = P(v_2|v_1), \quad (4.30)$$

$$Q[S](v) = P(v_4|v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1). \quad (4.31)$$

In the subgraph $G(S) = G_{S \cup U}$ (Figure 4.3(b)), V_1 is not an ancestor of $H = \{V_3, V_4\}$, and from Lemma 2, summing both sides of (4.31) over V_1 , we obtain

$$Q[H](v_2, v_3, v_4) = \sum_{v_1} P(v_4|v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1). \quad (4.32)$$

The subgraph $G(H) = G_{H \cup U}$ (Figure 4.3(c)) has two c-components $\{V_3\}$ and $\{V_4\}$. By Lemma 3, we have $Q[H] = Q[\{V_3\}]Q[\{V_4\}]$, and

$$Q[\{V_3\}](v_2, v_3) = \sum_{v_4} Q[H] = \sum_{v_1} P(v_3|v_2, v_1)P(v_1), \quad (4.33)$$

$$Q[\{V_4\}](v_3, v_4) = \frac{Q[H]}{\sum_{v_4} Q[H]} = \frac{\sum_{v_1} P(v_4|v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1)}{\sum_{v_1} P(v_3|v_2, v_1)P(v_1)}. \quad (4.34)$$

Eq. (4.34) implies a constraint on $P(v)$ that the right hand side is independent of v_2 .

From the preceding examples, we see that we may find constraints by alternatively applying Lemma 2 and 3. Next, we present a procedure that systematically looking for constraints.

4.4.2 Identifying constraints systematically

Let a topological order over V be $V_1 < \dots < V_n$, and let $V^{(i)} = \{V_1, \dots, V_i\}$, $i = 1, \dots, n$. For i from 1 to n , at each step, we will look for constraints that involve V_i and the variables ordered before V_i . At step i , we do the following:

- (A1) Consider the subgraph $G(V^{(i)})$. If $G(V^{(i)})$ has more than one c-component, assuming that V_i is in the c-component S_i of $G(V^{(i)})$, then by Corollary 1, $Q[S_i]$ is computable from $P(v)$ and may give a conditional independence constraint that V_i is independent of its predecessors given its effective parents, other variables in S_i , and the effective parents of other variables in S_i , that is, V_i is independent of $V^{(i)} \setminus Pa^+(S_i)$ given $Pa^+(S_i) \setminus \{V_i\}$.
- (A2) Consider $Q[S_i]$ in the subgraph $G(S_i)$. For each descendent set $D \subset S_i$ (D contains its own observed descendants) in $G(S_i)$ that does not contain V_i ,³ by Lemma 2 we have

$$\sum_d Q[S_i] = Q[S_i \setminus D]. \quad (4.35)$$

The left hand side of (4.35) is a function of $Pa^+(S_i) \setminus D$, while the right hand side is a function of $Pa^+(S_i \setminus D) \subseteq Pa^+(S_i) \setminus D$. Therefore, if some effective parents of D are not effective parents of $S_i \setminus D$, then (4.35) implies a constraint on the distribution $P(v)$ that the quantity $\sum_d Q[S_i]$ is independent of $(Pa^+(S_i) \setminus D) \setminus Pa^+(S_i \setminus D)$.

Let $D' = S_i \setminus D$. Next we consider $Q[D']$ in the subgraph $G(D')$. If $G(D')$ has more than one c-component, assuming that V_i is in the c-component E_i of $G(D')$, by Lemma 3, $Q[E_i]$ is computable from $Q[D']$, and $Q[D'] / \sum_{v_i} Q[D']$ is a function only of $Pa^+(E_i)$, which imposes a constraint on $P(v)$ if $Pa^+(D') \setminus Pa^+(E_i) \neq \emptyset$.

³We need to consider every descendent set D that does not contain V_i , because it is possible that for two descendent sets $D_1 \subset D_2$, the constraints from summing D_2 are not implied by that from D_1 , and vice versa.

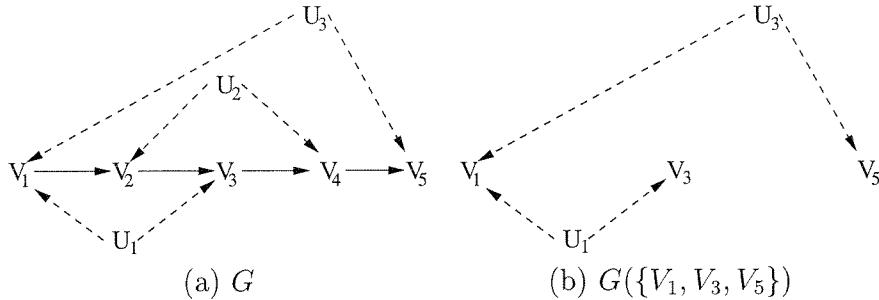


Figure 4.4: A model imposing functional constraints.

Finally we study $Q[E_i]$ by repeating the process (A2) with S_i now replaced by E_i .

The preceding analysis gives us a recursive procedure for systematically finding constraints. To illustrate this process, we consider the example in Figure 4.4(a). The only admissible order over V is $V_1 < \dots < V_5$. The constraints involving V_1 to V_4 are the same as in Figure 4.2, and here we look for constraints involving V_5 . V_5 is in the c-component $S = \{V_1, V_3, V_5\}$. By Corollary 1, $Q[S]$ is given by

$$Q[S](v) = P(v_5|v_4, v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1), \quad (4.36)$$

which implies no constraints. In the subgraph $G(S)$ (Figure 4.4(b)), the descendant sets not containing V_5 are $\{V_1\}$, $\{V_3\}$, and $\{V_1, V_3\}$.

(a) Summing both sides of (4.36) over v_1 , we obtain

$$Q[\{V_3, V_5\}](v_2, v_3, v_4, v_5) = \sum_{v_1} P(v_5|v_4, v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1), \quad (4.37)$$

which implies no constraints. The subgraph $G(\{V_3, V_5\})$ is partitioned into two c-components $\{V_3\}$ and $\{V_5\}$, and by Lemma 3, we have

$$\begin{aligned} Q[\{V_5\}](v_4, v_5) &= \frac{Q[\{V_3, V_5\}]}{\sum_{v_5} Q[\{V_3, V_5\}]} \\ &= \frac{\sum_{v_1} P(v_5|v_4, v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1)}{\sum_{v_1} P(v_3|v_2, v_1)P(v_1)}, \end{aligned} \quad (4.38)$$

which implies a constraint that the right hand side is independent of v_2 and v_3 .

(b) Summing both sides of (4.36) over v_3 , we obtain

$$Q[\{V_1, V_5\}](v_1, v_4, v_5) = \sum_{v_3} P(v_5|v_4, v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1), \quad (4.39)$$

which implies a constraint that the right hand side is independent of v_2 . $G(\{V_1, V_5\})$ can not be further partitioned into c-components.

(c) Summing both sides of (4.36) over v_1 and v_3 , we obtain

$$Q[\{V_5\}](v_4, v_5) = \sum_{v_1, v_3} P(v_5|v_4, v_3, v_2, v_1)P(v_3|v_2, v_1)P(v_1), \quad (4.40)$$

which implies a constraint that the right hand side is independent of v_2 . This constraint is implied by that obtained from Eq. (4.38).

4.5 Projection to Semi-Markovian Models

If, in a causal model with hidden variables, each hidden variable is a root node with exactly two observed children, then the corresponding model is a semi-Markovian model. The examples we have studied in Figure 4.1, 4.3, and 4.4 are semi-Markovian models while Figure 4.2 is not. Semi-Markovian models are easy to work with, and we will show that a causal model with arbitrary hidden variables can be converted to a semi-Markovian model with exactly the same set of constraints (that can be found through the procedure in Section 4.4.2) on the observed distribution $P(v)$.

In a semi-Markovian model, the observed distribution $P(v)$ is given by Eq. (1.5). And the function $Q[S](v)$ in (4.3) becomes

$$Q[S](v) = \sum_u \prod_{\{i|V_i \in S\}} P(v_i|pa_{v_i}) \prod_i P(u_i). \quad (4.41)$$

The appearance of hidden variables is represented by bidirected edges in the causal graph of a semi-Markovian model. It is easy to partition a graph with bidirected edges into c-components. Let a path composed entirely of bidirected edges be called a *bidirected path*. Two observed variables are in the same c-component if and only if they are connected by a bidirected path. Letting $Pa(S)$ denote the union of S and the set of parents of S , then it is clear that $Q[S]$ is a function of $Pa(S)$. For semi-Markovian models, Lemma 2 and 3 still hold, in which $G(C)$ ($G(H)$) will be replaced by G_C (G_H), and $Pa^+(\cdot)$ replaced by $Pa(\cdot)$.

A causal model with arbitrary hidden variables can be converted to a semi-Markovian model by constructing its *projection* [Ver93].

Definition 5 (Projection) *The projection of a DAG G over $V \cup U$ on the set V , denoted by $PJ(G, V)$, is a DAG over V with bidirected edges constructed as follows:*

1. Add each variable in V as a node of $PJ(G, V)$.
2. For each pair of variables $X, Y \in V$, if there is an edge between them in G , add the edge to $PJ(G, V)$.
3. For each pair of variables $X, Y \in V$, if there exists a directed path from X to Y in G such that every internal node on the path is in U , add edge $X \rightarrow Y$ to $PJ(G, V)$ (if it does not exist yet).
4. For each pair of variables $X, Y \in V$, if there exists a divergent path between X and Y in G such that every internal node on the path is in U ($X \leftarrow U_i \dashrightarrow Y$), add a bidirected edge $X \dashrightarrow Y$ to $PJ(G, V)$.

It is shown in [Ver93] that G and $PJ(G, V)$ have the same set of conditional independence relations among V . Next we show that the procedure presented in Section 4.4.2 will find the same sets of constraints on $P(v)$ in G and $PJ(G, V)$. To this purpose, we need to show that for any set $H \subseteq V$, G and $PJ(G, V)$ have the same arguments for $Q[H]$, the same topological relations over H , and the same sets of c-components.

Lemma 4 *For any set $H \subseteq V$, $Q[H]$ has the same arguments in G and $PJ(G, V)$, that is, $Pa^+(H)$ in G is equal to $Pa(H)$ in $PJ(G, V)$.*

Lemma 4 is obvious from Definition 5.

Lemma 5 *For any set $H \subseteq V$, and any two variables $V_i, V_j \in H$, V_i is an ancestor of V_j in $G(H)$ if and only if V_i is an ancestor of V_j in $PJ(G, V)_H$ (the subgraph of $PJ(G, V)$ composed only of variables in H).*

Lemma 5 has been shown in [Ver93].

Lemma 6 *For any set $H \subseteq V$, $G(H)$ is partitioned into the same set of c-components as $PJ(G, V)_H$.*

Proof: (1) If two variables $X, Y \in H$ are in the same c-component in $PJ(G, V)_H$, then there is a bidirected path between X and Y in $PJ(G, V)_H$:

$$X \dashrightarrow \dots \dashrightarrow V_i \dashrightarrow \dots \dashrightarrow Y$$

From the definition of a projection, there is a path between X and Y in $G(H)$ on which each observable is head-to-head:

$$X \leftarrow U_l \dashrightarrow V_j \leftarrow \dots \rightarrow V_i \leftarrow \dots \rightarrow V_k \leftarrow U_m \dashrightarrow Y$$

Therefore X and Y are in the same c-component in $G(H)$.

(2) If $X, Y \in H$ are in the same c-component in $G(H)$, then there exist U_i and U_j such that U_i is a parent of X , U_j is a parent of Y , and $U_i = U_j$ or there is a path p between U_i and U_j such that every observable on p is head-to-head and every hidden variable on p is in $U(H)$. We prove that X and Y are in the same c-component in $PJ(G, V)_H$ by induction on the number k of head-to-head nodes on p .

Base: $k = 0$. There is no head-to-head node on p , then there is a divergent path between X and Y in G :

$$X \leftarrow \dots \leftarrow U_k \rightarrow \dots \rightarrow Y.$$

Therefore there is a bidirected edge $X \dashv\vdash Y$ in $PJ(G, V)_H$, and X and Y are in the same c-component in $PJ(G, V)_H$.

Induction hypothesis: If there are k head-to-head nodes on p , X and Y are in the same c-component in $PJ(G, V)_H$.

If there are $k + 1$ head-to-head nodes on p , let W be the head-to-head node closest to X on p . If W is an observable, let $V_i = W$, otherwise let V_i be an observable descendant of W such that there is a directed path from W to V_i on which all internal nodes are hidden variables. From the base case, X and V_i are in the same c-component in $PJ(G, V)_H$, and from the induction hypothesis, V_i and Y are in the same c-component in $PJ(G, V)_H$, hence we have that X and Y are in the same c-component in $PJ(G, V)_H$. \square

By Lemma 4–6, we conclude that the procedure presented in Section 4.4.2 will find the same sets of constraints on $P(v)$ in G and $PJ(G, V)$. Since it is easier to work in a semi-Markovian model, we can always convert a Bayesian network with arbitrary hidden variables to a semi-Markovian model before searching for constraints on the distribution $P(v)$.

4.6 Conclusion

This chapter develops a systematic procedure of identifying functional constraints induced by causal Bayesian networks with hidden variables. The procedure can be used for devising tests for validating causal models, and for inferring the structures of such models from observed data. At this stage of research we cannot ascertain whether *all* functional constraints can be identified by our procedure; however, we could not rule out this possibility.

CHAPTER 5

Identification of Causal Effects

5.1 Introduction

This chapter explores the feasibility of inferring cause effect relationships from various combinations of data and theoretical assumptions. The assumptions considered will be represented in the form of an acyclic causal diagram which contains both arrows and bi-directed arcs [Pea95a, Pea00]. The arrows represent the potential existence of direct causal relationships between the corresponding variables, and the bi-directed arcs represent spurious dependencies due to unmeasured confounders. Our main task will be to decide whether the assumptions represented in any given diagram are sufficient for assessing the strength of causal effects from nonexperimental data and, if sufficiency is proven, to express the target causal effect in terms of estimable quantities.

It is well known that, in the absence of unmeasured confounders, all causal effects are *identifiable*, that is, the joint response of any set Y of variables to intervention on a set T of treatment variables $P_t(y)$ can be estimated consistently from nonexperimental data [Rob86, SGS93, Pea93]. If some confounders are not measured, then the question of identifiability arises, and whether the desired quantity can be estimated depends critically on the precise locations (in the diagram) of those confounders vis a vis the sets T and Y . Sufficient graphical conditions for ensuring the identification of $P_t(y)$ were established by several authors [SGS93, Pea93, Pea95a] and are summarized in [Pea00, Chapters 3 and 4]. For example, a criterion called “back-door” permits one to determine whether a given causal effect $P_t(y)$ can be obtained by “adjustment”, that is, whether a set C of covariates exists such that

$$P_t(y) = \sum_c P(y|c, t)P(c) \quad (5.1)$$

When there exists no set of covariates that is sufficient for adjustment, causal effects can sometimes be estimated by invoking multi-stage adjustments, through a criterion called “front-door” [Pea95a]. More generally, identifiability can be decided using *do*-calculus derivations [Pea95a], that is, a sequence of syntactic transformations capable of reducing expressions of the type $P_t(y)$ to subscript-free

expressions. Using *do*-calculus as a guide, [GP95] devised a graphical criterion for identifying $P_x(y)$ (where X and Y are singletons) that combines and expands the “front-door” and “back-door” criteria (see [Pea00, pp. 114-8]).¹ [PR95] further derived a graphical condition under which it is possible to identify $P_t(y)$ where T consists of an arbitrary set of variables. This permits one to predict the effect of time varying treatments from longitudinal data, in the presence of unmeasured confounders, some of which are affected by previous treatments. This criterion was further extended by [Rob97b] and [KM99].

This chapter develops new graphical identification criteria that generalize and simplify existing criteria in several ways. In Sections 5.2-5.5, we study the identifiability problem in semi-Markovian models. Section 5.2 concerns the identification of $P_x(v)$, where X is a singleton and V is the set of all variables excluding X . It asserts that $P_x(v)$ is identifiable if and only if there is no consecutive sequence of confounding arcs between X and X ’s children in the graph. When interest lies in the effect of X on a subset S of outcome variables, not on the entire set V , it is possible, however, that $P_x(s)$ would be identifiable even though $P_x(v)$ is not. Section 5.3 first gives a sufficient criterion for identifying $P_x(s)$, which is an extension of the criterion for identifying $P_x(v)$. It says that $P_x(s)$ is identifiable if there is no consecutive sequence of confounding arcs between X and X ’s children in the subgraph composed of the ancestors of S . Other than this requirement, the diagram may have an arbitrary structure, including any number of confounding arcs between X and S . This simple criterion is shown to cover all criteria reported in the literature (with X singleton), including the “back-door”, “front-door”, and those developed by [GP95]. However, the criterion is not necessary for identifying $P_x(s)$. Section 5.3 further devises a procedure for the identification and computation of $P_x(s)$, based on systematic removal of certain nonessential nodes from G . This procedure is shown to be more powerful than the one devised by [GP95] ([Pea00, pp. 114-8]). Section 5.4 deals with the identification of general causal effects, $P_t(s)$, where T and S are arbitrary subsets of variables, representing multiple interventions and multiple outcomes, such as those encountered in the management of time varying treatments. The criterion established in this section extends those of [PR95] and [KM99], and also provides a criterion for the identification of *direct* effects, that is, the effect of one variable on another when all other variables are held fixed (Section 5.4.4). Section 5.5 deals with the identification of general conditional causal effects $P_t(s|c)$. Finally, in Section 5.6, we show that causal effects in a Markovian model with arbitrary sets of unobserved variables can be identified by first converting the model into a semi-Markovian model while keeping the identifiability properties.

¹[GP95] claimed their graphical criterion to embrace all cases where identification is verifiable by *do*-calculus. We show in this chapter (Section 5.3.7) that their criterion is *not* complete in this sense.

5.2 Identification of $P_x(v)$

Let X be a singleton variable. In this section we study the problem of identifying the causal effects of X on $V \setminus \{X\}$, (namely, on all other variables in V), a quantity denoted by $P_x(v)$.

5.2.1 The easiest case

Theorem 13 *If there is no bidirected edge connected to X , then $P_x(v)$ is identifiable and is given by*

$$P_x(v) = P(v|x, pa_x)P(pa_x) \quad (5.2)$$

Proof: Since there is no bidirected edge connected to X , we have that the term $P(x|pa_x, u^x) = P(x|pa_x)$ in Eq. (1.5) can be moved ahead of the summation, giving

$$\begin{aligned} P(v) &= P(x|pa_x) \sum_u \prod_{\{i|V_i \neq X\}} P(v_i|pa_i, u^i)P(u) \\ &= P(x|pa_x)P_x(v). \end{aligned} \quad (5.3)$$

Hence,

$$P_x(v) = P(v)/P(x|pa_x) = P(v|x, pa_x)P(pa_x) \quad (5.4)$$

□

Theorem 13 also follows from Theorem 3.2.5 of [Pea00] which states that for any disjoint sets S and T in a Markovian model M , if the parents of T are measured, then $P_t(s)$ is identifiable. Indeed, when the parents of X are measured, there would be no bidirected edge entering X in the semi-Markovian representation of M and the identification of $P_x(v)$ is insured.

5.2.2 A more interesting case

The case where there is no bidirected edge connected to any child of X is also easy to handle. As an example, consider the graph given in Figure 1.2. We have

$$P(v) = P(z|x) \sum_u P(x|u)P(y|z, u)P(u), \quad (5.5)$$

$$P_x(v) = P(z|x) \sum_u P(y|z, u)P(u). \quad (5.6)$$

From Eq. (5.5), we have

$$\sum_u P(x|u)P(y|z, u)P(u) = P(v)/P(z|x), \quad (5.7)$$

hence,

$$\sum_u P(y|z, u)P(u) = \sum_x \sum_u P(x|u)P(y|z, u)P(u) = \sum_x P(v)/P(z|x). \quad (5.8)$$

Substituting Eq. (5.8) into Eq. (5.6), we obtain

$$P_x(y, z) = P(z|x) \sum_{x'} P(x', y, z)/P(z|x') = P(z|x) \sum_{x'} P(y|x', z)P(x'). \quad (5.9)$$

This derivation can be generalized to the case where X has several children. Letting Ch_x denote the set of X 's children, we have the following theorem.

Theorem 14 *If there is no bidirected edge connected to any child of X , then $P_x(v)$ is identifiable and is given by*

$$P_x(v) = \left(\prod_{\{i|V_i \in Ch_x\}} P(v_i|pa_i) \right) \sum_x \frac{P(v)}{\prod_{\{i|V_i \in Ch_x\}} P(v_i|pa_i)} \quad (5.10)$$

Proof: Let $S = V \setminus (Ch_x \cup \{X\})$. Since there is no bidirected edge connected to any child of X , the factors corresponding to the variables in Ch_x can be moved ahead of the summation in Eqs. (1.5) and (1.6), and we have

$$P(v) = \left(\prod_{\{i|V_i \in Ch_x\}} P(v_i|pa_i) \right) \sum_u P(x|pa_x, u^x) \prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)P(u), \quad (5.11)$$

and

$$P_x(v) = \left(\prod_{\{i|V_i \in Ch_x\}} P(v_i|pa_i) \right) \sum_u \prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)P(u). \quad (5.12)$$

The variable X does not appear in the factors of $\prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)$, hence we augment $\prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)$ with the term $\sum_x P(x|pa_x, u^x) = 1$, and write

$$\begin{aligned} \sum_u \prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)P(u) &= \sum_x \sum_u P(x|pa_x, u^x) \prod_{\{i|V_i \in S\}} P(v_i|pa_i, u^i)P(u) \\ &= \sum_x \frac{P(v)}{\prod_{\{i|V_i \in Ch_x\}} P(v_i|pa_i)}. \quad (\text{by (5.11)}) \end{aligned} \quad (5.13)$$

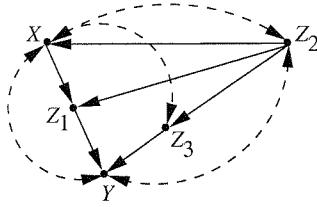


Figure 5.1: Theorem 14 is applicable for identifying $P_x(z_1, z_2, z_3, y)$.

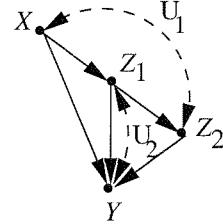


Figure 5.2: The problem of identifying $P_x(z_1, z_2, y)$.

Substituting this expression into Eq. (5.12) leads to Eq. (5.10). \square

The usefulness of Theorem 14 can be demonstrated in the model of Figure 5.1. Although the diagram is quite complicated, Theorem 14 is applicable, and readily gives

$$\begin{aligned} P_x(z_1, z_2, z_3, y) &= P(z_1|x, z_2) \sum_{x'} \frac{P(x', z_1, z_2, z_3, y)}{P(z_1|x', z_2)} \\ &= P(z_1|x, z_2) \sum_{x'} P(y, z_3|x', z_1, z_2) P(x', z_2). \end{aligned} \quad (5.14)$$

Note that this expression remains valid when we add bidirected edges between Z_3 and Y and between Z_3 and Z_2 .

5.2.3 The general case

When there are bidirected edges connected to the children of X , it may still be possible to identify $P_x(v)$. To illustrate, consider the graph in Figure 5.2, for

which we have

$$P(v) = \sum_{u_1} P(x|u_1)P(z_2|z_1, u_1)P(u_1) \sum_{u_2} P(z_1|x, u_2)P(y|x, z_1, z_2, u_2)P(u_2), \quad (5.15)$$

and

$$P_x(v) = \sum_{u_1} P(z_2|z_1, u_1)P(u_1) \sum_{u_2} P(z_1|x, u_2)P(y|x, z_1, z_2, u_2)P(u_2). \quad (5.16)$$

Let

$$Q_1 = \sum_{u_1} P(x|u_1)P(z_2|z_1, u_1)P(u_1), \quad (5.17)$$

$$(5.18)$$

and

$$Q_2 = \sum_{u_2} P(z_1|x, u_2)P(y|x, z_1, z_2, u_2)P(u_2). \quad (5.19)$$

Eq. (5.15) can then be written as

$$P(v) = Q_1 \cdot Q_2, \quad (5.20)$$

and Eq. (5.16) as

$$P_x(v) = Q_2 \sum_x Q_1. \quad (5.21)$$

Thus, if Q_1 and Q_2 can be computed from $P(v)$, then $P_x(v)$ is identifiable and given by Eq. (5.21). In fact, it is enough to show that Q_1 can be computed from $P(v)$ (i.e., identifiable); Q_2 would then be given by $P(v)/Q_1$. To show that Q_1 can indeed be obtained from $P(v)$, we sum both sides of Eq. (5.15) over y , and get

$$P(x, z_1, z_2) = Q_1 \cdot \sum_{u_2} P(z_1|x, u_2)P(u_2). \quad (5.22)$$

Summing both sides of (5.22) over z_2 , we get

$$P(x, z_1) = P(x) \sum_{u_2} P(z_1|x, u_2)P(u_2), \quad (5.23)$$

hence,

$$\sum_{u_2} P(z_1|x, u_2)P(u_2) = P(z_1|x). \quad (5.24)$$

From Eqs. (5.24) and (5.22),

$$Q_1 = P(x, z_1, z_2)/P(z_1|x) = P(z_2|x, z_1)P(x), \quad (5.25)$$

and from Eq. (5.20),

$$Q_2 = P(v)/Q_1 = P(y|x, z_1, z_2)P(z_1|x). \quad (5.26)$$

Finally, from Eq. (5.21), we obtain

$$P_x(v) = P(y|x, z_1, z_2)P(z_1|x) \sum_{x'} P(z_2|x', z_1)P(x'). \quad (5.27)$$

From the preceding example, we see that because the two bidirected arcs in Figure 5.2 do not share a common node, the set of factors (of $P(v)$) containing U_1 is disjoint of those containing U_2 , and $P(v)$ can be decomposed into a product of two terms, each being a summation of products. This decomposition has been studied in Chapter 4, in which we introduced the ideas of “c-component” and “c-factor”.

5.2.3.1 C-component

Two variables are in the same c-component if and only if they are connected by a bidirected path, a path composed entirely of bidirected edges. We will use the $Q[.]$ notation defined in Chapter 4. For any set $C \subseteq V$, $Q[C](v)$ denotes the following function (see Eq. (4.41))

$$Q[C](v) = P_{v \setminus C}(c) = \sum_u \prod_{\{i|V_i \in C\}} P(v_i|pa_i, u^i)P(u). \quad (5.28)$$

For any set C , let G_C denote the subgraph of G composed only of variables in C . We rewrite Corollary 1 as a lemma in the following, tailored for semi-Markovian models.

Lemma 7 *Assuming that V is partitioned into c-components S_1, \dots, S_k , we have*

$$(i) P(v) = \prod_i Q[S_i].$$

(ii) *Let a topological order over V be $V_1 < \dots < V_n$, and let $V^{(i)} = \{V_1, \dots, V_i\}$, $i = 1, \dots, n$, and $V^{(0)} = \emptyset$. Then each c-factor $Q[S_j]$, $j = 1, \dots, k$, is identifiable and is given by*

$$Q[S_j] = \prod_{\{i|V_i \in S_j\}} P(v_i|v^{(i-1)}). \quad (5.29)$$

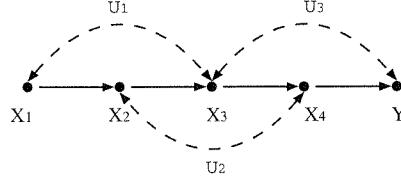


Figure 5.3: An example for applying Lemma 7.

(ii) Each factor $P(v_i|v^{(i-1)})$ can be expressed as

$$P(v_i|v^{(i-1)}) = P(v_i|pa(T_i) \setminus \{v_i\}), \quad (5.30)$$

where T_i is the c-component of $G_{V^{(i)}}$ that contains V_i .

We show the use of Lemma 7 by an example shown in Figure 5.3, which has two c-components $S_1 = \{X_2, X_4\}$ and $S_2 = \{X_1, X_3, Y\}$. $P(v)$ decomposes into

$$P(x_1, x_2, x_3, x_4, y) = Q[S_1]Q[S_2], \quad (5.31)$$

where

$$Q[S_1] = \sum_{u_2} P(x_2|x_1, u_2)P(x_4|x_3, u_2)P(u_2), \quad (5.32)$$

$$Q[S_2] = \sum_{u_1, u_3} P(x_1|u_1)P(x_3|x_2, u_1, u_3)P(y|x_4, u_3)P(u_1)P(u_3). \quad (5.33)$$

By Lemma 7, both $Q[S_1]$ and $Q[S_2]$ are identifiable. The only admissible order of variables is $X_1 < X_2 < X_3 < X_4 < Y$, and Eq. (5.29) gives

$$Q[S_1] = P(x_4|x_1, x_2, x_3)P(x_2|x_1), \quad (5.34)$$

$$Q[S_2] = P(y|x_1, x_2, x_3, x_4)P(x_3|x_1, x_2)P(x_1). \quad (5.35)$$

We can also check that the expressions obtained in Eq.s (5.25) and (5.26) for Figure 5.2 satisfy Lemma 7.

5.2.3.2 An identification criterion for $P_x(v)$

Lemma 7 has important implications on the general identifiability problem, and in this section we show how to use this property to identify $P_x(v)$.

Let X belong to the c-component S^X , and let other c-components be S_1, \dots, S_k . We have

$$P(v) = Q[S^X] \prod_i Q[S_i], \quad (5.36)$$

and

$$P_x(v) = Q[S^X \setminus \{X\}] \prod_i Q[S_i]. \quad (5.37)$$

Since all $Q[S_i]$'s are identifiable by Lemma 7, $P_x(v)$ is identifiable if and only if $Q[S^X \setminus \{X\}]$ is identifiable, and we have the following theorem.

Theorem 15 *If there is no bidirected path connecting X to any of its children, then $P_x(v)$ is identifiable and is given by*

$$P_x(v) = \frac{P(v)}{Q[S^X]} \sum_x Q[S^X], \quad (5.38)$$

where S^X is the c-component that contains X .

Proof: If there is no bidirected path connecting X to any of its children, then none of X 's children is in S^X . Under this condition, removing the term $P(x|pa_x, u^x)$ from $Q[S^X]$ is equivalent to summing $Q[S^X]$ over X , and we can write

$$Q[S^X \setminus \{X\}] = \sum_x Q[S^X]. \quad (5.39)$$

Hence from Eq.s (5.37) and (5.36), we obtain

$$P_x(v) = (\sum_x Q[S^X]) \prod_i Q_i = (\sum_x Q[S^X]) \frac{P(v)}{Q[S^X]}, \quad (5.40)$$

which proves the identifiability of $P_x(v)$. \square

We demonstrate the use of Theorem 15 by identifying $P_{x_1}(x_2, x_3, x_4, y)$ in Figure 5.3. The graph has two c-components $S_1 = \{X_2, X_4\}$ and $S_2 = \{X_1, X_3, Y\}$, with corresponding c-factors given in (5.34) and (5.35). Since X_1 is in S_2 and its child X_2 is not in S_2 , Theorem 15 ensures that $P_{x_1}(x_2, x_3, x_4, y)$ is identifiable and is given by

$$\begin{aligned} P_{x_1}(x_2, x_3, x_4, y) &= Q[S_1] \sum_{x_1} Q[S_2] \\ &= P(x_4|x_1, x_2, x_3) P(x_2|x_1) \sum_{x'_1} P(y|x'_1, x_2, x_3, x_4) P(x_3|x'_1, x_2) P(x'_1). \end{aligned} \quad (5.41)$$

More examples where Theorem 15 is applicable can be found in Figure 3.8 of [Pea00], some of which required complicated do-calculus derivations.

5.2.3.3 Necessity of the criterion

Next we will show that the condition given in Theorem 15 is also necessary for the identifiability of $P_x(v)$. To facilitate the proof of necessity, first we prove the following lemma.

Lemma 8 *Let $S, T \subseteq V$ be two disjoint sets of variables. If $P_t(s)$ is not identifiable in G , then $P_t(s)$ is not identifiable in the graph resulted from adding a directed or bidirected edge to G . Equivalently, if $P_t(s)$ is identifiable in G , then $P_t(s)$ is still identifiable in the graph resulted from removing a directed or bidirected edge from G .*

Proof: If $P_t(s)$ is not identifiable in G , then there exist two models with the same causal graph G , M_1 and M_2 , such that

$$P^{M_1}(v) = P^{M_2}(v) > 0, \text{ and } P_t^{M_1}(s) \neq P_t^{M_2}(s), \quad (5.42)$$

where

$$P^{M_k}(v) = \sum_u \prod_i P^{M_k}(v_i | pa_i, u^i) P^{M_k}(u), \quad k = 1, 2. \quad (5.43)$$

For a graph G' with extra edges added to G , we can always construct new models in such a way that the added edges are ineffective.

(i) Let G' be the graph identical to G except with an extra edge $Y \rightarrow V_j$. $P(v)$ decomposes as

$$P(v) = \sum_u P(v_j | pa_j, y, u^j) \prod_{i \neq j} P(v_i | pa_i, u^i) P(u). \quad (5.44)$$

We construct two models M'_1 and M'_2 with the causal graph G' as

$$P^{M'_k}(v_i | pa_i, u^i) = P^{M_k}(v_i | pa_i, u^i), \quad i \neq j, \quad k = 1, 2, \quad (5.45)$$

$$P^{M'_k}(v_j | pa_j, y, u^j) = P^{M_k}(v_j | pa_j, u^j), \quad k = 1, 2, \quad (5.46)$$

$$P^{M'_k}(u) = P^{M_k}(u), \quad k = 1, 2. \quad (5.47)$$

Clearly, if the pair (M_1, M_2) satisfies (5.42), so would the pair (M'_1, M'_2) . Hence $P_t(s)$ is not identifiable in G' .

(ii) Let G' be the graph identical to G except with an extra edge $V_l \longleftrightarrow V_j$. $P(v)$ decomposes as

$$P(v) = \sum_{u'} P(u') \sum_u P(v_j | pa_j, u^j, u') P(v_l | pa_l, u^l, u') \prod_{i \neq j, i \neq l} P(v_i | pa_i, u^i) P(u), \quad (5.48)$$

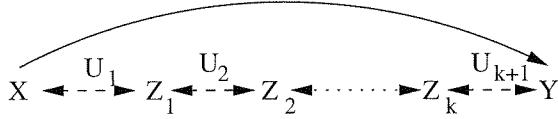


Figure 5.4: A graph used in proving Theorem 16.

where U' represents the new unobserved variable. We construct two models M'_1 and M'_2 with the causal graph G' as

$$P^{M'_k}(v_i|pa_i, u^i) = P^{M_k}(v_i|pa_i, u^i), \quad i \neq j, \quad i \neq l, \quad k = 1, 2, \quad (5.49)$$

$$P^{M'_k}(v_i|pa_i, u^i, u') = P^{M_k}(v_i|pa_i, u^i), \quad i = j, l, \quad k = 1, 2, \quad (5.50)$$

$$P^{M'_k}(u) = P^{M_k}(u), \quad k = 1, 2. \quad (5.51)$$

Again, if the pair (M_1, M_2) satisfies (5.42), so would the pair (M'_1, M'_2) . Hence $P_t(s)$ is not identifiable in G' . \square

Next we prove that the condition given in Theorem 15 is necessary.

Theorem 16 *If there is a bidirected path connecting X to any of its children in G , then $P_x(v)$ is not identifiable.*

Proof: Let Y be a child of X and assume that there is a bidirected path connecting X and Y with variables Z_1, \dots, Z_k on the path (see Figure 5.4). We will prove that, for any $k \geq 1$, $P_x(y, z_1, \dots, z_k)$ is not identifiable in the graph shown in Figure 5.4, which is a subgraph of G . By Lemma 8, if $P_x(y, z_1, \dots, z_k)$ is not identifiable in a subgraph of G , then it is not identifiable in G , and therefore $P_x(v)$ is not identifiable in G .

Let $U = \{U_1, \dots, U_{k+1}\}$. In Figure 5.4, we have

$$\begin{aligned} & P(x, y, z_1, \dots, z_k) \\ &= \sum_u P(x|u_1)P(y|x, u_{k+1})P(z_1|u_1, u_2)\cdots P(z_k|u_k, u_{k+1})P(u_1)\cdots P(u_{k+1}), \end{aligned} \quad (5.52)$$

and

$$\begin{aligned} & P_x(y, z_1, \dots, z_k) \\ &= \sum_u P(y|x, u_{k+1})P(z_1|u_1, u_2)\cdots P(z_k|u_k, u_{k+1})P(u_1)\cdots P(u_{k+1}). \end{aligned} \quad (5.53)$$

Let all variables $X, Y, Z_1, \dots, Z_k, U_1, \dots, U_{k+1}$ be binary variables. We will prove the nonidentifiability of $P_x(y, z_1, \dots, z_k)$ by constructing two models such that in both models,

$$P(x, y, z_1, \dots, z_k) = (1/2)^{k+2}, \quad \text{for all possible values of } x, y, z_1, \dots, z_k, \quad (5.54)$$

while $P_x(y, z_1, \dots, z_k)$ has different values in the two models. The construction involves the specification of all conditional probabilities in a parametric form, and shows two different parameterization both satisfying the set of 2^{k+2} equations in (5.54). We use the following parameterization, with five parameters, a, b, c, d , and e .

$$P(u_i) = 1/2, \quad u_i = 0, 1, \text{ and } i = 1, \dots, k + 1 \quad (5.55)$$

$$\begin{array}{cc|c} x & u_1 & P(x|u_1) \\ \hline 0 & 0 & 1/2 + a \\ 0 & 1 & 1/2 - a \end{array} \quad (5.56)$$

$$\begin{array}{ccc|c} y & x & u_{k+1} & P(y|x, u_{k+1}) \\ \hline 0 & 0 & 0 & 1/2 + b \\ 0 & 0 & 1 & 1/2 - b \\ 0 & 1 & 0 & 1/2 \\ 0 & 1 & 1 & 1/2 \end{array} \quad (5.57)$$

$$\begin{array}{ccc|c} z_1 & u_1 & u_2 & P(z_1|u_1, u_2) \\ \hline 0 & 0 & 0 & 1/2 + c \\ 0 & 0 & 1 & 1/2 - c \\ 0 & 1 & 0 & 1/2 + d \\ 0 & 1 & 1 & 1/2 - d \end{array} \quad (5.58)$$

$$\begin{array}{ccc|c} z_i & u_i & u_{i+1} & P(z_i|u_i, u_{i+1}) \\ \hline 0 & 0 & 0 & 1/2 + e \\ 0 & 0 & 1 & 1/2 - e \\ 0 & 1 & 0 & 1/2 - e \\ 0 & 1 & 1 & 1/2 + e \end{array} \quad i = 2, \dots, k. \quad (5.59)$$

Substituting (5.55) in (5.52), Eq. (5.54) becomes

$$\frac{1}{2} = \sum_u P(x|u_1)P(y|x, u_{k+1})P(z_1|u_1, u_2)\cdots P(z_k|u_k, u_{k+1}). \quad (5.60)$$

Next, we prove that if Eq. (5.60) is satisfied for $x = 0, y = 0, z_1 = 0, \dots, z_k = 0$, then it is satisfied for all possible values of x, y, z_1, \dots, z_k . We have that for any a, b, c, d, e , the parameterization given in Eqs. (5.55)–(5.59) satisfies the following properties

$$\sum_{u_1} P(x|u_1) = 1. \quad (5.61)$$

$$\sum_{u_{k+1}} P(y|x, u_{k+1}) = 1. \quad (5.62)$$

$$\sum_{u_{i+1}} P(z_i|u_i, u_{i+1}) = 1, i = 1, \dots, k. \quad (5.63)$$

$$\sum_{u_i} P(z_i|u_i, u_{i+1}) = 1, i = 2, \dots, k. \quad (5.64)$$

(a) For $x = 1$ and any values of y, z_1, \dots, z_k , Eq. (5.60) is satisfied:

$$\begin{aligned} & \sum_u P(x = 1|u_1)P(y|x = 1, u_{k+1})P(z_1|u_1, u_2)\cdots P(z_k|u_k, u_{k+1}) \\ &= \frac{1}{2} \sum_u P(x = 1|u_1)P(z_1|u_1, u_2)\cdots P(z_k|u_k, u_{k+1}) \quad (\text{by } P(y|x = 1, u_{k+1}) = 1/2) \\ &= \frac{1}{2} \quad (\text{by Eqs. (5.63) and (5.61)}) \end{aligned} \quad (5.65)$$

(b) If for a particular set of values x, y, z_1, \dots, z_k , Eq. (5.60) is satisfied, then for the set of values $x, 1 - y, z_1, \dots, z_k$, Eq. (5.60) is also satisfied:

$$\begin{aligned} & \sum_u P(x|u_1)P(1 - y|x, u_{k+1})P(z_1|u_1, u_2)P(z_2|u_2, u_3)\cdots P(z_k|u_k, u_{k+1}) \\ &= \sum_u P(x|u_1)(1 - P(y|x, u_{k+1}))P(z_1|u_1, u_2)P(z_2|u_2, u_3)\cdots P(z_k|u_k, u_{k+1}) \\ &= \sum_u P(x|u_1)P(z_1|u_1, u_2)P(z_2|u_2, u_3)\cdots P(z_k|u_k, u_{k+1}) - \frac{1}{2} \quad (\text{by Eq. (5.60)}) \\ &= 1 - \frac{1}{2} \quad (\text{by Eqs. (5.63) and (5.61)}) \\ &= \frac{1}{2} \end{aligned} \quad (5.66)$$

(c) If for a particular set of values x, y, z_1, \dots, z_k , Eq. (5.60) is satisfied, then for the set of values $x, y, z_1, \dots, z_{i-1}, 1 - z_i, z_{i+1}, \dots, z_k$, for $i = 1, \dots, k$, Eq. (5.60) is satisfied as well:

$$\begin{aligned}
& \sum_u P(x|u_1)P(y|x, u_{k+1})P(z_1|u_1, u_2) \cdots P(1 - z_i|u_i, u_{i+1}) \cdots P(z_k|u_k, u_{k+1}) \\
&= \sum_u P(x|u_1)P(y|x, u_{k+1})P(z_1|u_1, u_2) \cdots P(z_{i-1}|u_{i-1}, u_i)P(z_{i+1}|u_{i+1}, u_{i+2}) \\
&\quad \cdots P(z_k|u_k, u_{k+1}) - \frac{1}{2} \quad (\text{by Eq. (5.60)}) \\
&= 1 - \frac{1}{2} \quad (\text{by Eqs. (5.61)–(5.64)}) \\
&= \frac{1}{2}
\end{aligned} \tag{5.67}$$

From (a), (b), and (c), we obtain that if Eq. (5.60) is satisfied for $x = 0, y = 0, z_1 = 0, \dots, z_k = 0$, then it is satisfied for all possible values of x, y, z_1, \dots, z_k .

Next, we substitute the conditional probabilities given in Eqs. (5.55)–(5.59) into Eq. (5.60) for $x = 0, y = 0, z_1 = 0, \dots, z_k = 0$. Define

$$f_{u_2, u_{k+1}} = \sum_{u_3, \dots, u_k} P(z_2 = 0|u_2, u_3) \cdots P(z_k = 0|u_k, u_{k+1}) \tag{5.68}$$

We obtain

$$\begin{aligned}
f_{00} &= (1/2 + e)^{k-1} + \binom{k-1}{2} (1/2 + e)^{k-3} (1/2 - e)^2 \\
&\quad + \binom{k-1}{4} (1/2 + e)^{k-5} (1/2 - e)^4 + \cdots \\
&= \sum_{i=0}^{i < k/2} \binom{k-1}{2i} (1/2 + e)^{k-1-2i} (1/2 - e)^{2i}.
\end{aligned} \tag{5.69}$$

From Eq. (5.64), we have

$$\sum_{u_2} f_{u_2, u_{k+1}} = 1. \tag{5.70}$$

From Eq. (5.63), we have

$$\sum_{u_{k+1}} f_{u_2, u_{k+1}} = 1. \tag{5.71}$$

Let $f = f_{00} - 1/2$, then $f_{u_2, u_{k+1}}$ is given as

u_2	u_{k+1}	$f_{u_2, u_{k+1}}$
0	0	$1/2 + f$
0	1	$1/2 - f$
1	0	$1/2 - f$
1	1	$1/2 + f$

Therefore, for $x = 0, y = 0, z_1 = 0, \dots, z_k = 0$, Eq. (5.60) becomes

$$\begin{aligned}
\frac{1}{2} &= \sum_{u_1, u_{k+1}, u_2} P(x = 0|u_1)P(y = 0|x = 0, u_{k+1})P(z_1 = 0|u_1, u_2)f_{u_2, u_{k+1}} \\
&= (1/2 + a)(1/2 + b)[(1/2 + c)(1/2 + f) + (1/2 - c)(1/2 - f)] \\
&\quad + (1/2 + a)(1/2 - b)[(1/2 + c)(1/2 - f) + (1/2 - c)(1/2 + f)] \\
&\quad + (1/2 - a)(1/2 + b)[(1/2 + d)(1/2 + f) + (1/2 - d)(1/2 - f)] \\
&\quad + (1/2 - a)(1/2 - b)[(1/2 + d)(1/2 - f) + (1/2 - d)(1/2 + f)] \\
&= 1/2 + 2bf(c + d + 2ac - 2ad)
\end{aligned} \tag{5.72}$$

which leads to

$$bf(c + d + 2ac - 2ad) = 0. \tag{5.73}$$

$P_x(y, z_1, \dots, z_k)$ is computed as

$$\begin{aligned}
P_{x=0}(y = 0, z_1 = 0, \dots, z_k = 0) &= \frac{1}{2^{k+1}} \sum_{u_1, u_{k+1}, u_2} P(y = 0|x = 0, u_{k+1})P(z_1 = 0|u_1, u_2)f_{u_2, u_{k+1}} \\
&= \frac{1}{2^{k+1}}[1 + 4bf(c + d)]
\end{aligned} \tag{5.74}$$

Let $-1/2 < e_0 < 1/2$ be a number such that $f \neq 0$. Consider the following two models:

Model 1 $a = 1/4, b = 0, c = d = 1/4, e = e_0$.

Model 2 $a = 1/4, b = 1/4, c = 1/12, d = -1/4, e = e_0$.

Eq. (5.73) holds in both models, hence the two models have the same distribution $P(x, y, z_1, \dots, z_k) = (1/2)^{k+2}$. In Model 1, $P_{x=0}(y = 0, z_1 = 0, \dots, z_k = 0) = (1/2)^{k+1}$. In Model 2, $P_{x=0}(y = 0, z_1 = 0, \dots, z_k = 0) = (1/2)^{k+1}(1 - f/6)$. Since $f \neq 0$, we have that $P_{x=0}(y = 0, z_1 = 0, \dots, z_k = 0)$ takes different values in Model 1 and 2. \square

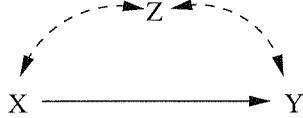


Figure 5.5: $P_x(y, z)$ is not identifiable but $P_x(y)$ is.

5.3 Identification of $P_x(s)$

Let X be a singleton variable and $S \subseteq V$ be a set of variables. In this section, we study the problem of identifying $P_x(s)$. Clearly, whenever $P_x(v)$ is identifiable, so is $P_x(s)$. However, there are obvious cases where $P_x(v)$ is not identifiable and still $P_x(s)$ is identifiable for some subsets S of V . The simplest such example can be seen in Figure 5.5, which is a special case of Figure 5.4 with $k = 1$. Here, variable Z can be ignored in the computation of $P_x(y)$, giving $P_x(y) = P(y|x)$ and $P_x(z) = P(z)$, while (by Theorem 16) $P_x(y, z)$ is not identifiable. This example suggests that a criterion similar to that of Theorem 15, applicable in some subgraphs of G , would establish the identifiability of $P_x(s)$. We will show indeed that $P_x(s)$ is identified when a systematic removal of certain nonessential nodes from G will lead to an identification criterion based on Theorem 15. First we give a criterion for identifying $P_x(s)$ which is a simple extension of Theorem 15.

5.3.1 A criterion for identifying $P_x(s)$

For any set $C \subseteq V$, let $An(C)$ denote the union of C and the set of ancestors of the variables in C . The nonancestors of S are nonessential for identifying $P_x(s)$ and we have the following lemma.

Lemma 9 $P_x(s)$ is identifiable if and only if in the subgraph $G_{An(S)}$, $P_x(s)$ is identifiable.

Proof: (only if) By Lemma 8.

(if) Summing both sides of Eq. (1.5) over $V \setminus An(S)$, we have that the marginal distribution $P(an(S))$ decomposes exactly according to the graph $G_{An(S)}$. Hence if $P_x(s)$ is identifiable in $G_{An(S)}$, then it is computable from $P(an(S))$, and therefore is identifiable in G . \square

From Lemma 9, a direct extension of Theorem 15 leads to the following criterion.

Theorem 17 $P_x(s)$ is identifiable if there is no bidirected path connecting X to any of its children in $G_{An(S)}$.

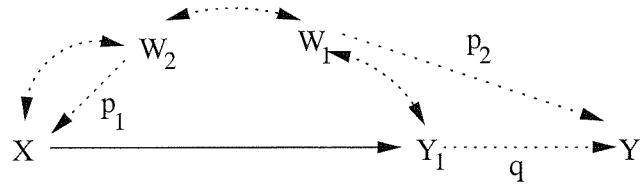


Figure 5.6: A graph used in proving Proposition 1.

When the condition in Theorem 17 is satisfied, we can compute $P_x(an(S))$ by applying Theorem 17 in $G_{An(S)}$, and $P_x(s)$ can be obtained by marginalizing over $P_x(an(S))$.

This simple criterion can classify correctly all the examples treated in the literature with X singleton, including those contrived by [GP95]. In fact, for X and S being singletons, we will show that if there is a bidirected path connecting X to one of its children such that every node on the path is in $An(S)$, then none of the “back-door”, “front-door”, and [GP95] criteria is applicable. The criterion in [GP95] (which will be called the G-P criterion) is for identifying $P_x(y)$ with X and Y being singletons, and it includes the “front-door” and “back-door” criteria as special cases (see [Pea00, pp. 114-8]).

Proposition 1 *If there is a bidirected path connecting X to one of its children such that every node on the path is an ancestor of Y , then the G-P criterion is not applicable.*

Proof: There are four conditions in the G-P criterion, among which Condition 1 is a special case of Condition 3, and Condition 2 is trivial. Therefore we only need to consider Condition 3 and 4.

Assume that there is a bidirected path p from X to its child Y_1 such that every node on p is an ancestor of Y , and that there is a directed path q from Y_1 to Y . We will show by contradiction that neither Condition 3 nor Condition 4 is applicable for identifying $P_x(y)$. For any set Z , a node will be called Z -active if it is in Z or any of its descendants is in Z , otherwise it will be called Z -inactive.

(Condition 3) Assume that there exists a set Z that blocks all back-door paths from X to Y so that $P_x(z)$ is identifiable.² If every internal node on p is an ancestor of X , or if every nonancestor of X on p is Z -active, then let $W_1 = Y_1$, otherwise let W_1 be the Z -inactive non-ancestor of X that is closest to X on p (see Figure 5.6). If every internal node on the subpath $p(W_1, X)$ ³ is Z -active,

²A path from X to Y is said to be a *back-door* path if it contains an arrow into X .

³We use $p(W_1, X)$ to represent the subpath of p from W_1 to X .

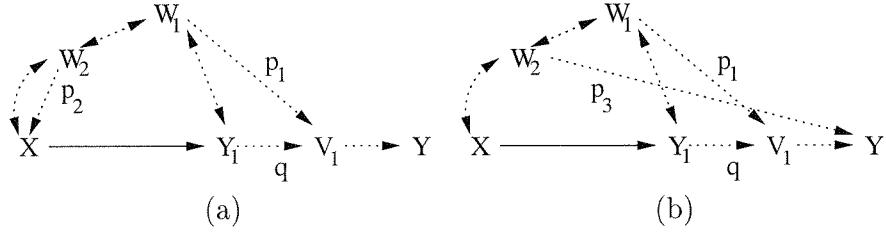


Figure 5.7: Graphs used in proving Proposition 1.

then let $W_2 = X$, otherwise let W_2 be the Z -inactive node that is closest to W_1 on $p(W_1, X)$. From the definition of W_1 and W_2 , W_2 must be an ancestor of X (or be X itself), and let p_1 be any directed path from W_2 to X . (i) If $W_1 \neq Y_1$, letting p_2 be any directed path from W_1 to Y , then from the definition of W_1 and W_2 the path $p' = (p_1(X, W_2), p(W_2, W_1), p_2(W_1, Y))$ is a back-door path from X to Y that is not blocked by Z (see Figure 5.6) since W_2 is Z -inactive, all internal nodes on $p(W_2, W_1)$ is Z -active, and W_1 is Z -inactive. (ii) If $W_1 = Y_1$, there are two situations:

(a) Z consists entirely of nondescendants of X . Then the path $p'' = (p_1(X, W_2), p(W_2, Y_1), q(Y_1, Y))$ is a back-door path from X to Y that is not blocked by Z .

(b) Z contains a variable Y' on $q(Y_1, Y)$ so that $P_x(z)$ is identifiable. By the definition of W_1 , every node on p is an ancestor of Z . $P_x(z)$ can not be identified by Theorem 17, and the G-P criterion is not applicable for identifying $P_x(z)$ if Z contains more than one variable. If Z contains only one variable Y' , then every node on p is an ancestor of Y' . If $P_x(y')$ is identifiable by Condition 3 of the G-P criterion (Condition 4 is not applicable as proved later), then from the preceding analysis there is a Y'' on the path $q(Y_1, Y')$ such that every node on p is an ancestor of Y'' and $P_x(y'')$ is identifiable. By induction, in the end we have every node on p is an ancestor of Y_1 and $P_x(y_1)$ is identifiable, which does not hold from the preceding analysis.

(Condition 4) Assume that there exist sets Z_1 and Z_2 that satisfy all (i)–(iv) conditions in Condition 4. Since Z_1 has to block the path $((X, Y_1), q(Y_1, Y))$, let V_1 be the variable in Z_1 that is closest to Y_1 on the path q (see Figure 5.7(a)). If none of the internal node on p is in $An(V_1) \setminus An(X)$ (the set of ancestors of V_1 that are not ancestors of X) or if every variable in $An(V_1) \setminus An(X)$ on p is Z_2 -active, then let $W_1 = Y_1$, otherwise let W_1 be the Z_2 -inactive variable in $An(V_1) \setminus An(X)$ that is closest to X on p . Let p_1 be any directed path from W_1 to V_1 . If every internal node on the subpath $p(W_1, X)$ is Z_2 -active, then let $W_2 = X$, otherwise let W_2 be the Z_2 -inactive node that is closest to W_1 on $p(W_1, X)$. Since W_2 must

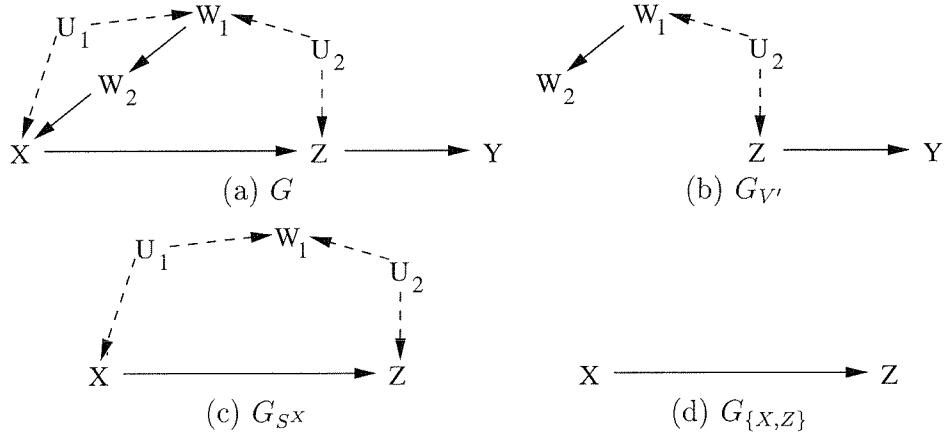


Figure 5.8: Subgraphs of G used in computing $P_x(y)$.

be an ancestor of Y , from the definition of W_1 and W_2 , there are two possible situations:

(a) W_2 is an ancestor of X or $W_2 = X$. Let p_2 be any directed path from W_2 to X (see Figure 5.7(a)). From the definition of W_1 and W_2 , the path $p' = (p_2(X, W_2), p(W_2, W_1), p_1(W_1, V_1))$ is a back-door path from X to $V_1 \in Z_1$ that is not blocked by Z_2 that does not contain any descendant of X (see Figure 5.7(a)).

(b) W_2 is an ancestor of Y but not ancestor of V_1 ($W_2 \in An(Y) \setminus An(V_1)$). Let p_3 be any directed path from W_2 to Y (see Figure 5.7(b)). From the definition of W_1 and W_2 , the path $p'' = (p_1(V_1, W_1), p(W_1, W_2), p_3(W_2, Y))$ is a back-door path from $V_1 \in Z_1$ to Y that is not blocked by Z_2 (see Figure 5.7(b)). \square

However, the criterion in Theorem 17 is not necessary for identifying $P_x(s)$. In the next section, we give an example in which $P_x(s)$ is identifiable but Theorem 17 is not applicable, and the process of computing $P_x(s)$ will give us hints on how to improve the criterion.

5.3.2 An example

To illustrate the general process of computing $P_x(s)$ making use of the factorization of $P(v)$ into c-factors, we work out an example in this section. Consider the problem of identifying $P_x(y)$ in Figure 5.8(a). Theorem 17 is not applicable, but we will show that $P_x(y)$ is identifiable. Let $V = \{X, Z, Y, W_1, W_2\}$ and $V' = \{Z, Y, W_1, W_2\}$. V is partitioned into three c-components: $S^X = \{X, Z, W_1\}$,

$\{W_2\}$, and $\{Y\}$. $P(v)$ can be decomposed into

$$P(v) = P(w_2|w_1)P(y|z)Q[S^X], \quad (5.75)$$

where

$$Q[S^X] = \sum_{u_1, u_2} P(x|w_2, u_1)P(w_1|u_1, u_2)P(z|x, u_2)P(u_1)P(u_2) \quad (5.76)$$

$$= P(v)/(P(w_2|w_1)P(y|z)) = P(z, x|w_2, w_1)P(w_1). \quad (5.77)$$

$P_x(v')$ is decomposed into

$$P_x(v') = Q[V'] = P(w_2|w_1)P(y|z) \sum_{u_1, u_2} P(w_1|u_1, u_2)P(z|x, u_2)P(u_1)P(u_2). \quad (5.78)$$

We want to compute $P_x(y)$:

$$\begin{aligned} P_x(y) &= \sum_{z, w_1, w_2} P_x(v') \\ &= \sum_{z, w_1, w_2} Q[V'] \\ &= \sum_{z, w_1} P(y|z) \sum_{u_1, u_2} P(w_1|u_1, u_2)P(z|x, u_2)P(u_1)P(u_2) \quad (\sum_{w_2} P(w_2|w_1) = 1) \\ &= \sum_z P(y|z) \sum_{u_1, u_2} P(z|x, u_2)P(u_1)P(u_2) \quad (\sum_{w_1} P(w_1|u_1, u_2) = 1) \\ &= \sum_z P(y|z)Q[\{Z\}]. \end{aligned} \quad (5.79)$$

Note that the key reason for the factors of W_1 and W_2 to be summed out is that $Q[V']$ factorizes according to the subgraph $G_{V'}$ and that W_1 and W_2 are not ancestors of Y in $G_{V'}$ (see Figure 5.8(b)). The problem of computing $P_x(y)$ is then reduced to computing $Q[\{Z\}]$, which may be computed from $Q[S^X]$. Again, noticing that W_1 is not an ancestor of Z in G_{S^X} (see Figure 5.8(c)), we sum W_1 over Eq. (5.76):

$$\sum_{w_1} Q[S^X] = Q[\{X, Z\}] \quad (5.80)$$

$$= \sum_{u_1, u_2} P(x|w_2, u_1)P(z|x, u_2)P(u_1)P(u_2) \quad (5.81)$$

$$= (\sum_{u_1} P(x|w_2, u_1)P(u_1))(\sum_{u_2} P(z|x, u_2)P(u_2)) \quad (5.82)$$

$$= Q[\{X\}]Q[\{Z\}] \quad (5.83)$$

To compute $Q[\{X\}]$ and $Q[\{Z\}]$, summing Z over Eq. (5.82), we obtain

$$Q[\{X\}] = \sum_{z,w_1} Q[S^X] = \sum_{w_1} P(x|w_2, w_1)P(w_1), \quad (5.84)$$

and from Eq. (5.83)

$$Q[\{Z\}] = \frac{\sum_{w_1} Q[S^X]}{Q[\{X\}]} = \frac{\sum_{w_1} P(z, x|w_2, w_1)P(w_1)}{\sum_{w_1} P(x|w_2, w_1)P(w_1)}. \quad (5.85)$$

Finally, substituting the expression for $Q[\{Z\}]$ (5.85) into Eq. (5.79), we obtain

$$P_x(y) = \sum_z P(y|z) \frac{\sum_{w_1} P(z, x|w_2, w_1)P(w_1)}{\sum_{w_1} P(x|w_2, w_1)P(w_1)}. \quad (5.86)$$

From this example, we see that the quantity $Q[C]$ defined in Eq. (5.28) plays an important role in identifying $P_x(y)$. The ingredients that allowed us to compute $P_x(y)$ were (i) our ability to sum out some factors from $Q[V']$ as in Eqs. (5.79), due to the fact that W_1 and W_2 are not ancestors of Y in $G_{V'}$; (ii) our ability to compute $Q[\{X\}]$ and $Q[\{Z\}]$ from $Q[\{X, Z\}]$, which is due to the decomposition of $Q[\{X, Z\}]$ into the product of $Q[\{X\}]$ and $Q[\{Z\}]$ (Eq. (5.83)) because in the graph $G_{\{X, Z\}}$ (Figure 5.8(d)), $\{X, Z\}$ is partitioned into two components $\{X\}$ and $\{Z\}$. These two points correspond to Lemma 2 and 3 in Chapter 4 respectively, which will be presented next.

5.3.3 Lemmas

Lemma 2 and 3 in Chapter 4 will be instrumental in facilitating the general computing of causal effects $P_x(s)$. The two lemmas are presented in the following tailored for the situation of semi-Markovian models.

Lemma 10 *Let $W \subseteq C \subseteq V$, and $W' = C \setminus W$. If W is an ancestral set in the subgraph G_C (that is, $An(W)_{G_C} = W$), or equivalently, if none of the parents of W is in W' ($Pa(W) \cap W' = \emptyset$), then*

$$\sum_{W'} Q[C] = Q[W]. \quad (5.87)$$

Lemma 10 provides a condition under which summing $Q[C]$ over some variables is equivalent to removing the corresponding factors. It also provides a condition under which we can compute $Q[W]$ from $Q[C]$, where W is a subset of C , by simply summing $Q[C]$ over the remaining variables (in $C \setminus W$), like ordinary marginalization in probability theory.

Lemma 11 Let $H \subseteq V$, and assume that H is partitioned into c -components H_1, \dots, H_l in the subgraph G_H . Then we have

(i) $Q[H]$ decomposes as

$$Q[H] = \prod_i Q[H_i]. \quad (5.88)$$

(ii) Let k be the number of variables in H , and let a topological order of the variables in H be $V_{h_1} < \dots < V_{h_k}$ in G_H . Let $H^{(i)} = \{V_{h_1}, \dots, V_{h_i}\}$ be the set of variables in H ordered before V_{h_i} (including V_{h_i}), $i = 1, \dots, k$, and $H^{(0)} = \emptyset$. Then each $Q[H_j]$, $j = 1, \dots, l$, is computable from $Q[H]$ and is given by

$$Q[H_j] = \prod_{\{i | V_{h_i} \in H_j\}} \frac{Q[H^{(i)}]}{Q[H^{(i-1)}]}, \quad (5.89)$$

where each $Q[H^{(i)}]$, $i = 0, 1, \dots, k$, is given by

$$Q[H^{(i)}] = \sum_{h \setminus h^{(i)}} Q[H]. \quad (5.90)$$

Lemma 11 generalizes Lemma 7 to proper subgraphs of G .

The use of Lemma 11 can be shown with the example studied in Section 5.3.2, where the subgraph $G_{\{X, Z\}}$ (Figure 5.8(d)) is partitioned into two c -components $\{X\}$ and $\{Z\}$, and therefore $Q[\{X\}]$ and $Q[\{Z\}]$ are both computable from $Q[\{X, Z\}]$. We can check that Eqs. (5.84) and (5.85) satisfy (5.89).

Next, we present a procedure for computing $P_x(s)$ based on Lemmas 7, 10, and 11.

5.3.4 Computing $P_x(s)$

Let V be partitioned into c -components S^X, S_1, \dots, S_k , where $X \in S^X$, and let $V' = V \setminus \{X\}$. We have

$$P(v) = Q[V] = Q[S^X] \prod_i Q[S_i], \quad (5.91)$$

and

$$P_x(v') = Q[V'] = Q[S^X \setminus \{X\}] \prod_i Q[S_i]. \quad (5.92)$$

We want to compute

$$P_x(s) = \sum_{V' \setminus S} P_x(v') = \sum_{V' \setminus S} Q[V']. \quad (5.93)$$

Let $D = An(S)_{G_{V'}}$. By Lemma 10, Eq. (5.93) becomes

$$P_x(s) = \sum_{D \setminus S} \sum_{V' \setminus D} Q[V'] = \sum_{D \setminus S} Q[D]. \quad (5.94)$$

Let $D^X = D \cap S^X$, and $D_i = D \cap S_i, i = 1, \dots, k$. From Eq. (5.92), $Q[D]$ can be written as

$$Q[D] = Q[D^X] \prod_i Q[D_i] \quad (5.95)$$

D_i is an ancestral set in G_{S_i} from its definition, hence by Lemma 10,

$$Q[D_i] = \sum_{S_i \setminus D_i} Q[S_i], \quad i = 1, \dots, k. \quad (5.96)$$

However, D^X may not be an ancestral set in G_{S^X} (although it is an ancestral set in $G_{S^X \setminus \{X\}}$), because X could be an ancestor of D^X . Combining Eqs. 5.94–5.96, we obtain

$$P_x(s) = \sum_{D \setminus S} Q[D^X] \prod_i \sum_{S_i \setminus D_i} Q[S_i]. \quad (5.97)$$

Assume that in the graph G_{D^X} , D^X is partitioned into c-components D_1^X, \dots, D_l^X . Then $Q[D^X] = \prod_j Q[D_j^X]$, and we obtain

$$P_x(s) = \sum_{D \setminus S} \prod_j Q[D_j^X] \prod_i \sum_{S_i \setminus D_i} Q[S_i]. \quad (5.98)$$

Since all the c-factors $Q[S_i]$'s are identifiable, we obtain that $P_x(s)$ is identifiable if all $Q[D_j^X]$'s are identifiable.

Since $D_j^X \subset S^X$, $Q[D_j^X]$ is identifiable if it is computable from $Q[S^X]$. Next, we study the conditions for $Q[D_j^X]$ to be computable from $Q[S^X]$. Let $F = An(D_j^X)_{G_{S^X}}$.

- If $F = D_j^X$, that is, if D_j^X is an ancestral set in G_{S^X} , then by Lemma 10, $Q[D_j^X]$ can be computed as

$$Q[D_j^X] = \sum_{S^X \setminus D_j^X} Q[S^X]. \quad (5.99)$$

- If $F = S^X$, we are unable to determine whether $Q[D_j^X]$ is computable from $Q[S^X]$ at this moment.

Function Identify(C, T, Q)

INPUT: $C \subseteq T \subseteq V$, $Q = Q[T]$. Assuming G_T is composed of one single c-component.

OUTPUT: Expression for $Q[C]$ in terms of Q or fail to determine.

Let $A = An(C)_{G_T}$.

- IF $A = C$, output $Q[C] = \sum_{T \setminus C} Q$.
- IF $A = T$, output FAIL.
- IF $C \subset A \subset T$
 1. Assume that in G_A , C is contained in a c-component T' .
 2. Compute $Q[T']$ from $Q[A] = \sum_{T \setminus A} Q$ by Lemma 11.
 3. Output Identify($C, T', Q[T']$).

Figure 5.9: A function determining if $Q[C]$ is computable from $Q[T]$.

- Assume that $D_j^X \subset F \subset S^X$. By Lemma 10, we have

$$Q[F] = \sum_{S^X \setminus F} Q[S^X]. \quad (5.100)$$

Assume that in the graph G_F , D_j^X is contained in a c-component H (the variables in D_j^X are connected by bidirected paths among themselves hence belong to one same c-component). By Lemma 11, $Q[H]$ can be computed from $Q[F]$ and thus is identifiable. We obtain that the problem of whether $Q[D_j^X]$ is computable from $Q[S^X]$ is reduced to that whether $Q[D_j^X]$ is computable from $Q[H]$.

The preceding analysis gives a recursive procedure for determining whether $Q[D_j^X]$ is computable from $Q[S^X]$; at each step, we either find an expression for $Q[D_j^X]$, find the problem indeterminable, or reduce the problem to a simpler one in the sense that $H \subset S^X$. In general, let $C \subseteq T \subseteq V$; a recursive algorithm for determining if $Q[C]$ is computable from $Q[T]$ is presented in Figure 5.9.

In summary, an algorithm for computing $P_x(s)$ is given in Figure 5.10. The procedure consists of three basic phases. In phase-1, we compute the expressions for all c-factors and find (graphically) the sets D_j^X from the graph G . In phase-2, we attempt to compute $Q[D_j^X]$'s from $Q[S^X]$ by calling the function Identify($D_j^X, S^X, Q[S^X]$) given in Figure 5.9. In phase-3, if all $Q[D_j^X]$'s are computable, we output the expression for $P_x(s)$ given in Eq. (5.98).

Algorithm 4 (Computing $P_x(s)$)

INPUT: a set $S \subset V$.

OUTPUT: the expression for $P_x(s)$ or fail to determine.

Phase-1:

1. Find the c-components of G : S^X, S_1, \dots, S_k , where $X \in S^X$.
2. Compute the c-factors $Q[S^X], Q[S_1], \dots, Q[S_k]$ by Lemma 7.
3. Let $D = An(S)_{G_{V \setminus \{X\}}}$, $D^X = D \cap S^X$.
4. Let the c-components of G_{D^X} be $D_j^X, j = 1, \dots, l$.

Phase-2:

For each set D_j^X :

Compute $Q[D_j^X]$ from $Q[S^X]$ by calling the function $Identify(D_j^X, S^X, Q[S^X])$ given in Figure 5.9. If the function returns FAIL, then stop and output FAIL.

Phase-3:

Output $P_x(s) = \sum_{D \setminus S} \prod_j Q[D_j^X] \prod_i \sum_{S_i \setminus D} Q[S_i]$.

Figure 5.10: An algorithm for computing $P_x(s)$

From the preceding analysis, we see that the problem of identifying $P_x(s)$ is reduced to that of computing $Q[C]$ from $Q[T]$ for some sets $C \subset T \subseteq V$, for which we give an algorithm in Figure 5.9. Now the open problem is: Is $Q[C]$ computable from $Q[T]$ if (i) G_C has only one c-component (C itself), (ii) G_T has only one c-component (T itself), and (iii) in G_T , all variables in $T \setminus C$ are ancestors of C ($An(C)_{G_T} = T$)?

5.3.5 Useful graphical criteria

We have given a procedure for determining the identifiability of $P_x(s)$ and finding its expression (when identifiable) in Figure 5.10. Next, we give some graphical criteria based on Algorithm 4 which can be used for quickly judging the identifiability of $P_x(s)$ by looking at the causal graph G .

The idea lies in systematically removing certain nonessential nodes from G till Theorem 17 is applicable (or no more nodes can be removed). First, Lemma 9 can be used to remove nonancestors of S from G . Next, we show that all variables that are not in the same c-components as X can be removed. To prove this conclusion, we present a utility lemma first. Let $A \subseteq B \subseteq V$. We use $Q[A]_{G_B}$ to denote the function $Q[A] = \sum_u \prod_{\{i|V_i \in A\}} P(v_i | pa'_i, u^i) P(u)$ where $PA'_i = PA_i \cap B$.

The difference between $Q[A]_{G_B}$ and $Q[A] = Q[A]_{G_V}$ is that some parents of A in G are removed in G_B .

Lemma 12 *Let $A \subseteq B \subseteq V$. $Q[A]$ is computable from $Q[B]$ if and only if $Q[A]_{G_B}$ is computable from $Q[B]_{G_B}$.*

Proof: (only if) By Lemma 8.

(if) Proof by contradiction. Assume that $Q[A]$ is not computable from $Q[B]$, then there exist two models, M_1 and M_2 , with the same causal graph G , satisfying

$$Q^{M_k}[B](b, c) = \sum_u \prod_{\{i|V_i \in B\}} P^{M_k}(v_i | pa'_i, c_i, u^i) P^{M_k}(u), \quad k = 1, 2, \quad (5.101)$$

where $PA'_i = PA_i \cap B$, $C_i = PA_i \setminus B$, and $C = \cup_i C_i$, such that

$$Q^{M_1}[B](b, c) = Q^{M_2}[B](b, c) > 0, \text{ for all values } b, c, \quad (5.102)$$

but

$$Q^{M_1}[A](b', c') \neq Q^{M_2}[A](b', c'), \text{ for some particular value } b', c'. \quad (5.103)$$

$Q[B]_{G_B}$ can be written as

$$Q[B]_{G_B}(b) = \sum_u \prod_{\{i|V_i \in B\}} P(v_i | pa'_i, u^i) P(u). \quad (5.104)$$

We construct two models, M'_1 and M'_2 , with the same causal graph G_B as

$$P^{M'_k}(v_i | pa'_i, u^i) = P^{M_k}(v_i | pa'_i, C_i = c'_i, u^i), \quad k = 1, 2, \quad (5.105)$$

$$P^{M'_k}(u) = P^{M_k}(u), \quad k = 1, 2. \quad (5.106)$$

Then we have

$$Q[B]_{G_B}^{M'_k}(b) = Q[B]^{M_k}(b, c'), \text{ and } Q[A]_{G_B}^{M'_k}(b) = Q[A]^{M_k}(b, c'), \quad k = 1, 2. \quad (5.107)$$

From Eqs. (5.107), (5.102) and (5.103), we obtain

$$Q^{M'_1}[B]_{G_B}(b) = Q^{M'_2}[B]_{G_B}(b) > 0, \text{ for all values } b, \quad (5.108)$$

and

$$Q^{M'_1}[A]_{G_B}(b') \neq Q^{M'_2}[A]_{G_B}(b'), \text{ for the value } b', \quad (5.109)$$

which says that $Q[A]_{G_B}$ is not computable from $Q[B]_{G_B}$. \square

Using Lemma 12, we obtain the following lemma which reduces the identifiability problem to some subgraph of G .

Lemma 13 Assume that X is in the c-component S^X , and let $D^X = An(S)_{G_{V \setminus \{X\}}} \cap S^X$. Then $P_x(s)$ is identifiable if in the graph G_{S^X} , $P_x(D^X)$ is identifiable.

Proof: From Eq. (5.97), $P_x(s)$ is identifiable if $Q[D^X]$ is identifiable. By Lemma 12, $Q[D^X]$ is identifiable if $Q[D^X]_{G_{S^X}}$ is identifiable. Let $E^X = (S^X \setminus D^X) \setminus \{X\}$. In G_{S^X} , we have

$$P_x(D^X) = \sum_{E^X} P_x(S^X \setminus \{X\}) = \sum_{E^X} Q[S^X \setminus \{X\}]_{G_{S^X}} = Q[D^X]_{G_{S^X}}, \quad (5.110)$$

where we used Lemma 10 in the last step. Hence we obtain that $P_x(s)$ is identifiable if $P_x(D^X)$ is identifiable in G_{S^X} . \square

Lemma 9 and 13 reduce the original problems of deciding the identifiability of $P_x(s)$ in G to (usually simpler) problems of identifying the causal effect of X on a different set of variables in some subgraphs of G . If the latter problem is not recognized to be identifiable (via Theorem 17), we can of course repeat the process and attempt to reduce it further, using Lemma 9 and 13 alternatively.⁴ Such recursive application of Lemma 9 and 13 is illustrated in the next example.

5.3.6 An example

Consider the problem of identifying $P_x(y)$ in Figure 5.11(a). By Lemma 9, $P_x(y)$ is identifiable in Figure 5.11(a) if it is identifiable in Figure 5.11(b), then by Lemma 13, if it is identifiable in Figure 5.11(c). After applying Lemma 9 and 13 again (see Figure 5.11(d) and (e)), the problem is finally reduced to whether $P_x(y)$ is identifiable in Figure 5.11(f), which is obviously true, and we conclude that $P_x(y)$ is identifiable in Figure 5.11(a).

We now demonstrate the use of Algorithm 4 by computing $P_x(y)$ in Figure 5.11(a).

Phase-1:

1. The whole graph is one c-component.
2. $D^X = D = An(\{Y\})_{G_{V \setminus \{X\}}} = \{Y\}$.
3. We want to compute $P_x(y) = Q[\{Y\}]$.

Phase-2:

⁴Note that some causal effects identified by Algorithm 4 may not be identified by repeatedly using Lemma 9 and 13 which are meant for quick judgement only.

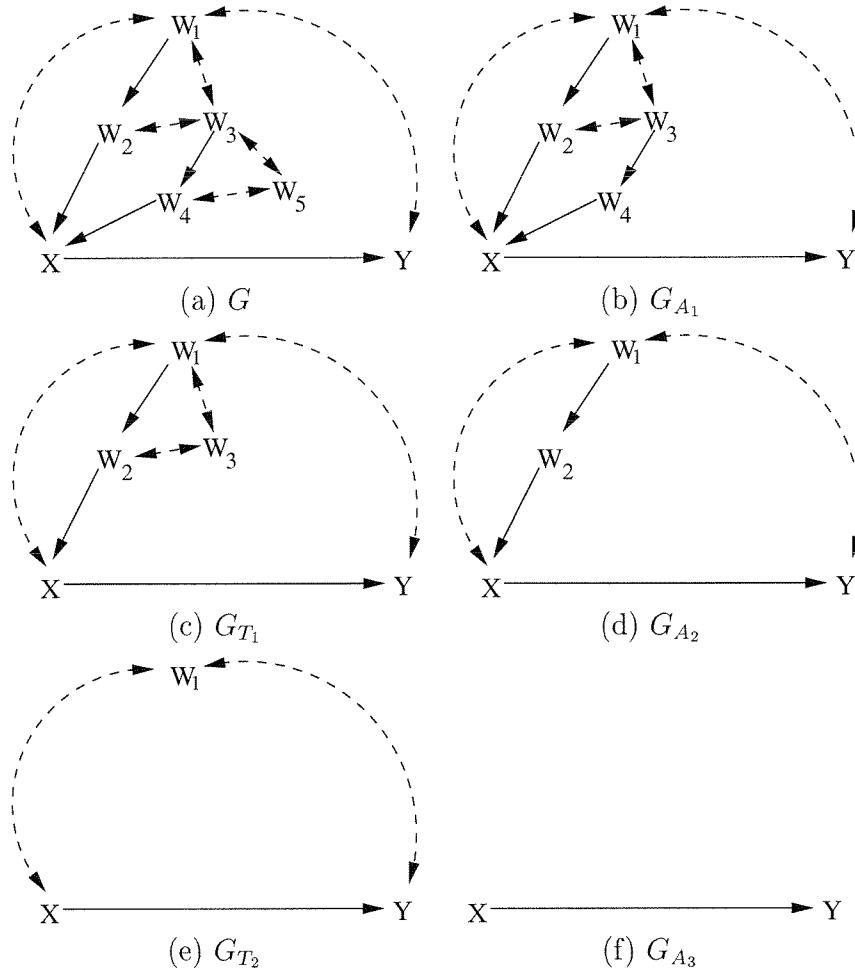


Figure 5.11: Subgraphs of G used in computing $P_x(y)$.

1. Compute $Q[\{Y\}]$ by calling the function $\text{Identify}(\{Y\}, V, P(v))$ in Figure 5.9. Let $A_1 = An(\{Y\})_G = \{X, Y, W_1, W_2, W_3, W_4\}$. We have $\{Y\} \subset A_1 \subset V$. The graph G_{A_1} (Figure 5.11(b)) has two c-components: $T_1 = \{X, Y, W_1, W_2, W_3\}$ and $\{W_4\}$, and we have

$$Q[A_1] = \sum_{w_5} P(v) = P(a_1) = Q[T_1]Q[\{W_4\}]. \quad (5.111)$$

A topological sort over A_1 is: $W_3 < W_4 < W_1 < W_2 < X < Y$. By Lemma 11, we obtain

$$Q[\{W_4\}] = \frac{Q[\{W_4, W_3\}]}{Q[\{W_3\}]} = \frac{\sum_{w_1, w_2, x, y} P(a_1)}{\sum_{w_4, w_1, w_2, x, y} P(a_1)} = P(w_4|w_3), \quad (5.112)$$

and from (5.111),

$$\begin{aligned} Q[T_1] &= P(a_1)/P(w_4|w_3) \\ &= P(x, y, w_1, w_2|w_3, w_4)P(w_3) \\ &= P(x, y|w_1, w_2, w_3, w_4)P(w_1, w_2, w_3). \end{aligned} \quad (5.113)$$

2. Call the function $\text{Identify}(\{Y\}, T_1, Q[T_1])$.

Let $A_2 = An(\{Y\})_{G_{T_1}} = \{X, Y, W_1, W_2\}$ (see Figure 5.11(c)). We have $\{Y\} \subset A_2 \subset T_1$. The graph G_{A_2} (Figure 5.11(d)) has two c-components: $T_2 = \{X, Y, W_1\}$ and $\{W_2\}$, and we have

$$Q[A_2] = \sum_{w_3} Q[T_1] = Q[T_2]Q[\{W_2\}]. \quad (5.114)$$

A topological sort over A_2 is: $W_1 < W_2 < X < Y$. By Lemma 11, we obtain

$$Q[\{W_2\}] = \frac{Q[\{W_2, W_1\}]}{Q[\{W_1\}]} = \frac{\sum_{x, y} Q[A_2]}{\sum_{w_2, x, y} Q[A_2]} = P(w_2|w_1), \quad (5.115)$$

and from (5.114) and (5.113),

$$\begin{aligned} Q[T_2] &= \sum_{w_3} Q[T_1]/P(w_2|w_1) \\ &= \sum_{w_3} P(x, y|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1). \end{aligned} \quad (5.116)$$

3. Call the function $\text{Identify}(\{Y\}, T_2, Q[T_2])$. Let $A_3 = An(\{Y\})_{G_{T_2}} = \{X, Y\}$ (see Figure 5.11(e)). We have $\{Y\} \subset A_3 \subset T_2$. The graph G_{A_3} (Figure 5.11(f)) has two c-components: $\{X\}$ and $\{Y\}$, and we have

$$Q[A_3] = \sum_{w_1} Q[T_2] = Q[\{X\}]Q[\{Y\}]. \quad (5.117)$$

The only admissible order over A_3 is: $X < Y$. By Lemma 11, we obtain

$$Q[\{X\}] = \sum_y \sum_{w_1} Q[T_2] = \sum_{w_1, w_3} P(x|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1), \quad (5.118)$$

and

$$\begin{aligned} Q[\{Y\}] &= (\sum_{w_1} Q[T_2])/Q[\{X\}] \\ &= \frac{\sum_{w_1, w_3} P(x, y|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1)}{\sum_{w_1, w_3} P(x|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1)}. \end{aligned} \quad (5.119)$$

Phase-3:

Finally, we obtain

$$P_x(y) = Q[\{Y\}] = \frac{\sum_{w_1, w_3} P(x, y|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1)}{\sum_{w_1, w_3} P(x|w_1, w_2, w_3, w_4)P(w_3|w_1, w_2)P(w_1)}. \quad (5.120)$$

5.3.7 Galles&Pearl's graphical criterion vs. *do*-calculus

[GP95] claimed that their graphical criterion will embrace all cases where identification is verifiable by *do*-calculus. Here we show that their criterion is not complete in this sense. Consider the problem of identifying $P_x(z)$ in Figure 5.8(a). Neither “back-door” nor “front-door” criterion is applicable. The graphical criterion in [GP95] also fails because there is no set which can block all back-door paths from X to Z . However we have that $P_x(z) = Q[\{Z\}]$ is identifiable and is given in Eq. (5.85). $P_x(z)$ can also be computed by *do*-calculus as

$$P(z|\hat{x}) = P(z|\hat{x}, \hat{w}_1) \quad (5.121)$$

$$= P(z|x, \hat{w}_1) \quad (5.122)$$

$$= P(z|x, w_2, \hat{w}_1) \quad (5.123)$$

$$= \frac{P(z, x, w_2|\hat{w}_1)}{P(x, w_2|\hat{w}_1)} \quad (5.124)$$

$$= \frac{\sum_{w_1} P(z, x|w_2, w_1)P(w_1)}{\sum_{w_1} P(x|w_2, w_1)P(w_1)} \quad (5.125)$$

Hence we see that the graphical criterion in [GP95] is *not* complete with respect to *do*-calculus. [GP95] may have failed to consider the possibility of removing a hat by transforming Eq. (5.123) to (5.124).

5.4 Identification of $P_t(s)$

So far, we have assumed that intervention is applied to a single variable X . In this section we study the problem of identifying $P_t(s)$ where S and T are arbitrary (disjoint) subsets of V . We will show that, as for identifying $P_x(s)$, the problem of identifying $P_t(s)$ is also reduced to identifying $Q[C]$ from $Q[C']$ for some sets $C \subset C'$, and we give a procedure for computing $P_t(s)$.

5.4.1 Computing $P_t(s)$

Let $T' = V \setminus T$, we want to compute

$$P_t(s) = \sum_{T' \setminus S} P_t(t') = \sum_{T' \setminus S} Q[T']. \quad (5.126)$$

Let $D = An(S)_{G_{T'}}$. Then by Lemma 10,

$$P_t(s) = \sum_{D \setminus S} \sum_{T' \setminus D} Q[T'] = \sum_{D \setminus S} Q[D]. \quad (5.127)$$

Let V be partitioned into c-components S_1, \dots, S_k , and let $D_i = D \cap S_i, i = 1, \dots, k$. Eq. (5.127) can be rewritten as

$$P_t(s) = \sum_{D \setminus S} \prod_i Q[D_i]. \quad (5.128)$$

We obtain that $P_t(s)$ is identifiable if all $Q[D_i]$'s are identifiable. Assume that the graph G_{D_i} is partitioned into c-components D_{i1}, \dots, D_{ik_i} . Then

$$Q[D_i] = \prod_j Q[D_{ij}], \quad i = 1, \dots, k. \quad (5.129)$$

We obtain

$$P_t(s) = \sum_{D \setminus S} \prod_i \prod_j Q[D_{ij}]. \quad (5.130)$$

Hence $P_t(s)$ is identifiable if all $Q[D_{ij}]$'s are identifiable. Whether $Q[D_{ij}]$ is identifiable can be determined by using the function $\text{Identify}(D_{ij}, S_i, Q[S_i])$ given in Figure 5.9.

Algorithm 5 (Computing $P_t(s)$)

INPUT: two disjoint sets $S, T \subset V$.

OUTPUT: the expression for $P_t(s)$ or fail to determine.

Phase-1:

1. Find the c-components of G : S_1, \dots, S_k .
2. Compute the c-factors $Q[S_1], \dots, Q[S_k]$ by Lemma 7.
3. Let $D = An(S)_{G_{V \setminus T}}$, $D_i = D \cap S_i$, $i = 1, \dots, k$.
4. Let the c-components of G_{D_i} be D_{ij} , $j = 1, \dots, k_i$, $i = 1, \dots, k$.

Phase-2:

For each set D_{ij} :

Compute $Q[D_{ij}]$ from $Q[S_i]$ by calling the function $Identify(D_{ij}, S_i, Q[S_i])$ in Figure 5.9. If the function returns FAIL, then stop and output FAIL.

Phase-3:

Output $P_t(s) = \sum_{D \setminus S} \prod_i \prod_j Q[D_{ij}]$.

Figure 5.12: An algorithm for computing $P_t(s)$

In summary, an algorithm for computing $P_t(s)$ is given in Figure 5.12. The procedure consists of three basic phases. In phase-1, we compute the expressions for all c-factors and find (graphically) the sets D_{ij} from the graph G . In phase-2, we attempt to compute $Q[D_{ij}]$'s by calling the function $Identify(D_{ij}, S_i, Q[S_i])$ given in Figure 5.9. In phase-3, if all $Q[D_{ij}]$'s are identifiable, we output the expression for $P_t(s)$ given in Eq. (5.130).

5.4.2 Useful graphical criteria

Next, we give some graphical criteria for quick judgement of the identifiability of $P_t(s)$ by looking at the causal graph G . First we give some graphical conditions for identifying $P_t(v) = P_t(v \setminus t)$, the causal effect of T on all other variables in V . The following criterion is a corollary of Lemma 7.

Theorem 18 *If there is no bidirected edge connecting variables in a set T to variables not in T , then $P_t(v)$ is identifiable. Let a topological order over V be $V_1 < \dots < V_n$, and let $V^{(i)} = \{V_1, \dots, V_i\}$, $i = 1, \dots, n$, and $V^{(0)} = \emptyset$. Then*

$P_t(v)$ is given by

$$P_t(v \setminus t) = \prod_{\{i|V_i \in V \setminus T\}} P(v_i | pa(C_i) \setminus \{v_i\}), \quad (5.131)$$

where C_i is the c-component of $G_{V^{(i)}}$ that contains V_i .

In general, let $T' = V \setminus T$, let V be partitioned into c-components S_1, \dots, S_k , and let $T_i = T \cap S_i, T'_i = T' \cap S_i, i = 1, \dots, k$. We have

$$P_t(t') = \prod_i Q[T'_i]. \quad (5.132)$$

Hence $P_t(t')$ is identifiable if and only if each $Q[T'_i]$ is computable from $Q[S_i]$. On the other hand, we have

$$P_{t_j}(v \setminus t_j) = Q[T'_j] \prod_{i \neq j} Q[S_i]. \quad (5.133)$$

Hence $P_{t_j}(v \setminus t_j)$ is identifiable if and only if $Q[T'_j]$ is computable from $Q[S_j]$. And we obtain the following lemma.

Lemma 14 *Let V be partitioned into c-components S_1, \dots, S_k , and let $T_i = T \cap S_i, i = 1, \dots, k$. $P_t(v)$ is identifiable if and only if each $P_{t_i}(v), i = 1, \dots, k$, is identifiable.*

In the subgraph G_{S_j} ,

$$P(s_j) = Q[S_j]_{G_{S_j}}, \text{ and } P_{t_j}(s_j \setminus t_j) = Q[T'_j]_{G_{S_j}}. \quad (5.134)$$

Hence by Lemma 12, $Q[C_j]$ is computable from $Q[S_j]$ if and only if $P_{t_j}(s_j \setminus t_j)$ is identifiable in G_{S_j} , which gives the following lemma.

Lemma 15 *Let S_i be a c-component of G , and $T_i \subseteq S_i$. $P_{t_i}(v)$ is identifiable if and only if $P_{t_i}(s_i)$ is identifiable in the graph G_{S_i} .*

One simple condition for $Q[T'_i]$ to be computable from $Q[S_i]$ is that T'_i is an ancestral set in G_{S_i} , or T_i contains its own descendants in G_{S_i} . Under this condition, by Lemma 10,

$$Q[T'_i] = \sum_{T_i} Q[S_i]. \quad (5.135)$$

And we obtain the following theorem.

Theorem 19 Let S_i be a c-component of G , and $T_i \subseteq S_i$. If the children of variables in T_i are either in T_i or outside of S_i (i.e. T_i contains its own descendants in G_{S_i}), then $P_{t_i}(v)$ is identifiable, and is given by

$$P_{t_i}(v \setminus t_i) = \frac{P(v)}{Q[S_i]} \sum_{T_i} Q[S_i]. \quad (5.136)$$

Next, we give some graphical conditions for quick judgment of the identifiability of $P_t(s)$.

Lemma 16 Let V be partitioned into c-components S_1, \dots, S_k . Let $T_i = T \cap S_i$, $D_i = An(S)_{G_{V \setminus T}} \cap S_i$, $i = 1, \dots, k$. Then $P_t(s)$ is identifiable if every $P_{t_i}(d_i)$ is identifiable in G_{S_i} for $i = 1, \dots, k$.

Proof: From Eq. (5.128), $P_t(s)$ is identifiable if each $Q[D_i]$ is identifiable. By Lemma 12, $Q[D_i]$ is computable from $Q[S_i]$ if $Q[D_i]_{G_{S_i}}$ is computable from $Q[S_i]_{G_{S_i}}$. Let $T'_i = S_i \setminus T_i$. In G_{S_i} , we have

$$P_{t_i}(d_i) = \sum_{T'_i \setminus D_i} P_{t_i}(t'_i) = \sum_{T'_i \setminus D_i} Q[T'_i]_{G_{S_i}} = Q[D_i]_{G_{S_i}}, \quad (5.137)$$

where we used Lemma 10 in the last step. Hence we obtain that $P_t(s)$ is identifiable if each $P_{t_i}(d_i)$ is identifiable in G_{S_i} . \square

Lemma 17 Let $T_1 = T \cap An(S)$. $P_t(s)$ is identifiable if and only if $P_{t_1}(s)$ is identifiable in $G_{An(S)}$.

Proof: It is well-known that $P_t(s) = P_{t_1}(s)$. The rest of the proof is the same as Lemma 9. \square

Lemma 16 and 17 reduce the original problems of deciding the identifiability of $P_t(s)$ in G to some (usually simpler) identifiability problems in subgraphs of G . They can be repeatedly applied to further reduce the problems, till inapplicable or till those problems are recognized to be identifiable (for example, via Theorem 17 or 19).

5.4.3 Examples

Next, we study some examples, to illustrate the use of Algorithm 5 and the graphical criteria in Section 5.4.2.

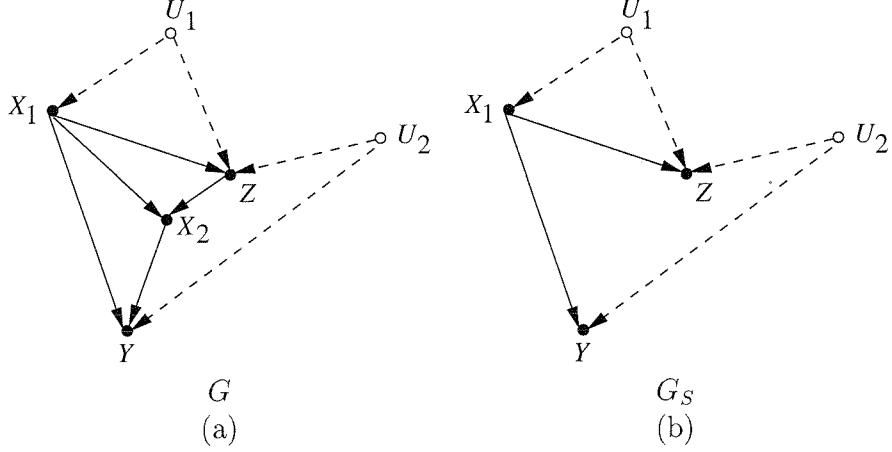


Figure 5.13: By Lemma 16, $P_{x_1x_2}(y)$ is identifiable if $P_{x_1}(y)$ is identifiable in G_S .

Consider the problem of identifying $P_{x_1x_2}(y)$ in Figure 5.13(a), which was studied in [PR95]. G has two c-components $S = \{X_1, Z, Y\}$ and $\{X_2\}$, and X_1 and X_2 are in different c-components. Letting $C = V \setminus \{X_1, X_2\} = \{Y, Z\}$, then $An(\{Y\})_{G_C} = \{Y\} \subset S$. By Lemma 16 we have that $P_{x_1x_2}(y)$ is identifiable if $P_{x_1}(y)$ is identifiable in the subgraph G_S (Figure 5.13(b)). Since the latter is true by Theorem 17, we conclude that $P_{x_1x_2}(y)$ is identifiable. Next we compute $P_{x_1x_2}(y)$. We have

$$P(v) = P(x_2|x_1, z)Q[S], \quad (5.138)$$

from which we obtain

$$Q[S] = P(v)/P(x_2|x_1, z) = P(y|x_1, x_2, z)P(x_1, z). \quad (5.139)$$

$P_{x_1x_2}(y)$ is computed as

$$P_{x_1x_2}(y) = \sum_z Q[\{Y, Z\}] = Q[\{Y\}], \quad (5.140)$$

which can be computed by calling $\text{Identify}(\{Y\}, S, Q[S])$ in Figure 5.9. Let $A = An(\{Y\})_{G_S} = \{X_1, Y\}$. We have $\{Y\} \subset A \subset S$. The graph G_A has two c-components: $\{X_1\}$ and $\{Y\}$, and we have

$$Q[A] = \sum_z Q[S] = Q[\{X_1\}]Q[\{Y\}]. \quad (5.141)$$

The only admissible order over A is: $X_1 < Y$. By Lemma 11, we obtain

$$Q[\{X_1\}] = \sum_y \sum_z Q[S] = P(x_1), \quad (5.142)$$

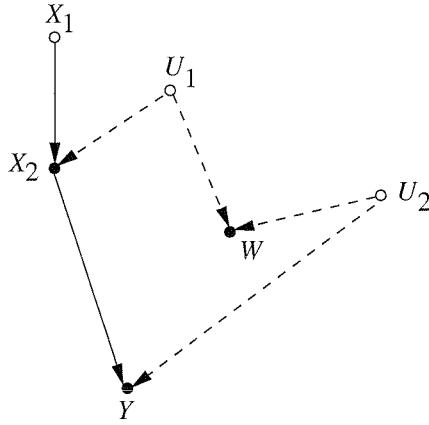


Figure 5.14: By Lemma 16, $P_{x_1x_2}(y)$ is identifiable if $P_{x_2}(y)$ is identifiable.

and

$$Q[\{Y\}] = \sum_z Q[S]/Q[\{X_1\}] = \sum_z P(y|x_1, x_2, z)P(z|x_1). \quad (5.143)$$

Finally, we obtain

$$P_{x_1x_2}(y) = Q[\{Y\}] = \sum_z P(y|x_1, x_2, z)P(z|x_1), \quad (5.144)$$

which coincides with Eq. (4.3) of [Pea00, page 122].

Consider the problem of identifying $P_{x_1x_2}(y)$ in Figure 5.14, which was studied in [PR95]. G has two c-components $S = \{X_2, W, Y\}$ and $\{X_1\}$, and X_1 and X_2 are in different c-components. Letting $C = V \setminus \{X_1, X_2\} = \{Y, W\}$, then $An(\{Y\})_{G_C} = \{Y\} \subset S$. By Lemma 16, $P_{x_1x_2}(y)$ is identifiable if $P_{x_2}(y)$ is identifiable in G_S . It is clear that $P_{x_2}(y)$ is identifiable (by Theorem 17), hence $P_{x_1x_2}(y)$ is identifiable.

Consider the problem of identifying $P_{x_1x_2}(y)$ in Figure 5.15, which was studied in [PR95]. G has three c-components $\{X_1\}, \{Y\}$, and $S = \{X_2, Z_1, Z'_1\}$, and X_1 and X_2 are in different c-components. By Lemma 14, $P_{x_1x_2}(v)$ is identifiable if both $P_{x_1}(v)$ and $P_{x_2}(v)$ are identifiable, which is true by Theorem 15. Therefore $P_{x_1x_2}(v)$ is identifiable. Next we compute $P_{x_1x_2}(v)$. We have

$$P(v) = P(x_1|z_1)P(y|x_2, z'_1)Q[S], \quad (5.145)$$

from which we obtain

$$Q[S] = P(v)/(P(x_1|z_1)P(y|x_2, z'_1)) = P(x_2, z'_1|x_1, z_1)P(z_1). \quad (5.146)$$

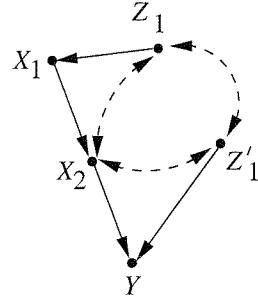


Figure 5.15: By Lemma 14, $P_{x_1 x_2}(v)$ is identifiable if both $P_{x_1}(v)$ and $P_{x_2}(v)$ are identifiable.

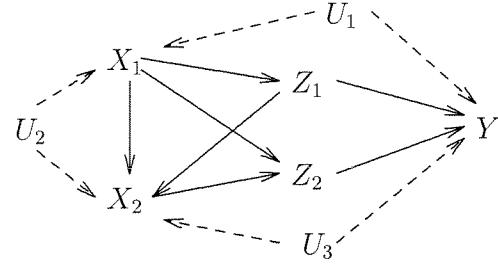


Figure 5.16: The problem of identifying $P_{x_1 x_2}(y)$ (from [KM99]).

$P_{x_1 x_2}(v)$ is computed as

$$\begin{aligned}
 P_{x_1 x_2}(y, z_1, z'_1) &= P(y|x_2, z'_1)Q[\{Z_1, Z'_1\}] \\
 &= P(y|x_2, z'_1) \sum_{x_2} Q[S] \\
 &= P(y|x_2, z'_1)P(z'_1|x_1, z_1)P(z_1) \\
 &= P(y|x_2, z'_1)P(z'_1, z_1).
 \end{aligned} \tag{5.147}$$

Next, consider the problem of identifying $P_{x_1 x_2}(y)$ in Figure 5.16, which was studied in [KM99]. X_1 and X_2 are in the same c-component $S = \{X_1, X_2, Y\}$, and their children other than X_2 itself are not in S , hence Theorem 19 is applicable and $P_{x_1 x_2}(v)$ is identifiable. We have

$$P(v) = P(z_1|x_1)P(z_2|x_1, x_2)Q[S], \tag{5.148}$$

from which we obtain

$$\begin{aligned}
 Q[S] &= P(v)/(P(z_1|x_1)P(z_2|x_1, x_2)) \\
 &= P(y|x_1, x_2, z_1, z_2)P(x_2|x_1, z_1)P(x_1).
 \end{aligned} \tag{5.149}$$

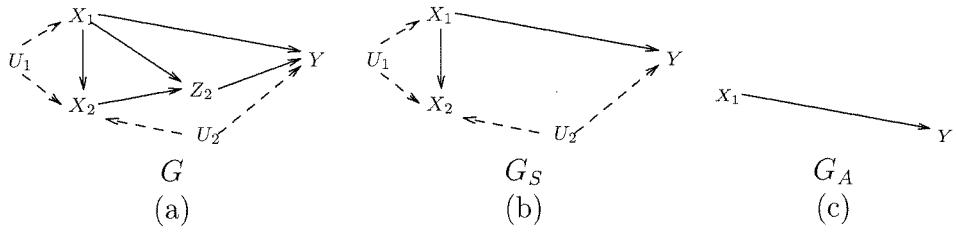


Figure 5.17: The problem of identifying $P_{x_1x_2}(y)$ (from [KM99]).

From Theorem 19, we have

$$\begin{aligned} P_{x_1x_2}(y, z_1, z_2) &= P(z_1|x_1)P(z_2|x_1, x_2) \sum_{x_1, x_2} Q[S] \\ &= P(z_1|x_1)P(z_2|x_1, x_2) \sum_{x'_1, x'_2} P(y|x'_1, x'_2, z_1, z_2)P(x'_2|x'_1, z_1)P(x'_1). \end{aligned} \quad (5.150)$$

We further obtain

$$P_{x_1x_2}(y) = \sum_{z_1, z_2} P(z_1|x_1)P(z_2|x_1, x_2) \sum_{x'_1, x'_2} P(y|x'_1, x'_2, z_1, z_2)P(x'_2|x'_1, z_1)P(x'_1), \quad (5.151)$$

which coincides with Eq. (3.12) of [KM99].

Consider the problem of identifying $P_{x_1x_2}(y)$ in Figure 5.17(a), which was studied in [KM99]. X_1 and X_2 are in the same c-component $S = \{X_1, X_2, Y\}$. By Lemma 15, $P_{x_1x_2}(v)$ is identifiable if $P_{x_1x_2}(y)$ is identifiable in G_S (Figure 5.17(b)). Let $A = An(\{Y\})_{G_S} = \{X_1, Y\}$. By Lemma 17, $P_{x_1x_2}(y)$ is identifiable in G_S if $P_{x_1}(y)$ is identifiable in the subgraph G_A (Figure 5.17(c)). Since $P_{x_1}(y)$ is obviously identifiable in G_A , we conclude that $P_{x_1x_2}(v)$ is identifiable. We have

$$P(v) = P(z_2|x_1, x_2)Q[S], \quad (5.152)$$

from which we obtain

$$Q[S] = P(v)/P(z_2|x_1, x_2) = P(y|z_2, x_1, x_2)P(x_1, x_2). \quad (5.153)$$

$P_{x_1x_2}(v)$ is computed as

$$P_{x_1x_2}(z_2, y) = P(z_2|x_1, x_2)Q[\{Y\}]. \quad (5.154)$$

$Q[\{Y\}]$ can be computed by calling $\text{Identify}(\{Y\}, S, Q[S])$ in Figure 5.9. We have

$$Q[A] = \sum_{x_2} Q[S] = Q[\{X_1\}]Q[\{Y\}], \quad (5.155)$$

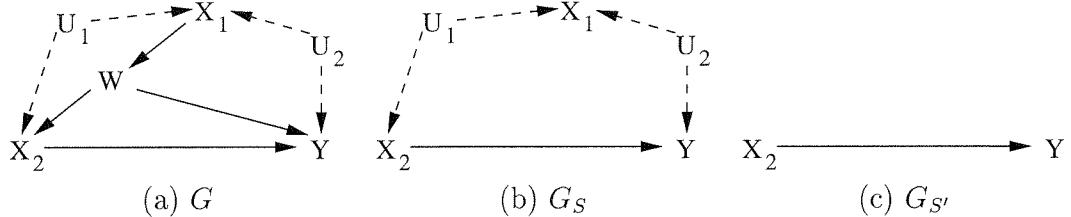


Figure 5.18: Subgraphs used in identifying $P_{x_1x_2}(w, y)$ in G .

from which we obtain

$$Q[\{X_1\}] = \sum_y Q[A] = P(x_1), \quad (5.156)$$

and

$$Q[\{Y\}] = \sum_{x_2} Q[S]/Q[\{X_1\}] = \sum_{x_2} P(y|z_2, x_1, x_2)P(x_2|x_1). \quad (5.157)$$

Finally, substituting (5.157) into (5.154), we obtain

$$P_{x_1x_2}(z_2, y) = P(z_2|x_1, x_2) \sum_{x'_2} P(y|z_2, x_1, x'_2)P(x'_2|x_1), \quad (5.158)$$

and

$$P_{x_1x_2}(y) = \sum_{z_2} P(z_2|x_1, x_2) \sum_{x'_2} P(y|z_2, x_1, x'_2)P(x'_2|x_1), \quad (5.159)$$

which coincides with Eq. (3.21) of [KM99].

In the examples studied so far, in Figure 5.13(a), 5.14, and 5.15, $P_{x_1x_2}(y)$ can be identified using the criteria given in [PR95]. In Figure 5.16 and 5.17(a), $P_{x_1x_2}(y)$ can be identified by the extended front-door criterion and the mixed-door criterion given in [KM99] respectively. Next we give an example shown in Figure 5.18(a), for which $P_{x_1x_2}(w, y)$ is identifiable, but none of the criteria in [PR95] and [KM99] is applicable. X_1 and X_2 are in the same c-component $S = \{X_1, X_2, Y\}$. By Lemma 15, $P_{x_1x_2}(v)$ is identifiable if $P_{x_1x_2}(y)$ is identifiable in G_S (Figure 5.18(b)). The latter is obviously true, hence we conclude that $P_{x_1x_2}(w, y)$ is identifiable. (Formally, let $S' = An(\{Y\})_{G_S} = \{X_2, Y\}$; by Lemma 17, $P_{x_1x_2}(y)$ is identifiable in G_S if $P_{x_2}(y)$ is identifiable in the subgraph $G_{S'}$ (Figure 5.17(c)), which is obvious.)

5.4.4 Identification of direct effects $P_{pa_y}(y)$

Let Y be a single variable and let $V_Y = V \setminus \{Y\}$ be the set of all other variables. A special case of the identifiability problem is to identify the *direct effect* $P_{v_y}(y)$. We have

$$P_{v_y}(y) = P_{pa_y}(y) = Q[\{Y\}]. \quad (5.160)$$

Let Y be in the c-component S^Y . In general, the identifiability of $P_{pa_y}(y)$ can be determined by using the function $\text{Identify}(\{Y\}, S^Y, Q[S^Y])$ in Figure 5.9. In this section we give some graphical criteria for determining whether $P_{pa_y}(y)$ is identifiable.

Theorem 20 *If Y is not connected to bidirected links, then $P_{pa_y}(y)$ is identifiable, and is given by*

$$P_{pa_y}(y) = P(y|pa_y). \quad (5.161)$$

Theorem 20 is obvious. The use of Theorem 20 can be shown by identifying the direct effect on Y in Figure 5.15. Theorem 20 says that $P_{x_2, z'_1}(y)$ is identifiable and is equal to $P(y|x_2, z'_1)$.

Theorem 21 *Let Y be in the c-component S^Y . If there is no bidirected path connecting Y and any of its parents (i.e., Y is not in the same c-components with any of its parents), then $P_{pa_y}(y)$ is identifiable, and is given by*

$$P_{pa_y}(y) = \sum_{S^Y \setminus \{Y\}} Q[S^Y]. \quad (5.162)$$

Proof: Since none of the variables in $S^Y \setminus \{Y\}$ is an ancestor of Y in the subgraph G_{S^Y} , by Lemma 10, $Q[\{Y\}] = \sum_{S^Y \setminus \{Y\}} Q[S^Y]$. \square

We demonstrate the use of Theorem 21 by identifying the direct effect on Y in Figure 5.16. Y is in the c-component $S = \{X_1, X_2, Y\}$, and $Q[S]$ is given in Eq. (5.149). By Theorem 21, $P_{z_1, z_2}(y)$ is identifiable and is given by

$$P_{z_1, z_2}(y) = \sum_{x_1, x_2} P(y|x_1, x_2, z_1, z_2)P(x_2|x_1, z_1)P(x_1). \quad (5.163)$$

Lemma 18 *The direct effect on Y is identifiable if and only if the direct effect on Y is identifiable in $G_{An(\{Y\})}$.*

Lemma 18 follows from Lemma 17.

Lemma 19 *Let Y be in the c-component S^Y . The direct effect on Y is identifiable if and only if the direct effect on Y is identifiable in G_{S^Y} .*

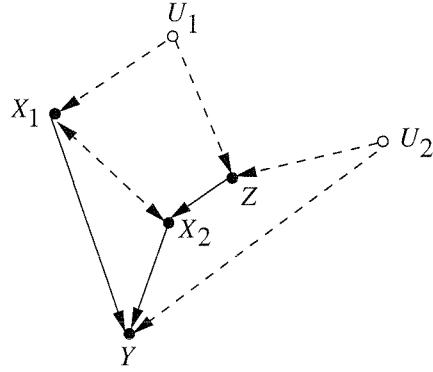


Figure 5.19: A graph in which the direct effect on Y is unidentified.

Proof: By Lemma 12, $Q[\{Y\}]$ is computable from $Q[S^Y]$ if and only if $Q[\{Y\}]_{G_{S^Y}}$ is computable from $Q[S^Y]_{G_{S^Y}}$. \square

Lemma 18 and 19 can be applied alternatively to remove nodes from a graph, until it is clear that the direct effect on Y is identifiable or until neither lemmas is applicable. This leads to the following criterion.

Theorem 22 *The direct effect on Y is identifiable if there exists no subgraph G_S of G satisfying all of the following: (i) $Y \in S$; (ii) G_S has only one c-component, S itself; (iii) All variables in S are ancestors of Y in G_S .*

The graph in Figure 5.19 satisfies conditions (i)-(iii), and for general graphs of such a type, we are unable to determine the identifiability of the direct effect on Y .

5.5 Identification of $P_t(s|c)$

Let $T, S, C \subseteq V$. In this section, we study the problem of identifying $P_t(s|c)$. This problem is important for the identifiability of conditional plans, where action T is taken in response to observation C [Pea00, chapter 4].

We have

$$P_t(s|c) = \frac{P_t(s, c)}{P_t(c)}. \quad (5.164)$$

Therefore, $P_t(s|c)$ can be identified by identifying $P_t(s, c)$ and $P_t(c)$ using the method in Section 5.4. $P_t(s|c)$ is identifiable if $P_t(s, c)$ is identifiable. $P_t(s|c)$ is

not identifiable if $P_t(s, c)$ is not identifiable but $P_t(c)$ is. If neither $P_t(s, c)$ nor $P_t(c)$ is identifiable, $P_t(s|c)$ may still be identifiable if the non-identifiable terms are canceled out in the expressions for $P_t(s, c)$ and $P_t(c)$ computed as shown in Section 5.4.1. Next, we study conditions for this canceling out to happen.

First we compute an expression for $P_t(s, c)$ with the procedure shown in Section 5.4.1. Assume that V is partitioned into c-components S_1, \dots, S_k . Let $D = An(S \cup C)_{G_{V \setminus T}}$, $F = D \setminus (S \cup C)$, and let $D_i = D \cap S_i, i = 1, \dots, k$. Assume that each subgraph G_{D_i} is partitioned into c-components D_{i1}, \dots, D_{ik_i} . Then we have (see Eq. (5.130))

$$P_t(s, c) = \sum_F \prod_{i,j} Q[D_{ij}]. \quad (5.165)$$

The identifiability of $Q[D_{ij}]$'s can be determined by calling the function $\text{Identify}(D_{ij}, S_i, Q[S_i])$ given in Figure 5.9. Let D_{ij} 's be put into two sets: in H^i if $Q[D_{ij}]$ is identified and in H^n if not identified (via the function $\text{Identify}(., ., .)$). Eq. (5.165) can be rewritten as

$$P_t(s, c) = \sum_F (\prod_{D_{ij} \in H^n} Q[D_{ij}]) (\prod_{D_{ij} \in H^i} Q[D_{ij}]) \quad (5.166)$$

This summation over F can sometimes be decomposed into a product of summations as

$$P_t(s, c) = (\sum_{F_0} \prod_{D_{ij} \in H^n \cup H^0} Q[D_{ij}]) (\sum_{F_1} \prod_{D_{ij} \in H^1} Q[D_{ij}]), \quad (5.167)$$

where F is partitioned into two sets F_0 and F_1 , and H^i is partitioned into two sets H^0 and H^1 . This partition of F and H^i can be determined as follows, using the fact that each $Q[D_{ij}]$ is a function of $Pa(D_{ij})$.

1. Let $F_0 = F \cap \cup_{D_{ij} \in H^n} Pa(D_{ij})$, $F_1 = F \setminus F_0$, and $H^1 = H^i$.
2. For each $D_{ij} \in H^1$, if $Pa(D_{ij}) \cap F_0 \neq \emptyset$, then remove D_{ij} from H^1 and put it into H^0 .
3. Let $G = F_1 \cap \cup_{D_{ij} \in H^0} Pa(D_{ij})$. If G is not empty, remove variables in G from F_1 and put them into F_0 . Then go back to step 2. If G is empty, then stop, and the partition process is finished.

Now since $P_t(c) = \sum_s P_t(s, c)$, if none of the variables in S appears in the terms in $\prod_{D_{ij} \in H^n \cup H^0} Q[D_{ij}]$, that is, if $S \cap \cup_{D_{ij} \in H^n \cup H^0} Pa(D_{ij}) = \emptyset$, then

$$P_t(c) = (\sum_{F_0} \prod_{D_{ij} \in H^n \cup H^0} Q[D_{ij}]) \sum_S (\sum_{F_1} \prod_{D_{ij} \in H^1} Q[D_{ij}]), \quad (5.168)$$

and $P_t(s|c)$ is identifiable as

$$P_t(s|c) = \frac{P_t(s, c)}{P_t(c)} = \frac{\sum_{F_1} \prod_{D_{ij} \in H^1} Q[D_{ij}]}{\sum_S (\sum_{F_1} \prod_{D_{ij} \in H^1} Q[D_{ij}])}. \quad (5.169)$$

In summary, an algorithm for computing $P_t(s|c)$ is given in Figure 5.20. The procedure consists of four basic phases. In phase-1, we compute the expressions for all c-factors and find (graphically) the sets D_{ij} and F from the graph G . In phase-2, we attempt to compute $Q[D_{ij}]$'s by calling the function $\text{Identify}(D_{ij}, S_i, Q[S_i])$, and put D_{ij} 's into two sets: H^i if identifiable and H^n if not. In phase-3, we partition F into two sets and H^i into two sets. In phase-4, when certain conditions are met, we output the expression for $P_t(s|c)$ given in Eq. (5.169).

5.6 Beyond Semi-Markovian Models

In Sections 5.2-5.5 we have studied the identifiability problem in semi-Markovian models. Our method is based on the decomposition of $P(v)$ into a product of c-factors and Lemmas 7, 10, and 11. Chapter 4 has shown that, in a Markovian model with arbitrary sets of unobserved variables, $P(v)$ can still be decomposed into a product of c-factors and that properties as given in Lemmas 7, 10, and 11 hold as well (see Corollary 1, Lemma 2, and Lemma 3). Therefore, we can use the same method developed in Sections 5.2-5.5 to identify causal effects in a Markovian model with arbitrary sets of unobserved variables. In fact, instead of working directly with a complicated model with arbitrary unobserved variables, we may work with its semi-Markovian projection defined in Section 4.5. It is shown in Section 4.5 that G and its projection $PJ(G, V)$ have the same topological relations over V and the same partition of V into c-components. Based on these results, we conclude that if $P_t(s)$ is identified in $PJ(G, V)$ (using the methods in Sections 5.2-5.5), then it is identified in G with the same expression.

In summary, to identify a causal effect $P_t(s)$ in a model with arbitrary unobserved variables, we first construct the projection graph $PJ(G, V)$, then attempt to compute $P_t(s)$ in $PJ(G, V)$. If $P_t(s)$ is computable in $PJ(G, V)$, then $P_t(s)$ is identifiable in G with the same expression.

5.7 Conclusion

We developed a new method for inferring causal effects based on the concept of c-component. Using the method, we established some powerful graphical criteria

Algorithm 6 (Computing $P_t(s|c)$)

INPUT: three disjoint sets $T, S, C \subset V$.

OUTPUT: the expression for $P_t(s|c)$ or fail to determine.

Phase-1:

1. Find the c -components of G : S_1, \dots, S_k .
2. Compute the c -factors $Q[S_1], \dots, Q[S_k]$ by Lemma 7.
3. Let $D = An(S \cup C)_{G_{V \setminus T}}$, $F = D \setminus (S \cup C)$, $D_i = D \cap S_i$, $i = 1, \dots, k$.
4. Let the c -components of G_{D_i} be D_{ij} , $j = 1, \dots, k_i$, $i = 1, \dots, k$.

Phase-2:

1. For each set D_{ij} :
Call the function $Identify(D_{ij}, S_i, Q[S_i])$ in Figure 5.9. If the function returns FAIL, then put D_{ij} into the set H^n , otherwise put D_{ij} into the set H^i .
2. If H^n is empty, then stop and output

$$P_t(s|c) = \frac{\sum_F \prod_{i,j} Q[D_{ij}]}{\sum_S \sum_F \prod_{i,j} Q[D_{ij}]}$$

Phase-3:

1. Let $F_0 = F \cap \bigcup_{D_{ij} \in H^n} Pa(D_{ij})$, $F_1 = F \setminus F_0$, and $H^1 = H^i$.
2. For each $D_{ij} \in H^1$: if $Pa(D_{ij}) \cap F_0 \neq \emptyset$, then remove D_{ij} from H^1 and put it into H^0 .
3. Let $G = F_1 \cap \bigcup_{D_{ij} \in H^0} Pa(D_{ij})$.
If $G \neq \emptyset$, then remove variables in G from F_1 and put them into F_0 . Go back to step 2.
If $G = \emptyset$, go to Phase-4.

Phase-4:

If $S \cap \bigcup_{D_{ij} \in H^n \cup H^0} Pa(D_{ij}) = \emptyset$, then output the expression for $P_t(s|c)$ as given in Eq. (5.169), otherwise output FAIL.

Figure 5.20: An algorithm for computing $P_t(s|c)$

for ensuring the identifiability of causal effects and developed procedures that systematically identifies causal effects.

CHAPTER 6

Identification of Causal Effects in Linear Models

In Chapter 5, we studied the identification problem in nonparametric models, that is, we did not make any assumptions about the functional forms of how the variables interact with each other. In this chapter, we study the identification problem in linear models, in which we assume that all interactions among variables are linear. We will show how the identifiability results in nonparametric models can be used to identify causal effects in linear models.

6.1 Linear Models

A linear recursive model over a set of variables $V = \{V_1, \dots, V_n\}$ is given by a set of linear equations

$$V_i = \sum_{j < i} c_{ij} V_j + \epsilon_i, \quad i = 1, \dots, n, \tag{6.1}$$

where c_{ij} is called a *path coefficient*, and ϵ_i represents an “error” term and is assumed to have normal distribution. Without loss of generality, we assume that the model is standardized as

$$E[V_i] = E[\epsilon_i] = 0, \quad i = 1, \dots, n, \tag{6.2}$$

where $E[.]$ represents the expectation.

A linear model can be represented by a DAG G with bidirected links, called a *causal graph*, as follows. There is a direct link from V_j to V_i in G if the coefficient of V_j in the equation for V_i is not zero ($c_{ij} \neq 0$). There is a bidirected link between V_i and V_j if the error terms ϵ_i and ϵ_j have non-zero correlation. Figure 6.1 shows a simple linear model and the corresponding causal graph in which each link is annotated by the corresponding path coefficient.

In linear models, the observed distribution $P(v)$ is fully specified by a covariance matrix Σ over V . The identification problem is that whether a path coefficient c_{ij} is computable from the covariance Σ given the causal graph. The problem has been under study for half a century. Some existing methods include the rank and order conditions [Fis66], the instrumental variable method [BT84],

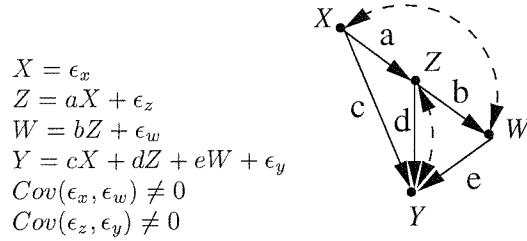


Figure 6.1: A linear model.

and graphical methods [McD97, Pea00, BP02]. In the next section, we show how causal effects ($P_t(s)$) are related to path coefficients, and thus provide a tool for the identification problem, extending the results in [Pea00].

6.2 Causal Effects

In linear models, we define three types of causal effects as follows. The path coefficient c_{ij} quantifies the direct causal effect of V_j on V_i , and is called a *direct effect*. Assume that there is a directed path p from V_k to V_i in the causal graph G , then the product of path coefficients along the path p is called the *partial effect* of V_k on V_i along the path p , and is denoted by $PE(p)$. Let $\Gamma(V_k, V_i)$ be the set of directed paths from V_k to V_i , and let $\gamma \subseteq \Gamma(V_k, V_i)$. Then $\sum_{p \in \gamma} PE(p)$ is called the partial effect of V_k on V_i along the set of paths γ and is denoted by $PE(\gamma)$. In particular, $PE(\Gamma(V_k, V_i))$ is called the *total effect* of V_k on V_i and is denoted by $TE(V_k, V_i)$.

The direct effects, partial effects, and total effects as defined above can be computed from appropriate causal effects $P_t(s)$ by computing expectations. Let $E[.]|do(t)]$ denote the expectations in the post-intervention distribution $P_t(.)$. The following proposition is obvious.

Proposition 2 (Total Effects) *The total effect of V_k on V_i can be computed as*

$$TE(V_k, V_i) = E[V_i|do(v_k)]/v_k. \quad (6.3)$$

Let $Pa_i = \{V_{i_1}, \dots, V_{i_l}\}$ be the set of parents of V_i , we have

$$E[V_i|do(pa_i)] = \sum_j c_{ii_j} v_{i_j}, \quad (6.4)$$

from which we have the following proposition.

Proposition 3 (Direct Effects) *The direct effect of V_k on V_i can be computed as*

$$c_{ik} = \frac{\partial}{\partial v_k} E[V_i | do(pa_i)], \quad V_k \in Pa_i. \quad (6.5)$$

Let $S = \{V_{i_1}, \dots, V_{i_m}\}$ be a set of variables that does not contain V_i . Let γ_j be the set of directed paths from V_{i_j} to V_i that does not pass any variables in $S \setminus \{V_{i_j}\}$. Then we have

$$E[V_i | do(s)] = \sum_j PE(\gamma_j) v_{i_j}, \quad (6.6)$$

where we define $PE(\emptyset) = 0$. Eq. (6.6) leads to the following proposition.

Proposition 4 (Partial Effects) *Given a set of directed paths $\gamma \subseteq \Gamma(V_k, V_i)$, assuming that there exists a set of variables S that does not contain variables lying in the paths in γ but contains a variable lying in each path in $\Gamma(V_k, V_i) \setminus \gamma$, the partial effect $PE(\gamma)$ can be computed as*

$$PE(\gamma) = \frac{\partial}{\partial v_k} E[V_i | do(s), do(v_k)]. \quad (6.7)$$

Note that such a set S may not exist for some γ .

6.3 Identifying Causal Effects

Next, we show how to compute those expectations with respect to post-intervention distributions given causal effects expressed in terms of the observed joint $P(v)$. For two variables V_i and V_j , and a set of variables S , the coefficient of V_j in the linear regression of V_i on V_j and S is called a *partial regression coefficient*, and is denoted by $\beta_{V_i V_j . S}$ (Note that the order of the subscripts in $\beta_{V_i V_j . S}$ is important). Partial regression coefficients can be expressed in terms of covariance matrices as follows:

$$\beta_{V_i V_j . S} = \frac{\sigma_{V_i V_j} - C_{V_i S}^T C_{S S}^{-1} C_{V_j S}}{\sigma_{V_j V_j} - C_{V_j S}^T C_{S S}^{-1} C_{V_j S}}, \quad (6.8)$$

where $C_{S S}$ etc. represents covariance matrices. Let $S = \{V_{i_1}, \dots, V_{i_m}\}$ and $S_j = S \setminus \{V_{i_j}\}$. We have the following formula for conditional expectations

$$E[V_i | s] = \sum_j \beta_{V_i V_{i_j} . S_j} v_{i_j}. \quad (6.9)$$

Eq. (6.9) provides the foundation for computing expectations in post-intervention distributions expressed in terms of $P(v)$. Whenever a causal effect $P_t(s)$ is determined as identifiable (in a nonparametric model), we can use Eqs. (6.3)–(6.7) to compute the causal effects in the corresponding linear model.

Next we study some examples. The “back-door” criterion [Pea93] says that if a set of variables Z satisfies the back-door criterion relative to (X, Y) , then $P_x(y)$ is identifiable and is given by

$$P_x(y) = \sum_z P(y|x, z)P(z). \quad (6.10)$$

Let $Z = \{Z_1, \dots, Z_k\}$ and $Z^i = Z \setminus \{Z_i\}$. Eq. (6.10) leads to

$$\begin{aligned} E[Y|do(x)] &= \sum_z E[Y|x, z]P(z) \\ &= \sum_z (\beta_{YX.Z} x + \sum_i \beta_{YZ_i.XZ^i} z_i)P(z) \quad (\text{by Eq. (6.9)}) \\ &= \beta_{YX.Z} x \quad (E[Z_i] = 0) \end{aligned} \quad (6.11)$$

Therefore, by Proposition 2, if a set of variables Z satisfies the back-door criterion relative to (X, Y) , then the total effect of X on Y is given by $\beta_{YX.Z}$. This result is given as Theorem 5.3.2 in [Pea00, p. 152].

Consider the “front-door” criterion [Pea95a], which says that if a set of variables Z satisfies the front-door criterion relative to (X, Y) , then $P_x(y)$ is identifiable and is given by

$$P_x(y) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \quad (6.12)$$

Let $Z = \{Z_1, \dots, Z_k\}$ and $Z^i = Z \setminus \{Z_i\}$. We have

$$\begin{aligned} E[Y|do(x)] &= \sum_z P(z|x) \sum_{x'} E[Y|x', z]P(x') \\ &= \sum_z P(z|x) \sum_{x'} (\beta_{YX.Z} x' + \sum_i \beta_{YZ_i.XZ^i} z_i)P(x') \\ &= \sum_z P(z|x) \sum_i \beta_{YZ_i.XZ^i} z_i \\ &= \sum_i \beta_{YZ_i.XZ^i} E[Z_i|x] \\ &= \sum_i \beta_{YZ_i.XZ^i} \beta_{Z_i X} x \end{aligned} \quad (6.13)$$

Therefore, if a set of variables Z satisfies the front-door criterion relative to (X, Y) , then the total effect of X on Y is given by $\sum_i \beta_{YZ_i.XZ^i} \beta_{Z_i X}$.

In general, the identifiability of $P_x(y)$ may be decided by using Theorem 17 or Algorithm 4 in Chapter 5.3. We can then identify the total effect of X on Y by computing expectations using Eq. (6.9). Next we show a few examples in which we can identify path coefficients by identifying direct effect $P_{p_{xy}}(y)$. Consider the problem of identifying direct effects on Y in Figure 5.13(a). It is shown in Chapter 5.4.3 that the direct effect $P_{x_1x_2}(y)$ is identifiable and is given in Eq. (5.144) rewritten in the following

$$P_{x_1x_2}(y) = \sum_z P(y|x_1, x_2, z)P(z|x_1). \quad (6.14)$$

We have

$$\begin{aligned} E[Y|do(x_1, x_2)] &= \sum_z (\beta_{YX_1.X_2Z} x_1 + \beta_{YX_2.X_1Z} x_2 + \beta_{YZ.X_1X_2} z)P(z|x_1) \\ &= \beta_{YX_1.X_2Z} x_1 + \beta_{YX_2.X_1Z} x_2 + \beta_{YZ.X_1X_2} \beta_{ZX_1} x_1 \\ &= (\beta_{YX_1.X_2Z} + \beta_{YZ.X_1X_2} \beta_{ZX_1})x_1 + \beta_{YX_2.X_1Z} x_2. \end{aligned} \quad (6.15)$$

Therefore, by Proposition 3, we have that the direct effects of X_1 and X_2 on Y are both identifiable and are given by

$$c_{YX_1} = \beta_{YX_1.X_2Z} + \beta_{YZ.X_1X_2} \beta_{ZX_1}, \quad (6.16)$$

$$c_{YX_2} = \beta_{YX_2.X_1Z}. \quad (6.17)$$

Consider the problem of identifying direct effects on Y in Figure 5.16. $P_{z_1,z_2}(y)$ is identifiable and is given in Eq. (5.163) rewritten in the following

$$P_{z_1,z_2}(y) = \sum_{x_1,x_2} P(y|x_1, x_2, z_1, z_2)P(x_2|x_1, z_1)P(x_1), \quad (6.18)$$

which leads to

$$\begin{aligned} E[Y|do(z_1, z_2)] &= \sum_{x_1,x_2} (\beta_{YX_1.X_2Z_1Z_2} x_1 + \beta_{YX_2.X_1Z_1Z_2} x_2 + \beta_{YZ_1.X_1X_2Z_2} z_1 \\ &\quad + \beta_{YZ_2.X_1X_2Z_1} z_2)P(x_2|x_1, z_1)P(x_1) \\ &= \beta_{YX_2.X_1Z_1Z_2} \sum_{x_1} (\beta_{X_2X_1.Z_1} x_1 + \beta_{X_2Z_1.X_1} z_1)P(x_1) \\ &\quad + \beta_{YZ_1.X_1X_2Z_2} z_1 + \beta_{YZ_2.X_1X_2Z_1} z_2 \\ &= (\beta_{YZ_1.X_1X_2Z_2} + \beta_{YX_2.X_1Z_1Z_2} \beta_{X_2Z_1.X_1})z_1 + \beta_{YZ_2.X_1X_2Z_1} z_2. \end{aligned} \quad (6.19)$$

Therefore we obtain that the direct effects of Z_1 and Z_2 on Y are both identifiable and are given by

$$c_{YZ_1} = \beta_{YZ_1.X_1X_2Z_2} + \beta_{YX_2.X_1Z_1Z_2}\beta_{X_2Z_1.X_1}, \quad (6.20)$$

$$c_{YZ_2} = \beta_{YZ_2.X_1X_2Z_1}. \quad (6.21)$$

This method of translating identifiability results in nonparametric models to linear models provides a new tool for the identification problem in linear models. First, the method may identify some path coefficients that can not be identified by the instrumental variable approach. Second, the method may directly identify some total effects and partial effects even though some individual path coefficients involved are not identifiable, while standard instrumental variable approach focuses on the identification of individual path coefficients.

6.4 Identifying Causal Effects Systematically

For the purpose of identifying individual path coefficients, we suggest the following systematic process. Let a topological order over V be $V_1 < \dots < V_n$, and let $V^{(j)} = \{V_1, \dots, V_j\}$, $j = 1, \dots, n$. For j from 2 to n , at each step, we consider the subgraph $G_{V^{(j)}}$ and try to identify the path coefficients associated with links pointing at V_j . At step j , the causal effects involving V_j can be computed as follows. Assuming that V_j is in the c-component S_j of $G_{V^{(j)}}$, by Lemma 7, $Q[S_j] = P_{v \setminus s_j}(s_j)$ is identifiable. Therefore, we can obtain some partial effects on V_j by computing $E[V_j|do(v \setminus s_j)]$. We may get further information about causal effects on V_j by looking for subset S of S_j that contains V_j such that $Q[S]$ is identifiable. The maximum information is achieved by finding the minimum subset S_{min} of S_j that contains V_j such that $Q[S_{min}] = P_{v \setminus s_{min}}(s_{min})$ is identifiable and computing $E[V_j|do(v \setminus s_{min})]$.

Such a minimum set can be found out by slightly modifying the function Identify(C, T, Q) in Figure 5.9 used to determine if, for any set $C \subset T$, $Q[C]$ is computable from $Q[T]$. The modified function Identify_Min(S, T, Q) is given in Figure 6.2, which, given $Q[T]$ and a set $S \subset T$, finds the minimum set S_{min} that contains S such that $Q[S_{min}]$ is computable from $Q[T]$.

Therefore, at step j , we call the function Identify_Min($\{V_j\}, S_j, Q[S_j]$) to find out S_{min} and $Q[S_{min}] = P_{v \setminus s_{min}}(s_{min})$. Then we compute $E[V_j|do(v \setminus s_{min})]$ to get some partial effects on V_j . Let $Z = \{Z_1, \dots, Z_k\}$ be the set of variables in $V \setminus S_{min}$ such that for each Z_i there exists a directed path from Z_i to V_j that does not pass any other variables in $V \setminus S_{min}$. Let γ_i be the set of directed paths from Z_i to V_j that do not pass any other variables in $V \setminus S_{min}$. By Proposition 4,

Function Identify_Min(S, T, Q)

INPUT: $S \subseteq T \subseteq V$. $Q = Q[T]$. Assuming G_T is composed of one single c-component.

OUTPUT: The minimum set $S_{min} \supseteq S$ such that $Q[S_{min}]$ is computable from $Q[T]$.

Let $A = An(S)_{G_T}$.

- IF $A = S$, output $S_{min} = S$ and $Q[S] = \sum_{T \setminus S} Q$.
- IF $A = T$, output $S_{min} = T$ and $Q[T]$.
- IF $S \subset A \subset T$
 1. Assume that, in G_A , S is contained in a c-component T' .
 2. Compute $Q[T']$ from $Q[A] = \sum_{T \setminus A} Q$ by Lemma 11.
 3. Output $\text{Identify_Min}(S, T', Q[T'])$.

Figure 6.2: A function finding the minimum set $S_{min} \supseteq S$ such that $Q[S_{min}]$ is identifiable from $Q[T]$.

the partial effect $PE(\gamma_i)$ is identifiable and is given by

$$PE(\gamma_i) = \frac{\partial}{\partial z_i} E[V_j | do(v \setminus s_{min})]. \quad (6.22)$$

Let the set of parents of V_j be $Pa_j = \{Y_1, \dots, Y_l\}$. Then the partial effect $PE(\gamma_i)$ as a summation of products of path coefficients along some paths from Z_i to V_j can be decomposed into

$$PE(\gamma_i) = \sum_{m, Y_m \in S_{min}} PE(\delta_{im}) c_{V_j Y_m}, \quad \text{for } Z_i \notin Pa_j, \quad (6.23)$$

or when Z_i is a parent of V_j ,

$$PE(\gamma_i) = c_{V_j Y_i} + \sum_{m, Y_m \in S_{min}} PE(\delta_{im}) c_{V_j Y_m}, \quad \text{for } Z_i = Y_i, \quad (6.24)$$

where δ_{im} is the set of directed paths from Z_i to Y_m that do not pass any other variables in $V \setminus S_{min}$. The summation is for $Y_m \in S_{min}$ because γ_i only contains paths that do not pass variables in $V \setminus S_{min}$. Since $P_{v \setminus s_{min}}(s_{min})$ is identifiable, by Proposition 4, the partial effect $PE(\delta_{im})$ is identifiable and is given by

$$PE(\delta_{im}) = \frac{\partial}{\partial z_i} E[Y_m | do(v \setminus s_{min})], \quad \text{for } Y_m \in S_{min}, \quad (6.25)$$

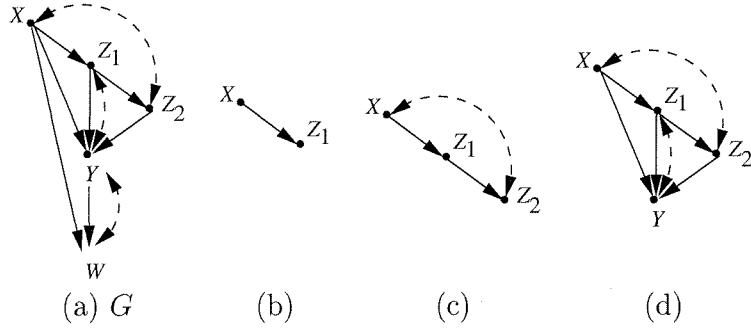


Figure 6.3: Subgraphs for identifying path coefficients in G .

which would have been identified before the step j . From Eqs. (6.22)–(6.25), we conclude that, at step j , we will obtain a set of equations which are linear in the set of path coefficients $c_{V_j Y_m}$ associated with links pointing at V_j and in which those path coefficients are the only unknowns.

In summary, at step j , we do the following

1. Find the c-component S_j of $G_{V^{(j)}}$.
2. Find the expression for $Q[S_j]$ by Lemma 7.
3. Call the function $\text{Identify_Min}(\{V_j\}, S_j, Q[S_j])$ to find out S_{min} and $Q[S_{min}] = P_{v \setminus s_{min}}(s_{min})$.
4. Compute $E[V_j | do(v \setminus s_{min})]$ to get a set of equations linear in path coefficients associated with links pointing at V_j .
5. Try to solve the set of linear equations.

Next, we demonstrate this procedure by some examples. Consider the identification problem in Figure 6.3(a). The only admissible order of variables is $X < Z_1 < Z_2 < Y < W$. At step 1, we consider the subgraph in Figure 6.3(b). It is obvious that $P_x(z_1) = P(z_1|x)$, and we obtain

$$c_{Z_1 X} = E[Z_1 | do(x)]/x = \beta_{Z_1 X}. \quad (6.26)$$

At step 2, we consider the subgraph in Figure 6.3(c). Z_2 is in the c-component $\{X, Z_2\}$ and Lemma 7 gives

$$Q[\{X, Z_2\}] = P(z_2|z_1, x)P(x). \quad (6.27)$$

Calling the function `Identify_Min`($\{Z_2\}$, $\{X, Z_2\}$, $Q[\{X, Z_2\}]$), we obtain

$$Q[\{Z_2\}] = P_{z_1}(z_2) = \sum_x P(z_2|z_1, x)P(x). \quad (6.28)$$

Therefore,

$$\begin{aligned} c_{Z_2 Z_1} &= E[Z_2|do(z_1)]/z_1 \\ &= \sum_x (\beta_{Z_2 Z_1.X} z_1 + \beta_{Z_2 X.Z_1} x)P(x)/z_1 \\ &= \beta_{Z_2 Z_1.X}. \end{aligned} \quad (6.29)$$

At step 3, we consider the subgraph in Figure 6.3(d). Y is in the c-component $\{Y, Z_1\}$ and Lemma 7 gives

$$Q[\{Y, Z_1\}] = P_{xz_2}(y, z_1) = P(y|z_2, z_1, x)P(z_1|x). \quad (6.30)$$

Calling the function `Identify_Min`($\{Y\}$, $\{Y, Z_1\}$, $Q[\{Y, Z_1\}]$) returns $\{Y, Z_1\}$ as the minimum set ($Q[\{Y\}]$ is not identifiable). We then compute the expectation

$$\begin{aligned} E[Y|do(x, z_2)] &= \sum_{z_1} (\beta_{YZ_2.Z_1X} z_2 + \beta_{YZ_1.Z_2X} z_1 + \beta_{YX.Z_2Z_1} x)P(z_1|x) \\ &= \beta_{YZ_2.Z_1X} z_2 + (\beta_{YZ_1.Z_2X}\beta_{Z_1X} + \beta_{YX.Z_2Z_1})x \end{aligned} \quad (6.31)$$

Therefore, we obtain the path coefficient

$$c_{YZ_2} = \beta_{YZ_2.Z_1X}, \quad (6.32)$$

and the following partial effect

$$c_{YX} + c_{Z_1X}c_{YZ_1} = \beta_{YX.Z_2Z_1} + \beta_{Z_1X}\beta_{YZ_1.Z_2X}, \quad (6.33)$$

where c_{Z_1X} is identified in Eq. (6.26).

Finally, at the last step, we consider the graph in Figure 6.3(a). W is in the c-component $\{W, Y, Z_1\}$ and Lemma 7 gives

$$Q[\{W, Y, Z_1\}] = P_{xz_2}(w, y, z_1) = P(w|y, z_2, z_1, x)P(y|z_2, z_1, x)P(z_1|x). \quad (6.34)$$

Calling the function `Identify_Min`($\{W\}$, $\{W, Y, Z_1\}$, $Q[\{W, Y, Z_1\}]$) returns

$\{W, Y, Z_1\}$ as the minimum set. We then compute the expectation

$$\begin{aligned}
& E[W|do(x, z_2)] \\
&= \sum_{y, z_1} (\beta_{WY.Z_2Z_1X} y + \beta_{WZ_2.YZ_1X} z_2 + \beta_{WZ_1.YZ_2X} z_1 + \beta_{WX.YZ_2Z_1} x) \\
&\quad P(y|z_2, z_1, x)P(z_1|x) \\
&= \beta_{WY.Z_2Z_1X} E[Y|do(x, z_2)] + \beta_{WZ_2.YZ_1X} z_2 + \beta_{WZ_1.YZ_2X} \beta_{Z_1X} x \\
&\quad + \beta_{WX.YZ_2Z_1} x \\
&= (\beta_{WY.Z_2Z_1X} \beta_{YZ_2.Z_1X} + \beta_{WZ_2.YZ_1X}) z_2 \\
&\quad + [\beta_{WY.Z_2Z_1X} (\beta_{YZ_1.Z_2X} \beta_{Z_1X} + \beta_{YX.Z_2Z_1}) \\
&\quad + \beta_{WZ_1.YZ_2X} \beta_{Z_1X} + \beta_{WX.YZ_2Z_1}] x, \quad (\text{substitute (6.31) in}) \quad (6.35)
\end{aligned}$$

from which we obtain the total effect of Z_2 on W :

$$c_{YZ_2} c_{WY} = \beta_{WY.Z_2Z_1X} \beta_{YZ_2.Z_1X} + \beta_{WZ_2.YZ_1X}, \quad (6.36)$$

and the following partial effect of X on W :

$$\begin{aligned}
& c_{WX} + (c_{YX} + c_{Z_1X} c_{YZ_1}) c_{WY} \\
&= \beta_{WY.Z_2Z_1X} (\beta_{YZ_1.Z_2X} \beta_{Z_1X} + \beta_{YX.Z_2Z_1}) + \beta_{WZ_1.YZ_2X} \beta_{Z_1X} + \beta_{WX.YZ_2Z_1}. \quad (6.37)
\end{aligned}$$

The path coefficient c_{YZ_2} is identifiable and is given in (6.32), and therefore by Eq. (6.36), the path coefficient c_{WY} is identifiable and is given by

$$c_{WY} = \beta_{WY.Z_2Z_1X} + \frac{\beta_{WZ_2.YZ_1X}}{\beta_{YZ_2.Z_1X}}. \quad (6.38)$$

Then from Eqs. (6.37), (6.33), and (6.38), the path coefficient c_{WX} is identifiable and is given by

$$c_{WX} = \beta_{WX.YZ_2Z_1} + \beta_{WZ_1.YZ_2X} \beta_{Z_1X} - \frac{\beta_{WZ_2.YZ_1X}}{\beta_{YZ_2.Z_1X}} (\beta_{YZ_1.Z_2X} \beta_{Z_1X} + \beta_{YX.Z_2Z_1}) \quad (6.39)$$

6.5 Conclusion

We show how the identifiability results in nonparametric models presented in Chapter 5 can be used to identify causal effects in linear models. The method may directly identify some total effects and partial effects even though some individual path coefficients involved are not identifiable. The method is useful in models with few bidirected (confounding) links.

CHAPTER 7

Probabilities of Causation: Bounds and Identification

7.1 Introduction

Assessing the likelihood that one event *was the cause* of another guides much of what we understand about (and how we act in) the world. For example, few of us would take aspirin to combat headache if it were not for our conviction that, with high probability, it was aspirin that “actually caused” relief in previous headache episodes. Likewise, according to common judicial standard, judgment in favor of plaintiff should be made if and only if it is “more probable than not” that the defendant’s action was a *cause* for the plaintiff’s injury (or death). This chapter deals with the question of estimating the probability of causation from statistical data.

Causation has two faces, *necessary* and *sufficient*. The most common conception of causation – that the effect E would not have occurred in the absence of the cause C – captures the notion of “necessary causation”. Competing notions such as “sufficient cause” and “necessary-and-sufficient cause” are also of interest in a number of applications, and this chapter analyzes the relationships among the three notions. Although the distinction between necessary and sufficient causes goes back to J.S. Mill [Mil43], it has received semi-formal explications only in the 1960s – via conditional probabilities [Goo61] and logical implications [Mac65]. These explications suffer from basic semantical difficulties [Kim71] [Pea00, pp. 249–256, 313–316], and they do not yield effective procedures for computing probabilities of causes. This chapter defines probabilities of causes in a language of counterfactuals that is based on a simple model-theoretic semantics (to be formulated in Section 7.2).

[RG89] gave a counterfactual definition for the probability of necessary causation taking counterfactuals as primitives, and assuming that one is in possession of a consistent joint probability function on both ordinary and counterfactual events. [Pea99] gave definitions for the probabilities of necessary or sufficient causation (or both) based on structural model semantics, which defines counterfactuals as quantities derived from modifiable sets of functions [GP97, GP98, Hal98, Pea00].

The structural models semantics, as we shall see in Section 7.2, leads to effective procedures for computing probabilities of counterfactual expressions from a given causal theory [BP94, BP95]. Additionally, this semantics can be characterized by a complete set of axioms [GP98, Hal98], which we will use as inference rules in our analysis.

The central aim of this chapter is to estimate probabilities of causation from frequency data, as obtained in experimental and observational statistical studies. In general, such probabilities are *non-identifiable*, that is, non-estimable from frequency data alone. One factor that hinders identifiability is confounding – the cause and the effect may both be influenced by a third factor. Moreover, even in the absence of confounding, probabilities of causation are sensitive to the data-generating process, namely, the functional relationships that connect causes and effects [RG89, BP94]. Nonetheless, useful information in the form of *bounds* on the probabilities of causation can be extracted from empirical data without actually knowing the data-generating process. These bounds improve when data from observational and experimental studies are combined. Additionally, under certain assumptions about the data-generating process (such as exogeneity and monotonicity), the bounds may collapse to point estimates, which means that the probabilities of causation are identifiable – they can be expressed in terms of probabilities of observed quantities. These estimates will be recognized as familiar expressions that often appear in the literature as measures of *attribution*. Our analysis thus explicates the assumptions about the data-generating process that must be ascertained before those measures can legitimately be interpreted as probabilities of causation.

The analysis of this chapter leans heavily on results reported in [Pea99] [Pea00, pp. 283–308]. Pearl derived bounds and identification conditions under certain assumptions of exogeneity and monotonicity, and this chapter improves on Pearl’s results by narrowing his bounds and weakening his assumptions. In particular, we show that for most of Pearl’s results, the assumption of strong exogeneity can be replaced by weak exogeneity (to be defined in Section 7.4.3). Additionally, we show that the point estimates that Pearl obtained under the assumption of monotonicity (Definition 19) constitute valid lower bounds when monotonicity is not assumed. Finally, we prove that the bounds derived by Pearl, as well as those provided in this chapter are *sharp*, that is, they cannot be improved without strengthening the assumptions.

The rest of the chapter is organized as follows. Section 7.2 reviews the structural model semantics of actions, counterfactuals and probability of counterfactuals. In Section 7.3 we present formal definitions for the probabilities of causation and briefly discuss their applicability in epidemiology, artificial intelligence, and legal reasoning. In Section 7.4 we systematically investigate the maximal infor-

mation (about the probabilities of causation) that can be obtained under various assumptions and from various types of data. Section 7.5 illustrates, by example, how the results presented in this chapter can be applied to resolve issues of attribution in legal settings. Section 7.6 illustrates the use of our results in personal decision making. Section 7.7 concludes the chapter.

7.2 Structural Model Semantics

In Chapters 1–5, we assumed probabilistic relations between variables in the model. In this chapter, we assume deterministic, functional relations between variables, and the causal model will be called *functional*, which, in addition to interventions, supports counterfactual readings. This section presents a brief summary of the structural-equation semantics of counterfactuals as defined in [BP95, GP97, GP98, Hal98]. Related approaches have been proposed in [SR66] (see footnote 5) and [Rob86]. For detailed exposition of the structural account and its applications see [Pea00].

Structural models are generalizations of the structural equations used in engineering, biology, economics and social science.¹ World knowledge is represented as a collection of stable and autonomous relationships called “mechanisms,” each represented as a function, and changes due to interventions or hypothetical eventualities are treated as local modifications of these functions.

A causal model is a mathematical object that assigns truth values to sentences involving causal relationships, actions, and counterfactuals. We will first define functional causal models, then discuss how causal sentences are evaluated in such models. We will restrict our discussion to recursive (or feedback-free) models; extensions to non-recursive models can be found in [GP97, GP98, Hal98].

Definition 6 (*functional causal model*)

A functional causal model is a triple

$$M = \langle U, V, F \rangle$$

where

- (i) *U is a set of variables, called exogenous. (These variables will represent background conditions, that is, variables whose values are determined outside the model.)*

¹Similar models, called “neuron diagrams” [Lew86, Hal02] are used informally by philosophers to illustrate chains of causal processes.

- (ii) V is an ordered set $\{V_1, V_2, \dots, V_n\}$ of variables, called endogenous. (These represent variables that are determined in the model, namely, by variables in $U \cup V$.)
- (iii) F is a set of functions $\{f_1, f_2, \dots, f_n\}$ where each f_i is a mapping from $U \times (V_1 \times \dots \times V_{i-1})$ to V_i . In other words, each f_i tells us the value of V_i given the values of U and all predecessors of V_i . Symbolically, the set of equations F can be represented by writing ²

$$v_i = f_i(pa_i, u_i) \quad i = 1, \dots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in V (connoting parents) sufficient for representing f_i ³. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that is sufficient for representing f_i .

Every functional causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i (by convention, the exogenous variables are usually not shown explicitly in the graph). We call such a graph the *causal graph* associated with M . This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i .

Basic of our analysis are sentences involving actions or external interventions, such as, “ p will be true if we do q ” where q is any elementary proposition. To evaluate such sentences we need the notion of “submodel.”

Definition 7 (Submodel)

Let M be a functional causal model, X be a set of variables in V , and x be a particular assignment of values to the variables in X . A submodel M_x of M is the functional causal model

$$M_x = \langle U, V, F_x \rangle$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \quad (7.1)$$

² We use capital letters (e.g., X, Y) as names of variables and sets of variables, and lower-case letters (e.g., x, y) for specific values (called realizations) of the corresponding variables.

³ A set of variables X is sufficient for representing a given function $y = f(x, z)$ if f is trivial in Z —that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Submodels represent the effect of actions and hypothetical changes, including those dictated by counterfactual antecedents. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u , then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X . The transformation from M to M_x modifies the algebraic content of F , which is the reason for the name *modifiable structural equations* used in [GP98].⁴

Definition 8 (*Effect of action*)

Let M be a functional causal model, X be a set of variables in V , and x be a particular realization of X . The effect of action $do(X = x)$ on M is given by the submodel M_x .

Definition 9 (*Potential response*)

Let Y be a variable in V , let X be a subset of V , and let u be a particular value of U . The potential response of Y to action $do(X = x)$ in situation u , denoted $Y_x(u)$, is the (unique) solution for Y of the set of equations F_x .

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form “ $do(X = x)$ if $Z = z$ ” can be formalized using the replacement of equations by functions of Z , rather than by constants [Pea94]. We will not consider disjunctive actions, of the form “ $do(X = x \text{ or } X = x')$ ”, since these complicate the probabilistic treatment of counterfactuals.

Definition 10 (*Counterfactual*)

Let Y be a variable in V , and let X be a subset of V . The counterfactual expression “The value that Y would have obtained, had X been x ” is interpreted as denoting the potential response $Y_x(u)$.

Definition 5 thus interprets the counterfactual phrase “had X been x ” in terms of a hypothetical external action that modifies the actual course of history and enforces the condition “ $X = x$ ” with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals [BP94], as it permits x to

⁴Structural modifications date back to [Mar50] and [Sim53]. An explicit translation of interventions into “wiping out” equations from the model was first proposed by [SW60] and later used in [Fis70], [Sob90], [SGS93], and [Pea95a]. A similar notion of sub-model is introduced in [Fin85], though not specifically for representing actions and counterfactuals.

differ from the actual value $X(u)$ of X without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.⁵

It can easily be shown [GP97] that the counterfactual relationship just defined, $Y_x(u)$, satisfies the following two properties:

Effectiveness:

For any two disjoint sets of variables, Y and W , we have

$$Y_{yw}(u) = y. \quad (7.2)$$

In words, setting the variables in W to w has no effect on Y , once we set the value of Y to y .

Composition:

For any two disjoint sets of variables X and W , and any set of variables Y ,

$$W_x(u) = w \implies Y_{xw}(u) = Y_x(u). \quad (7.3)$$

In words, once we set X to x , setting the variables in W to the same values, w , that they would attain (under x) should have no effect on Y . Furthermore, effectiveness and composition are *complete* whenever M is recursive (i.e., $G(M)$ is acyclic) [GP98, Hal98], that is, every property of counterfactuals that follows from the structural model semantics can be derived by repeated application of effectiveness and composition.

A corollary of composition is a property called *consistency* by [Rob87]:

$$(X(u) = x) \implies (Y_x(u) = Y(u)) \quad (7.4)$$

Consistency states that, if in a certain context u we find variable X at value x , and we intervene and set X to that same value, x , we should not expect any change in the response variable Y . This property will be used in several derivations of Section 7.3 and 7.4.

The structural formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 11 (*Probabilistic functional causal model*)

A probabilistic functional causal model is a pair

$$\langle M, P(u) \rangle$$

where M is a functional causal model and $P(u)$ is a probability function defined over the domain of U .

⁵Simon and Rescher [SR66, p. 339] did not include this step in their account of counterfactuals and noted that backward inferences triggered by the antecedents can lead to ambiguous interpretations.

$P(u)$, together with the fact that each endogenous variable is a function of U , defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \mid Y(u)=y\}} P(u) \quad (7.5)$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x . For example, the *causal effect* of x on y is defined as:

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \quad (7.6)$$

Likewise, a probabilistic functional causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W , not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u \mid Y_x(u)=y \& X(u)=x'\}} P(u) \quad (7.7)$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y \& Y_{x'}(u)=y'\}} P(u). \quad (7.8)$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement “ Y would be y if $X = x$ and Y would be y' if $X = x'$.” Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Daw97]. The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U , demonstrates that joint probabilities of counterfactuals have solid mathematical and conceptual underpinning and, moreover, these probabilities can be encoded rather parsimoniously using $P(u)$ and F .

In particular, the probabilities of causation analyzed in this chapter (see Eqs. (7.10)-(7.12)) require the evaluation of expressions of the form $P(Y_{x'} = y' | X = x, Y = y)$ with x and y incompatible with x' and y' , respectively. Eq. (7.7) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y' | X = x, Y = y) &= \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y') P(u | x, y) \end{aligned} \quad (7.9)$$

In other words, we first update $P(u)$ to obtain $P(u|x, y)$, then we use the updated distribution $P(u|x, y)$ to compute the expectation of the propositional variable $Y_{x'}(u) = y'$.⁶

7.3 Probabilities of Causation: Definitions

In this section, we present the definitions for the three aspects of causation as defined in [Pea99]. We use the counterfactual language and the structural model semantics introduced in Section 7.2. For notational simplicity, we limit the discussion to binary variables; extension to multi-valued variables are straightforward (see [Pea00], page 286, footnote 5).

Definition 12 (*Probability of necessity (PN)*)

Let X and Y be two binary variables in a functional causal model M , let x and y stand for the propositions $X = \text{true}$ and $Y = \text{true}$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} PN &\triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \\ &\triangleq P(y'_{x'}|x, y) \end{aligned} \tag{7.10}$$

In other words, PN stands for the probability that event y would not have occurred in the absence of event x , $y'_{x'}$, given that x and y did in fact occur.⁷

This quantity has applications in epidemiology, legal reasoning, and artificial intelligence (AI). Epidemiologists have long been concerned with estimating the probability that a certain case of disease is *attributable* to a particular exposure, which is normally interpreted counterfactually as “the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur.” This counterfactual notion, which Robins and Greenland (1989) called the “probability of causation”, measures how *necessary* the cause

⁶In our deterministic model, $P(Y_{x'}(u) = y')$ takes on the values zero and one, but in models involving intrinsic nondeterminism (see Section 7.7), or memoryless stochastic fluctuations, $P(Y_{x'}(u) = y')$ expresses the residual uncertainty in Y , under the setting $X = x'$, in situation $U = u$. Eq. (7.9) then captures the uncertainty associated with the effect of action $\text{do}(X = x')$, conditioned on the pre-action evidence $X = x$ and $Y = y$.

⁷Note a slight change in notation relative to that used Section 7.2. Lower case letters (e.g., x, y) denoted arbitrary values of variables in Section 7.2, and now stand for propositions (or events). Note also the abbreviations y_x for $Y_x = \text{true}$ and y'_x for $Y_x = \text{false}$. Readers accustomed to writing “ $A > B$ ” for the counterfactual “ B if it were A ” can translate Eq. (7.10) to read $PN \triangleq P(x' > y'|x, y)$.

is for the production of the effect. It is used frequently in lawsuits, where legal responsibility is at the center of contention (see Section 7.5).

Definition 13 (*Probability of sufficiency (PS)*)

$$PS \triangleq P(y_x|y', x') \quad (7.11)$$

PS measures the capacity of x to *produce* y and, since “production” implies a transition from the absence to the presence of x and y , we condition the probability $P(y_x)$ on situations where x and y are both absent. Thus, mirroring the necessity of x (as measured by PN), PS gives the probability that setting x would produce y in a situation where x and y are in fact absent.

PS finds applications in policy analysis, AI, and psychology. A policy maker may well be interested in the dangers that a certain exposure may present to the healthy population [KFG89]. Counterfactually, this notion is expressed as the “probability that a healthy unexposed individual would have gotten the disease had he/she been exposed.” In psychology, PS serves as the basis for Cheng’s (1997) causal power theory, which attempts to explain how humans judge causal strength among events. In AI, PS plays a major role in the generation of explanations [Pea00, pp. 221-223].

Definition 14 (*Probability of necessity and sufficiency (PNS)*)

$$PNS \triangleq P(y_x, y'_{x'}) \quad (7.12)$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y .

As illustrated above, PS assesses the presence of an active causal process capable of producing the effect, while PN emphasizes the absence of alternative processes, not involving the cause in question, that are capable of explaining the effect. In legal settings, where the occurrence of the cause, x , and the effect, y , are fairly well established, PN is the measure that draws most attention, and the plaintiff must prove that y would not have occurred *but for* x [Rob97a]. Still, lack of sufficiency may weaken arguments based on PN [Goo93, Mic00].

Although none of these quantities is sufficient for determining the others, they are not entirely independent, as shown in the following lemma.

Lemma 20 *The probabilities of causation satisfy the following relationship:*

$$PNS = P(x, y)PN + P(x', y')PS \quad (7.13)$$

Proof of Lemma 20

Using the consistency condition of Eq. (7.4),

$$x \Rightarrow (y_x = y), \quad x' \Rightarrow (y_{x'} = y), \quad (7.14)$$

we can write

$$\begin{aligned} y_x \wedge y'_{x'} &= (y_x \wedge y'_{x'}) \wedge (x \vee x') \\ &= (y_x \wedge x \wedge y'_{x'}) \vee (y_x \wedge y'_{x'} \wedge x') \\ &= (y \wedge x \wedge y'_{x'}) \vee (y_x \wedge y' \wedge x') \end{aligned}$$

Taking probabilities on both sides, and using the disjointness of x and x' , we obtain:

$$\begin{aligned} P(y_x, y'_{x'}) &= P(y'_{x'}, x, y) + P(y_x, x', y') \\ &= P(y'_{x'}|x, y)P(x, y) + P(y_x|x', y')P(x', y') \end{aligned}$$

which proves Lemma 20. \square

Definition 15 (Identifiability)

Let $Q(M)$ be any quantity defined on a functional causal model M . Q is identifiable in a class \mathbf{M} of models iff any two models M_1 and M_2 from \mathbf{M} that satisfy $P_{M_1}(v) = P_{M_2}(v)$ also satisfy $Q(M_1) = Q(M_2)$. In other words, Q is identifiable if it can be determined uniquely from the probability distribution $P(v)$ of the endogenous variables V .

The class \mathbf{M} that we will consider when discussing identifiability will be determined by assumptions that one is willing to make about the model under study. For example, if our assumptions consist of the structure of a causal graph G_0 , \mathbf{M} will consist of all models M for which $G(M) = G_0$. If, in addition to G_0 , we are also willing to make assumptions about the functional form of some mechanisms in M , \mathbf{M} will consist of all models M that incorporate those mechanisms, and so on.

Since all the causal measures defined above invoke conditionalization on y , and since y is presumed affected by x , the antecedent of the counterfactual y_x , we

know that none of these quantities is identifiable from knowledge of the structure $G(M)$ and the data $P(v)$ alone, even under condition of no confounding. However, useful information in the form of bounds may be derived for these quantities from $P(v)$, especially when knowledge about causal effects $P(y_x)$ and $P(y_{x'})$ are also available⁸. Moreover, under some general assumptions about the data-generating process, these quantities may even be identified.

To formulate precisely what it means to identify a counterfactual quantity from various types of data, we now generalize Definition 15 to capture the notion of “identification from experiments.” By *experiment* we mean a prescribed modification of the underlying functional causal model, together with the probability distribution that the modified model induces on the variables observed in the experiment.

Definition 16 (*Identifiability from experiments*)

Let $Q(M)$ be any quantity defined on a functional causal model M , let M^{\exp} be a modification of M induced by some experiment, \exp , and let Y be a set of variables observed under \exp . We say that Q is identifiable from experiment \exp in a class \mathbf{M} of models iff any two models M_1 and M_2 from \mathbf{M} that satisfy $P_{M_1^{\exp}}(y) = P_{M_2^{\exp}}(y)$ also satisfy $Q(M_1) = Q(M_2)$. In other words, Q is identifiable from \exp if it can be determined uniquely from the probability distribution that the observed variables Y attain under the experimental conditions created by \exp .

In the sequel, we will consider standard controlled experiments, in which the values of the control variable X are assigned at random. The outcomes of such experiments are the causal effects probabilities, $P(y_x)$ and $P(y_{x'})$, which are also induced by the submodels M_x and $M_{x'}$, respectively. However, Definition 16 is applicable to a much broader class of experimental designs, corresponding to both deletion and replacement of the model equations. Note that standard identifiability (Definition 15) is a special case of identifiability from experiments, where $Y = V$ and $M^{\exp} = M$.

7.4 Bounds and Conditions of Identification

In this section we estimate the three probabilities of causation defined in Section 7.3 when given experimental or nonexperimental data (or both) and additional assumptions about the data-generating process. We will assume that experimental data will be summarized in the form of the causal effects $P(y_x)$ and

⁸The causal effects $P(y_x)$ and $P(y_{x'})$ can be estimated reliably from controlled experimental studies, and from certain observational (i.e., nonexperimental) studies (see Chapter 5).

$P(y_{x'})$, and nonexperimental data will be summarized in the form of the joint probability function: $P_{XY} = \{P(x, y), P(x', y), P(x, y'), P(x', y')\}$.⁹

7.4.1 Linear programming formulation

In principle, in order to compute the probability of any counterfactual sentence involving variables X and Y we need to specify a functional causal model, namely, the functional relation between X and Y and the probability distribution on U . However, since every such model induces a joint probability distribution on the four binary variables: X , Y , Y_x and $Y_{x'}$, specifying the sixteen parameters of this distribution would suffice. Moreover, since Y is a deterministic function of the other three variables, the problem is fully specified by the following set of eight parameters:

$$\begin{aligned} p_{111} &= P(y_x, y_{x'}, x) = P(x, y, y_{x'}) \\ p_{110} &= P(y_x, y_{x'}, x') = P(x', y, y_x) \\ p_{101} &= P(y_x, y'_{x'}, x) = P(x, y, y'_{x'}) \\ p_{100} &= P(y_x, y'_{x'}, x') = P(x', y', y_x) \\ p_{011} &= P(y'_x, y_{x'}, x) = P(x, y', y_{x'}) \\ p_{010} &= P(y'_x, y_{x'}, x') = P(x', y, y'_x) \\ p_{001} &= P(y'_x, y'_{x'}, x) = P(x, y', y'_{x'}) \\ p_{000} &= P(y'_x, y'_{x'}, x') = P(x', y', y'_x) \end{aligned}$$

where we have used the consistency condition Eq. (7.14). These parameters are constrained by the probabilistic constraints

$$\sum_{i=0}^1 \sum_{j=0}^1 \sum_{k=0}^1 p_{ijk} = 1$$

$p_{ijk} \geq 0$ for $i, j, k \in \{0, 1\}$ (7.15)

In addition, the nonexperimental probabilities P_{XY} impose the constraints:

$$\begin{aligned} p_{111} + p_{101} &= P(x, y) \\ p_{011} + p_{001} &= P(x, y') \\ p_{110} + p_{010} &= P(x', y) \end{aligned} \quad (7.16)$$

⁹For example, if x represents a specific exposure and y represents the outcome of a specific individual I , then P_{XY} is estimated from sampled frequency counts in a population that is deemed representative of the relevant characteristics of I . The choice of an appropriate reference population is usually based on causal consideration (often suppressed), and involves matching the characteristics of I against the causal model $(M, P(u))$ judged to govern the population.

and the causal effects, $P(y_x)$ and $P(y_{x'})$, impose the constraints:

$$\begin{aligned} P(y_x) &= p_{111} + p_{110} + p_{101} + p_{100} \\ P(y_{x'}) &= p_{111} + p_{110} + p_{011} + p_{010} \end{aligned} \quad (7.17)$$

The quantities we wish to bound are:

$$PNS = p_{101} + p_{100} \quad (7.18)$$

$$PN = p_{101}/P(x, y) \quad (7.19)$$

$$PS = p_{100}/P(x', y') \quad (7.20)$$

In the following sections we obtain bounds for these quantities by solving various linear programming problems. For example, given both experimental and nonexperimental data, the lower (and upper) bounds for PNS are obtained by minimizing (or maximizing, respectively) $p_{101} + p_{100}$ subject to the constraints (7.15), (7.16) and (7.17). The bounds obtained are guaranteed to be sharp because the optimization is global.

Optimizing the functions in (7.18)–(7.20), subject to equality constraints, defines a linear programming (LP) problem that lends itself to closed-form solution. [Bal95, Appendix B] describes a computer program that takes symbolic descriptions of LP problems and returns symbolic expressions for the desired bounds. The program works by systematically enumerating the vertices of the constraint polygon of the dual problem. The bounds reported in this chapter were produced (or tested) using Balke's program, and will be stated here without proofs; their correctness can be verified by manually enumerating the vertices as described in [Bal95, Appendix B].

7.4.2 Bounds with no assumptions

7.4.2.1 Given nonexperimental data

Given P_{XY} , constraints (7.15) and (7.16) induce the following upper bound on PNS:

$$0 \leq PNS \leq P(x, y) + P(x', y'). \quad (7.21)$$

However, PN and PS are not constrained by P_{XY} .

These constraints also induce bounds on the causal effects $P(y_x)$ and $P(y_{x'})$:

$$\begin{aligned} P(x, y) &\leq P(y_x) \leq 1 - P(x, y') \\ P(x', y) &\leq P(y_{x'}) \leq 1 - P(x', y') \end{aligned} \quad (7.22)$$

7.4.2.2 Given causal effects

Given constraints (7.15) and (7.17), the bounds induced on PNS are:

$$\max[0, P(y_x) - P(y_{x'})] \leq PNS \leq \min[P(y_x), P(y'_{x'})] \quad (7.23)$$

with no constraints on PN and PS.

7.4.2.3 Given both nonexperimental data and causal effects

Given the constraints (7.15), (7.16) and (7.17), the following bounds are induced on the three probabilities of causation:

$$\max \left\{ \begin{array}{l} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{array} \right\} \leq PNS \leq \min \left\{ \begin{array}{l} P(y_x) \\ P(y'_{x'}) \\ P(x, y) + P(x', y') \\ P(y_x) - P(y_{x'}) + P(x, y') + P(x', y) \end{array} \right\} \quad (7.24)$$

$$\max \left\{ \frac{0}{P(y) - P(y_{x'})} \right\} \leq PN \leq \min \left\{ \frac{1}{P(y'_{x'}) - P(x', y')} \right\} \quad (7.25)$$

$$\max \left\{ \frac{0}{P(y_x) - P(y)} \right\} \leq PS \leq \min \left\{ \frac{1}{P(y_x) - P(x, y)} \right\} \quad (7.26)$$

Thus we see that some information about PN and PS can be extracted without making any assumptions about the data-generating process. Furthermore, combined data from both experimental and nonexperimental studies yield information that neither study alone can provide.

7.4.3 Bounds under exogeneity (no confounding)

Definition 17 (*Exogeneity*)

A variable X is said to be exogenous for Y in model M iff

$$P(y_x) = P(y|x) \quad \text{and} \quad P(y_{x'}) = P(y|x'), \quad (7.27)$$

or, equivalently,

$$Y_x \perp\!\!\!\perp X \quad \text{and} \quad Y_{x'} \perp\!\!\!\perp X. \quad (7.28)$$

In words, the way Y would potentially respond to experimental conditions x or x' is independent of the actual value of X .

Eq. (7.27) has been given a variety of (equivalent) definitions and interpretations. Epidemiologists refer to this condition as “no-confounding” [RG89], statisticians call it “as if randomized,” and [RR83] call it “weak ignorability.” A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$ (more precisely, a common ancestor that is connected to Y through a path not containing X , including latent ancestors, which represent dependencies among variables in U). The classical econometric criterion for exogeneity (e.g., [Dhr70, p. 169]) states that X be independent of the error term (u) in the equation for Y .¹⁰ We will use the term “exogeneity”, since it was under this term that the relations given in (7.27) first received their precise definition (by economists).

Combining Eq. (7.27) with the constraints of (7.15)–(7.17), the linear programming optimization (Section 7.4.1) yields the following results:

Theorem 23 *Under condition of exogeneity, the three probabilities of causation are bounded as follows:*

$$\max[0, P(y|x) - P(y|x')] \leq PNS \leq \min[P(y|x), P(y'|x')] \quad (7.29)$$

$$\frac{\max[0, P(y|x) - P(y|x')]}{P(y|x)} \leq PN \leq \frac{\min[P(y|x), P(y'|x')]}{P(y|x)} \quad (7.30)$$

$$\frac{\max[0, P(y|x) - P(y|x')]}{P(y'|x')} \leq PS \leq \frac{\min[P(y|x), P(y'|x')]}{P(y'|x')} \quad (7.31)$$

The bounds expressed in Eq. (7.30) were first derived by [RG89]; a more elaborate proof can be found in [FS99]. [Pea99] derived Eqs. (7.29)–(7.31) under a stronger condition of exogeneity (see Definition 18). We see that under the condition of no-confounding the lower bound for PN can be expressed as

$$PN \geq 1 - \frac{1}{P(y|x)/P(y|x')} \triangleq 1 - \frac{1}{RR} \quad (7.32)$$

where $RR = P(y|x)/P(y|x')$ is the *risk ratio* (also called *relative risk*) in epidemiology. Courts have often used the condition $RR > 2$ as a criterion for legal responsibility [BG94]. Eq. (7.32) shows that this practice represents a conservative interpretation of the “more probable than not” standard (assuming no confounding); PN must indeed be higher than 0.5 if RR exceeds 2. [FS99] argue that, in general, epidemiological evidence may not be applicable as proof for specific causation [FS99] because such evidence cannot account for all characteristics

¹⁰This criterion has been the subject of relentless objections by modern econometricians [EHR83, Hen95, imb97], but see [Ald93] and [Pea00, pp. 169-170; 245-247] for a reconciliatory perspective on this controversy.

specific to the plaintiff. Freedman and Stark further imply that the appropriate way of interpreting the “more probable than not” criterion would be to consider the probability of causation in a restricted subpopulation, one that shares the plaintiff characteristics. Taken to extreme, such restrictive interpretation would insist on characterizing the plaintiff to minute detail, and would reduce PN to zero or one when all relevant details are accounted for. We doubt that this interpretation underlies the intent of judicial standards. We believe that, by using the wording “more probable than not,” law makers have instructed us to ignore specific features for which data is not available, and to base our determination on the most specific features for which reliable data is available (see footnote 9).¹¹ PN ensures us that two obvious features of the plaintiff will not be ignored: the exposure, x , and the injury, y . In contrast, these two features are ignored in the causal effect measure $P(y_x)$ which is a quantity averaged over the entire population, including unexposed and uninjured.

7.4.3.1 Bounds under strong exogeneity

The condition of exogeneity, as defined in Eq. (7.27) is testable by comparing experimental and nonexperimental data. A stronger version of exogeneity can be defined as the joint independence $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$ which was called “strong ignorability” by Rosenbaum and Rubin [RR83]. Though untestable, such joint independence is assumed to hold when we assert the absence of factors that simultaneously affect exposure and outcome.

Definition 18 (*Strong Exogeneity*)

A variable X is said to be strongly exogenous for Y in model M iff $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, that is,

$$\begin{aligned} P(y_x, y_{x'}|x) &= P(y_x, y_{x'}) \\ P(y_x, y'_{x'}|x) &= P(y_x, y'_{x'}) \\ P(y'_x, y_{x'}|x) &= P(y'_x, y_{x'}) \\ P(y'_x, y'_{x'}|x) &= P(y'_x, y'_{x'}) \end{aligned} \tag{7.33}$$

The four conditions in (7.33) are sufficient to represent $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, because for every event E we have

$$P(E|x) = P(E) \implies P(E|x') = P(E). \tag{7.34}$$

Remarkably, the added constraints introduced by strong exogeneity do not alter the bounds of Eqs. (7.29)–(7.31). They do, however, strengthen Lemma 20:

¹¹Our results remain valid when we condition P_{XY} on a set of covariates that characterize the specific case at hand.

Theorem 24 *If strong exogeneity holds, the probabilities PN, PS, and PNS are constrained by the bounds of Eqs. (7.29)–(7.31), and, moreover, PN, PS, and PNS are related to each other as follows [Pea99] :*

$$PN = \frac{PNS}{P(y|x)} \quad (7.35)$$

$$PS = \frac{PNS}{P(y'|x')} \quad (7.36)$$

7.4.4 Identifiability under monotonicity

Definition 19 (Monotonicity)

A variable Y is said to be monotonic relative to variable X in a functional causal model M iff

$$y'_x \wedge y_{x'} = \text{false} \quad (7.37)$$

Monotonicity expresses the assumption that a change from $X = \text{false}$ to $X = \text{true}$ cannot, under any circumstance make Y change from *true* to *false*. In epidemiology, this assumption is often expressed as “no prevention,” that is, no individual in the population can be helped by exposure to the risk factor. [BP97] used this assumption to tighten bounds of treatment effects from studies involving non-compliance. Glymour [Gly98] and Cheng [Che97] resort to this assumption in using disjunctive or conjunctive relationships between causes and effects, excluding functions such as exclusive-or, or parity.

In the linear programming formulation of Section 7.4.1, monotonicity narrows the feasible space to the manifold:

$$\begin{aligned} p_{011} &= 0 \\ p_{010} &= 0 \end{aligned} \quad (7.38)$$

7.4.4.1 Given nonexperimental data

Under the constraints (7.15), (7.16), and (7.38), we find the same bounds for PNS as the ones obtained under no assumptions (Eq. (7.21)). Moreover, there are still no constraints on PN and PS. Thus, with nonexperimental data alone, the monotonicity assumption does not provide new information.

However, the monotonicity assumption induces sharper bounds on the causal

effects $P(y_x)$ and $P(y_{x'})$:

$$\begin{aligned} P(y) &\leq P(y_x) \leq 1 - P(x, y') \\ P(x', y) &\leq P(y_{x'}) \leq P(y) \end{aligned} \quad (7.39)$$

Compared with Eq. (7.22), the lower bound for $P(y_x)$ and the upper bound for $P(y_{x'})$ are tightened. The importance of Eq. (7.39) lies in providing a simple necessary test for the assumption of monotonicity. These inequalities are sharp, in the sense that every combination of experimental and non-experimental data that satisfy these inequalities can be generated from some functional causal model in which Y is monotonic in X .

That the commonly made assumption of “no-prevention” is not entirely exempt from empirical scrutiny should come as a relief to many epidemiologists. Alternatively, if the no-prevention assumption is theoretically unassailable, the inequalities of Eq. (7.39) can be used for testing the compatibility of the experimental and non-experimental data, namely, whether subjects used in clinical trials were sampled from the same target population, characterized by the joint distribution P_{XY} .

7.4.4.2 Given causal effects

Constraints (7.15), (7.17), and (7.38) induce no constraints on PN and PS, while the value of PNS is fully determined:

$$PNS = P(y_x, y'_{x'}) = P(y_x) - P(y_{x'})$$

That is, under the assumption of monotonicity, PNS can be determined by experimental data alone, despite the fact that the joint event $y_x \wedge y'_{x'}$ can never be observed.

7.4.4.3 Given both nonexperimental data and causal effects

Under the constraints (7.15)–(7.17) and (7.38), the values of PN, PS, and PNS are all determined precisely.

Theorem 25 *If Y is monotonic relative to X , then PNS, PN, and PS are given by*

$$PNS = P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}) \quad (7.40)$$

$$PN = P(y'_{x'}|x, y) = \frac{P(y) - P(y_{x'})}{P(x, y)} \quad (7.41)$$

$$PS = P(y_x|x', y') = \frac{P(y_x) - P(y)}{P(x', y')} \quad (7.42)$$

Corollary 2 *If Y is monotonic relative to X , then PNS, PN, and PS are identifiable whenever the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable,*

Eqs. (7.40)–(7.42) are applicable to situations where, in addition to observational probabilities, we also have information about the causal effects $P(y_x)$ and $P(y_{x'})$. Such information may be obtained either directly, through separate experimental studies, or indirectly, from observational studies in which certain identifying assumptions are deemed plausible (e.g., assumptions that permits identification through adjustment of covariates). Note that the identification of PN requires only $P(y_{x'})$ while that of PS requires $P(y_x)$. In practice, however, any method that yields the former also yields the latter.

One common class of models which permits the identification of $P(y_x)$ is called *Markovian*.

Definition 20 (*Markovian models*)

A functional causal model M is said to be Markovian if the graph $G(M)$ associated with M is acyclic, and if the exogenous factors u_i are mutually independent. A model is semi-Markovian iff $G(M)$ is acyclic and the exogenous variables are not necessarily independent. A functional causal model is said to be positive-Markovian if it is Markovian and $P(v) > 0$ for every v .

It is shown in [Pea93, Pea95a] that for every two variables, X and Y , in a positive-Markovian model M , the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable and are given by

$$\begin{aligned} P(y_x) &= \sum_{pa_X} P(y|pa_X, x)P(pa_X) \\ P(y_{x'}) &= \sum_{pa_X} P(y|pa_X, x')P(pa_X) \end{aligned} \quad (7.43)$$

where pa_X are (values of) the *parents* of X in the causal graph associate with M (see also [SGS93], [Rob86], and [Pea00, p. 73]). Thus, we can combine Eq. (7.43) with Theorem 25 and obtain a concrete condition for the identification of the probability of causation.

Corollary 3 *If in a positive-Markovian model M , the function $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable and are given by Eqs. (7.40)–(7.42), with $P(y_x)$ given in Eq. (7.43). If monotonicity cannot be ascertained, then PNS, PN and PS are bounded by Eqs. (7.24)–(7.26), with $P(y_x)$ given in Eq. (7.43).*

Broader identification conditions can be obtained through the use of the criteria for identifying $P_x(y)$ in Chapter 5. In particular, Theorem 17 leads to the following corollary:

Corollary 4 *Let \mathbf{GP} be the class of semi-Markovian models that satisfy the graphical criterion of Theorem 17. If $Y_x(u)$ is monotonic, then the probabilities of causation P_{NS} , PS and PN are identifiable in \mathbf{GP} and are given by Eqs. (7.40)–(7.42), with $P(y_x)$ determined by the topology of $G(M)$ through Theorem 17.*

7.4.5 Identifiability under monotonicity and exogeneity

Under the assumption of monotonicity, if we further assume exogeneity, then $P(y_x)$ and $P(y_{x'})$ are identified through Eq. (7.27), and from theorem 25 we conclude that P_{NS} , PN , and PS are all identifiable.

Theorem 26 (*Identifiability under exogeneity and monotonicity*)

If X is exogenous and Y is monotonic relative to X , then the probabilities PN , PS , and P_{NS} are all identifiable, and are given by

$$P_{NS} = P(y|x) - P(y|x') \quad (7.44)$$

$$PN = \frac{P(y) - P(y|x')}{P(x,y)} = \frac{P(y|x) - P(y|x')}{P(y|x)} \quad (7.45)$$

$$PS = \frac{P(y|x) - P(y)}{P(x',y')} = \frac{P(y|x) - P(y|x')}{P(y'|x')} \quad (7.46)$$

These expressions are to be recognized as familiar measures of attribution that often appear in the literature. The r.h.s. of (7.44) is called “risk-difference” in epidemiology, and is also misnamed “attributable risk” [HB87, p. 87]. The probability of necessity, PN , is given by the *excess-risk-ratio* (ERR)

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} = 1 - \frac{1}{RR} \quad (7.47)$$

often misnamed as the *attributable fraction* [Sch82], *attributable-rate percent* [HB87, p. 88], *attributed fraction for the exposed* [KWE96, p. 38], or *attributable proportion* [Col97]. The reason we consider these labels to be misnamed is that ERR invokes purely statistical relationships, hence it cannot in itself serve to measure attribution, unless fortified with some causal assumptions. Exogeneity and monotonicity are the causal assumptions that endow ERR with attributional interpretation, and these assumptions are rarely made explicit in the literature on attribution.

The expression for PS is likewise quite revealing

$$PS = [P(y|x) - P(y|x')]/[1 - P(y|x')], \quad (7.48)$$

as it coincides with what epidemiologists call the “relative difference” [She58], which is used to measure the *susceptibility* of a population to a risk factor x . It also coincides with what Cheng calls “causal power” [Che97], namely, the effect of x on y after suppressing “all other causes of y .” See [Pea99] for additional discussions of these expressions.

To appreciate the difference between Eqs. (7.41) and (7.47) we can rewrite Eq. (7.41) as

$$\begin{aligned} PN &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x,y)} \end{aligned} \quad (7.49)$$

The first term on the r.h.s. of (7.49) is the familiar ERR as in (7.47), and represents the value of PN under exogeneity. The second term represents the correction needed to account for X ’s non-exogeneity, i.e. $P(y_{x'}) \neq P(y|x')$. We will call the r.h.s. of (7.49) by corrected excess-risk-ratio (CERR).

From Eqs. (7.44)–(7.46) we see that the three notions of causation satisfy the simple relationships given by Eqs. (7.35) and (7.36) which we obtained under the strong exogeneity condition. In fact, we have the following theorem.

Theorem 27 *Monotonicity (7.37) and exogeneity (7.27) together imply strong exogeneity (7.33).*

Proof of Theorem 27:

From the monotonicity condition, we have

$$y_{x'} = y_{x'} \wedge (y_x \vee y'_x) = (y_{x'} \wedge y_x) \vee (y_{x'} \vee y'_x) = y_{x'} \wedge y_x. \quad (7.50)$$

Thus we can write

$$P(y_{x'}) = P(y_x, y_{x'}), \quad (7.51)$$

and

$$P(y|x') = P(y_{x'}|x') = P(y_x, y_{x'}|x') \quad (7.52)$$

where consistency condition (7.14) is used. The exogeneity condition (7.27) allows us to equate (7.51) and (7.52), and we obtain

$$P(y_x, y_{x'}|x') = P(y_x, y_{x'}), \quad (7.53)$$

Table 7.1: PN (the probability of necessary causation) as a function of assumptions and available data. ERR stands of the excess-risk-ratio $1 - P(y|x')/P(y|x)$ and CERR is given in Eq. (7.49). The non-entries (—) represent vacuous bounds, that is, $0 \leq PN \leq 1$.

Assumptions		Data Available		
Exogeneity	Monotonicity	Experimental	Nonexperimental	Combined
+	+	ERR	ERR	ERR
+	-	bounds	bounds	bounds
-	+	—	—	CERR
-	-	—	—	bounds

which implies the first of the four conditions in (7.33):

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}). \quad (7.54)$$

Combining Eq. (7.54) with

$$P(y_x) = P(y_x, y_{x'}) + P(y_x, y'_{x'}), \quad (7.55)$$

$$P(y|x) = P(y_x|x) = P(y_x, y_{x'}|x) + P(y_x, y'_{x'}|x), \quad (7.56)$$

and the exogeneity condition (7.27), we obtain the second equation in (7.33):

$$P(y_x, y'_{x'}|x) = P(y_x, y'_{x'}). \quad (7.57)$$

Both sides of the third equation in (7.33) are equal to zero from monotonicity condition and the last equation in (7.33) follows because the four quantities sum up to 1 on both sides of the four equations. \square

7.4.6 Summary of results

We now summarize the results from Section 7.4 that should be of value to practicing epidemiologists and policy makers. These results are shown in Table 7.1, which lists the best estimate of PN under various assumptions and various types of data—the stronger the assumptions, the more informative the estimates.

We see that the excess-risk-ratio (ERR), which epidemiologists commonly identify with the probability of causation, is a valid measure of PN only when two assumptions can be ascertained: exogeneity (i.e., no confounding) and monotonicity (i.e., no prevention). When monotonicity does not hold, ERR provides merely a lower bound for PN, as shown in Eq. (7.30). (The upper bound is usually unity.) In the presence of confounding, ERR must be corrected by the additive term $[P(y|x') - P(y_{x'})]/P(x,y)$, as stated in (7.49). In other words,

when confounding bias (of the causal effect) is positive, PN is higher than ERR by the amount of this additive term. Clearly, owing to the division by $P(x, y)$, the PN bias can be many times higher than the causal effect bias $P(y|x') - P(y_{x'})$. However, confounding results only from association between exposure and other factors that affect the outcome; one need not be concerned with associations between such factors and susceptibility to exposure, as is often assumed in the literature [KFG89, Gly98].

The last two rows in Table 7.1 correspond to no assumptions about exogeneity, and they yield vacuous bounds for PN when data come from either experimental or observational study. In contrast, informative bounds (7.25) or point estimates (7.49) are obtained when data from experimental and observational studies are combined. Concrete use of such combination will be illustrated in Section 7.5.

7.5 Example 1: Legal Responsibility

A lawsuit is filed against the manufacturer of drug x , charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve symptom S associated with disease D .

The manufacturer claims that experimental data on patients with symptom S show conclusively that drug x may cause only minor increase in death rates. The plaintiff argues, however, that the experimental study is of little relevance to this case, because it represents the effect of the drug on *all* patients, not on patients like Mr. A who actually died while using drug x . Moreover, argues the plaintiff, Mr. A is unique in that he used the drug on his own volition, unlike subjects in the experimental study who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes nonexperimental data indicating that most patients who chose drug x would have been alive if it were not for the drug. The manufacturer counter-argues by stating that: (1) counterfactual speculations regarding whether patients would or would not have died are purely metaphysical and should be avoided, and (2) nonexperimental data should be dismissed a priori, on the ground that such data may be highly biased; for example, incurable terminal patients might be more inclined to use drug x if it provides them greater symptomatic relief. The court must now decide, based on both the experimental and non-experimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The (hypothetical) data associated with the two studies are shown in Table

Table 7.2: Frequency data (hypothetical) obtained in experimental and nonexperimental studies, comparing deaths (in thousands) among drug users, x , and non-users, x' .

	Experimental		Nonexperimental	
	x	x'	x	x'
Deaths(y)	16	14	2	28
Survivals(y')	984	986	998	972

7.2. The experimental data provide the estimates

$$\begin{aligned} P(y_x) &= 16/1000 = 0.016 \\ P(y_{x'}) &= 14/1000 = 0.014 \\ P(y'_{x'}) &= 1 - P(y_{x'}) = 0.986 \end{aligned}$$

The non-experimental data provide the estimates

$$\begin{aligned} P(y) &= 30/2000 = 0.015 \\ P(x, y) &= 2/2000 = 0.001 \\ P(x', y') &= 972/2000 = 0.486 \end{aligned}$$

Since both the experimental and nonexperimental data are available, we can obtain bounds on all three probabilities of causation through Eqs. (7.24)–(7.26) without making any assumptions about the underlying mechanisms. The data in Table 7.2 imply the following numerical results:

$$0.002 \leq PNS \leq 0.016 \quad (7.58)$$

$$1.0 \leq PN \leq 1.0 \quad (7.59)$$

$$0.002 \leq PS \leq 0.031 \quad (7.60)$$

These figures show that although surviving patients who didn't take drug x have only less than 3.1% chance to die had they taken the drug, there is 100% assurance (barring sample errors) that those who took the drug and died would have survived had they not taken the drug. Thus the plaintiff was correct; drug x was in fact responsible for the death of Mr. A.

If we assume that drug x can only cause, but never prevent, death, Theorem 25 is applicable and Eqs. (7.40)–(7.42) yield

$$PNS = 0.002 \quad (7.61)$$

$$PN = 1.0 \quad (7.62)$$

$$PS = 0.002 \quad (7.63)$$

Thus, we conclude that drug x was responsible for the death of Mr. A, with or without the no-prevention assumption.

Note that a straightforward use of the experimental excess-risk-ratio would yield a much lower (and incorrect) result:

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125 \quad (7.64)$$

Evidently, what the experimental study does not reveal is that, given a choice, terminal patients stay away from drug x . Indeed, if there were any terminal patients who would choose x (given the choice), then the control group (x') would have included some such patients (due to randomization) and so the proportion of deaths among the control group $P(y_{x'})$ would have been higher than $P(x', y)$, the population proportion of terminal patients avoiding x . However, the equality $P(y_{x'}) = P(y, x')$ tells us that no such patients were present in the control group, hence (by randomization) no such patients exist in the population at large and therefore none of the patients who freely chose drug x was a terminal case; all were susceptible to x .

The numbers in Table 7.2 were obviously contrived to show the usefulness of the bounds in Eqs. (7.24)-(7.26). Nevertheless, it is instructive to note that a combination of experimental and non-experimental studies may unravel what experimental studies alone will not reveal. In addition, such combination may provide a test for the assumption of no-prevention, as outlined in Section 7.4.4.1. For example, if the frequencies in Table 2 were slightly different, they could easily violate the inequalities of Eq. (7.39). Such violation may be due either to nonmonotonicity or to incompatibility of the experimental and nonexperimental groups.

This last point may warrant a word of explanation, lest the reader wonders why two data sets, taken from two separate groups under different experimental conditions, should constrain one another. The explanation is that certain quantities in the two subpopulations are expected to remain invariant to all these differences, provided that the two subpopulations were sampled properly from the same general population. In fact, every quantity of the form $P(Q)$, where Q is computable from a functional causal model M , enjoys this invariance property, because the two subpopulations are assumed to be governed by the same functional causal model. Thus, the question whether two data sets, obtained under different experimental conditions, should constrain one another reduces to a purely mathematical question of whether the quantities that represent the two experimental conditions, $P(Q)$ and $P(Q')$, necessarily constrain one another in the same functional causal model considered. In our case, the quantities in question are simply the causal effects probabilities, $P(y_{x'})$ and $P(y_x)$. Although

these probabilities were not measured in the nonexperimental group, they must nevertheless be the same as those measured in the experimental group. The invariance of these quantities is the basic axiom of controlled experimentation, without which *no* inference would be possible from experimental studies to general behavior of the population. This invariance, together with monotonicity, imply the inequalities of (7.39).

7.6 Example 2: Personal Decision Making

Consider the case of Mr. B, who is one of the surviving patients in the observational study of Table 7.2. Mr. B wonders how safe it would be for him to take drug x , given that he has refrained thus far from taking the drug and that he managed to survive the disease. His argument for switching to the drug rests on the observation that only 2 out of 1000 drug users died in the observational study, which he considers a rather small risk to take, given the effectiveness of the drug as a pain killer.

Conventional wisdom instructs us to warn Mr. B against consulting a non-experimental study in matters of decisions, since such studies are marred with uncontrolled factors, which tend to bias effect estimates. Specifically, the death rate of 0.002 among drug users may be indicative of low tolerance to discomfort, or of membership in a medically-informed socio-economic group. Such factors do not apply to Mr. B, who did not use the drug in the past (be it by choice, instinct or ignorance), and who is now considering switching to the drug by rational deliberation. Conventional wisdom urges us to refer Mr. B to the randomized experimental study of Table 7.2, from which the death rate under controlled administration of the drug was evaluated to be $P(y_x) = 0.016$, eight times higher than 0.002.

What would his risk of death be, if Mr. B decides to start taking the drug? 0.2 percent or 1.6 percent?

The answer is that neither number is correct. Mr. B cannot be treated as a random patient in either study, because his history of not using the drug and his survival thus far puts him in a unique category of patients, for which the effect of the drug was not studied.¹² These two attributes provide extra evidence about Mr. B's sensitivity to the drug. This became clear already in Example 1, where we discovered definite relationships among these attributes – for some obscure reasons, terminal patients chose not to use the drug.

¹²The appropriate experimental design for measuring the risk of interest is to conduct a randomized clinical trial on patients in the category of Mr. B, that is, to subject a random sample of non-users to a period of drug treatment and measure their rate of survival.

To properly account for this additional evidence, the risk should be measured through the counterfactual expression $PS = P(y_x|x', y')$; the probability that a patient who survived with no drug would have died had he/she taken the drug. The appropriate bound for this probability is given in Eq. (7.60):

$$0.002 \leq PS \leq 0.031$$

Thus, Mr. B's risk of death (upon switching to drug usage) can be as high as 3.1 percent; more than 15 times his intuitive estimate of 0.2 percent, and almost twice the naive estimate obtained from the experimental study.

However, if the drug can safely be assumed to have no death-preventing effects, then monotonicity applies, and the appropriate bound is given by Eq. (7.63), $PS = 0.002$, which coincides with Mr. B's intuition.

7.7 Conclusion

This chapter shows how useful information about probabilities of causation can be obtained from experimental and observational studies, with weak or no assumptions about the data-generating process. We have shown that, in general, bounds for the probabilities of causation can be obtained from combined experimental and nonexperimental data. These bounds were proven to be sharp and, therefore, they represent the ultimate information that can be extracted from statistical methods. We have further illustrated the applicability of these results to problems in epidemiology and legal reasoning, and we have clarified the two basic assumptions – exogeneity and monotonicity – that must be ascertained before statistical measures such as excess-risk-ratio could represent attributional quantities such as probability of causation.

It is appropriate at this point to discuss the relation between the assumptions in the example of Section 7.5 (where we have population probabilities and available experiments) with the general framework with which the chapter begins (where we have exogenous variables that determine everything and the probabilities enter as an add-on feature). Traditional statisticians might judge the deterministic model incompatible with the stochastic nature of the data, and would be tempted to start the analysis at Section 7.3 (see [RG89] and [FS99]), without the counterfactual model expounded in Section 7.2. However, traditional statistical analysis cannot commence without explicating the quantity we wish to estimate (that is, PN), for which we have no empirical data and for which we have no statistical definition. Instead, our target quantity is defined verbally by law makers as a mixture of probabilistic and deterministic components: “it is more *probable* than not, that the plaintiff injury would not have occurred *but*

for the defender action". The "more probable than not" criterion is probabilistic while the "but for" criterion is deterministic, implying counterfactual necessity.

The structural approach expounded in this chapter gives a clear semantics to this mixture, typical of counterfactual expressions, and relates it in a natural way to empirical data. The stochastic nature of the data is viewed as emerging from our ignorance of the detailed experimental conditions that prevailed in the study. The exogenous variables in U represent these missing details, and include the physiology and previous history of each person, his/her mental and spiritual attitude, as well as the time and manner in which the exposure occurred. In short, U summarizes all the factors which "determine" in the classical physical sense the outcome of the study. $P(u)$ summarizes our ignorance of those factors.

The main application of our analysis to artificial intelligence lies in the automatic generation of causal explanations, where the distinction between necessary and sufficient causes has important ramifications. As can be seen from the definitions and examples discussed in this chapter, necessary causation is a concept tailored to a specific event under consideration (singular causation), whereas sufficient causation is based on the general tendency of certain event *types* to produce other event types. Adequate explanations should respect both aspects. If we base explanations solely on generic tendencies (i.e., sufficient causation) then we lose important scenario-specific information. For instance, aiming a gun at and shooting a person from 1,000 meters away will not qualify as an explanation for that person's death, owing to the very low tendency of shots fired from such long distances to hit their marks. This stands contrary to common sense, for when the shot does hit its mark on that singular day, regardless of the reason, the shooter is an obvious culprit for the consequence. If, on the other hand, we base explanations solely on singular-event considerations (i.e., necessary causation), then ambient factors that are normally present in the world would awkwardly qualify as explanations. For example, the presence of oxygen in the room would qualify as an explanation for the fire that broke out, simply because the fire would not have occurred were it not for the oxygen. That we judge the match struck, not the oxygen, to be the more adequate explanation of the fire indicates that we go beyond necessity considerations.

Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact, the PN is actually higher for the oxygen if we allow for alternative ways of igniting a spark). Thus, it must be the sufficiency component that endows the match with greater explanatory power than the oxygen. If the probabilities associated with striking a match and the presence of oxygen are denoted p_m and p_o , respectively, then the PS measures associated with these explanations evaluate to $PS(\text{match}) = p_o$ and $PS(\text{oxygen}) = p_m$, clearly favoring the match

when $p_o \gg p_m$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

Clearly, some balance must be made between the necessary and the sufficient components of causal explanation, and the present chapter illuminates this balance by formally explicating the basic relationships between the two components. In Pearl (2000, chapter 10) it is further shown that PN and PS are too crude for capturing probabilities of causation in multi-stage scenarios, and that the structure of the intermediate process leading from cause to effect must enter the definitions of causation and explanation. Such considerations will be the subject of future investigation (See [HP00]).

Another important application of probabilities of causation is found in decision making problems, such as those encountered in medicine, system maintenance, and planning under uncertainty. As was pointed out in [Pea00, p. 217-219], the counterfactual “ y would have been true if x were true” can often be translated into a conditional action claim “given that currently x and y are false, y will be true if we do x .” The evaluation of such conditional predictions, and the probabilities of such predictions, are commonplace in decision making situations, where actions are brought into focus by certain eventualities that demand remedial correction. In troubleshooting, for example, we observe undesirable effects $Y = y$ that are potentially caused by other conditions $X = x$ and we wish to predict whether an action that brings about a change in X would remedy the situation. The information provided by the evidence y and x is extremely valuable, and it must be processed (using the updated distribution $P(u|x, y)$, as in Eq. (7.9)) before we can predict the effect of any action¹³. Thus, the expressions developed in this chapter constitute bounds on the effectiveness of pending policies, when full knowledge of the current state of affairs (u) is not available, yet the current states of the decision variable (X) and the outcome variable (Y) are measured.

For these bounds to be valid in policy making, the context u must be time-invariant, that is, the probability $P(u)$ should represent epistemic uncertainty about a static, albeit unknown context $U = u$. The constancy of u is well justified in the control and diagnosis of physical systems, where u represents fixed, but unknown physical characteristics of devices or subsystems. The constancy approximation is also justified in the health sciences where patients’ genetic attributes and physical characteristics can be assumed relatively constant between observation and treatment.

The constancy assumption is less justified in economic systems, where agents are bombarded by rapidly fluctuating stream of external forces (“shocks” in econometric terminology) as well as by inter-agents communication messages.

¹³Such processing have been applied indeed to the evaluation of economic policies [BP95] and to repair-test strategies in troubleshooting [BH96]

These exogenous factors may vary substantially during the policy making interval and they require, therefore, time-dependent analysis. The canonical violation of the constancy assumption occurs, of course, in quantum mechanical systems, where the indeterminism associated with U is “intrinsic”, and the existence of a deterministic relationship between U and V is no longer a good approximation. A method of incorporating such intrinsic indeterminism into counterfactual analysis is outlined in [Pea00, p. 220], and leads to Eq. (7.9), where $P(Y_{x'}(u) = y')$ represents the intrinsic uncertainty in Y associated with the macroscopic state $U = u$, under the action $do(X = x)$ (see footnote 6).

REFERENCES

- [Ald93] J. Aldrich. “Cowles’ Exogeneity and Core Exogeneity.” Technical Report Discussion Paper 9308, Department of Economics, University of Southampton, England, 1993.
- [Bal95] A. Balke. *Probabilistic Counterfactuals: Semantics, Computation, and Applications*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 1995.
- [BGG94] L. A. Bailey, L. Gordis, and M. Green. “Reference guide on epidemiology.” *Reference Manual on Scientific Evidence*, 1994. Federal Judicial Center. Available online at http://www.fjc.gov/EVIDENCE/science/sc_ev_sec.html.
- [BH96] J.S. Breese and D. Heckerman. “Decision-theoretic troubleshooting: A framework for repair and experiment.” In E. Horvitz and F. Jensen, editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pp. 124–132. Morgan Kaufmann, San Francisco, CA, 1996.
- [BP94] A. Balke and J. Pearl. “Probabilistic evaluation of counterfactual queries.” In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume I, pp. 230–237. MIT Press, Menlo Park, CA, 1994.
- [BP95] A. Balke and J. Pearl. “Counterfactuals and Policy Analysis in Structural Models.” In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pp. 11–18. Morgan Kaufmann, San Francisco, 1995.
- [BP97] A. Balke and J. Pearl. “Nonparametric bounds on causal effects from partial compliance data.” *Journal of the American Statistical Association*, **92**(439):1–6, September 1997.
- [BP02] C. Brito and J. Pearl. “Generalized Instrumental Variables.” In *Proceedings of the Uncertainty in Artificial Intelligence*, 2002.
- [BSC89] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and Cooper G. F. “The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks.” In *Proceedings of the second European conference on Artificial Intelligence in Medicine*, pp. 247–256, London, England, 1989.

- [BT84] R.J. Bowden and D.A. Turkington. *Instrumental Variables*. Cambridge University Press, Cambridge, England, 1984.
- [CH92] G. F. Cooper and E. Herskovits. “A Bayesian method for the induction of probabilistic networks from data.” *Machine Learning*, **9**:309–347, 1992.
- [Che97] P.W. Cheng. “From covariation to causation: A causal power theory.” *Psychological Review*, **104**(2):367–405, 1997.
- [Col97] P. Cole. “Causality in epidemiology, health policy, and law.” *Journal of Marketing Research*, **27**:10279–10285, 1997.
- [Coo99] G.F. Cooper. “An overview of the representation and discovery of causal relationships using Bayesian networks.” In Glymour C. and Cooper G.F., editors, *Computation, Causation, and Discovery*, Menlo Park, CA, 1999. AAAI Press and MIT Press.
- [CY99] G. F. Cooper and C. Yoo. “Causal Discovery from a Mixture of Experimental and Observational Data.” In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 116–125, San Francisco, CA, 1999. Morgan Kaufmann Publishers.
- [Daw97] A.P. Dawid. “Causal inference without counterfactuals.” Technical report, Department of Statistical Science, University College London, UK, 1997. Forthcoming, with discussion, *Journal of the American Statistical Association*, 2000.
- [Daw02] A.P. Dawid. “Influence diagrams for causal modelling and inference.” *International Statistical Review*, **70**(2), 2002.
- [Des99] B. Desjardins. *On the theoretical limits to reliable causal inference*. PhD thesis, University of Pittsburgh, 1999.
- [Dhr70] P.J. Dhrymes. *Econometrics*. Springer-Verlag, New York, 1970.
- [EHR83] R.F. Engle, D.F. Hendry, and J.F. Richard. “Exogeneity.” *Econometrica*, **51**:277–304, 1983.
- [Fin85] K. Fine. *Reasoning with Arbitrary Objects*. B. Blackwell, New York, 1985.
- [Fis66] F.M. Fisher. *The Identification Problem in Econometrics*. McGraw-Hill, 1966.

- [Fis70] F.M. Fisher. “A correspondence principle for simultaneous equations models.” *Econometrica*, **38**(1):73–92, January 1970.
- [FS99] D. A. Freedman and P. B. Stark. “The swine flu vaccine and Guillain-Barré syndrome: A case study in relative risk and specific causation.” *Evaluation Review*, **23**(6):619–647, Dec. 1999.
- [Gei95] D. Heckerman, D. Geiger, and D.M. Chickering. “Learning Bayesian networks: The combination of knowledge and statistical data.” *Machine Learning*, **20**:197–243, 1995.
- [Gly98] C. Glymour. “Psychological and Normative Theories of Causal Power and the Probabilities of Causes.” In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pp. 166–172. Morgan Kaufmann, San Francisco, CA, 1998.
- [GM98] Dan Geiger and Christopher Meek. “Graphical Models and Exponential Families.” In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pp. 156–165, San Francisco, CA, 1998. Morgan Kaufmann Publishers.
- [GM99] Dan Geiger and Christopher Meek. “Quantifier Elimination for Statistical Problems.” In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pp. 226–235, San Francisco, CA, 1999. Morgan Kaufmann Publishers.
- [Goo61] I.J. Good. “A Causal Calculus, I.” *British Journal for the Philosophy of Science*, **11**:305–318, 1961.
- [Goo93] I.J. Good. “A tentative measure of probabilistic causation relevant to the philosophy of the law.” *J. Statist. Comput. and Simulation*, **47**:99–105, 1993.
- [GP95] D. Galles and J. Pearl. “Testing identifiability of causal effects.” In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pp. 185–195. Morgan Kaufmann, San Francisco, 1995.
- [GP97] D. Galles and J. Pearl. “Axioms of Causal Relevance.” *Artificial Intelligence*, **97**(1-2):9–43, 1997.
- [GP98] D. Galles and J. Pearl. “An axiomatic characterization of causal counterfactuals.” *Foundations of Science*, **3**(1):151–182, 1998.
- [GPR99] S. Greenland, J. Pearl, and J.M Robins. “Causal diagrams for epidemiologic research.” *Epidemiology*, **10**:37–48, 1999.

- [Hal98] J.Y. Halpern. “Axiomatizing Causal Reasoning.” In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pp. 202–210. Morgan Kaufmann, San Francisco, CA, 1998.
- [Hal02] N. Hall. “Two concepts of causation.” In J. Collins, N. Hall, and I. Paul, editors, *Counterfactuals and Causation*. MIT Press, 2002.
- [HB87] C.H. Hennekens and J.E. Buring. *Epidemiology in Medicine*. Brown, Little, Boston, 1987.
- [Hec95] D. Heckerman. “A Bayesian Approach to Learning Causal Networks.” In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pp. 274–284, San Francisco, CA, 1995. Morgan Kaufmann Publishers.
- [Hen95] David F. Hendry. *Dynamic Econometrics*. Oxford University Press, New York, 1995.
- [HMC97] D. Heckerman, C. Meek, and G. Cooper. “A Bayesian approach to causal discovery.” Technical Report MSR-TR-97-05, Microsoft Research, 1997.
- [Hoo90] K.D. Hoover. “The logic of causal inference.” *Economics and Philosophy*, **6**:207–234, 1990.
- [HP00] J. Y. Halpern and J. Pearl. “Causes and Explanations: A Structural-Model Approach.” Technical Report R-266, Cognitive System Laboratory, Department of Computer Science, University of California, Los Angeles, March, 2000.
- [HS95] D. Heckerman and R. Shachter. “Decision-Theoretic Foundations for Causal Reasoning.” *Journal of Artificial Intelligence Research*, **3**:405–430, 1995.
- [imb97] G.W. Imbens. “Book Reviews.” *Journal of Applied Econometrics*, **12**, 1997.
- [Jor98] M.I. Jordan. *Learning in Graphical Models*. Series D: Behavioural and Social Sciences – Vol. 89. Kluwer Academic Publishers, Dordrecht, 1998.
- [KFG89] M.J. Khoury, W.D. Flanders, S. Greenland, and M.J. Adams. “On the measurement of susceptibility in epidemiologic studies.” *American Journal of Epidemiology*, **129**(1):183–190, 1989.

- [Kim71] J. Kim. “Causes and events: Mackie on causation.” *Journal of Philosophy*, **68**:426–471, 1971. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.
- [KM99] M. Kuroki and M. Miyakawa. “Identifiability criteria for causal effects of joint interventions.” *Journal of the Japan Statistical Society*, **29**(2):105–117, 1999.
- [KSC84] H. Kiiveri, T.P. Speed, and J.B. Carlin. “Recursive causal models.” *Journal of Australian Math Society*, **36**:30–52, 1984.
- [KWE96] J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D. Thompson. *Methods in Observational Epidemiology*. Oxford University Press, New York, 1996.
- [Lau00] S. Lauritzen. “Graphical models for causal inference.” In O.E. Barndorff-Nielsen, D. Cox, and C. Kluppelberg, editors, *Complex Stochastic Systems*, chapter 2, pp. 67–112. Chapman and Hall/CRC Press, London/Boca Raton, 2000.
- [Lew86] D. Lewis. *Philosophical Papers*. Oxford University Press, New York, 1986.
- [Mac65] J.L. Mackie. “Causes and conditions.” *American Philosophical Quarterly*, **2/4**:261–264, 1965. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.
- [Mar50] J. Marschak. “Statistical inference in economics.” In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pp. 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.
- [McD97] R.P. McDonald. “Haldane’s Lungs: A Case study in Path Analysis.” *Multivariate Behavioral Research*, **32**(1):1–38, 1997.
- [Mee95] C. Meek. “Causal inference and causal explanation with background knowledge.” In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pp. 403–410. Morgan Kaufmann, San Francisco, 1995.
- [Mic00] D. Michie. “Adapting Good’s Q theory to the causation of individual events.” *Machine Intelligence*, **15**, 2000.
- [Mil43] J.S. Mill. *System of Logic*, volume 1. John W. Parker, London, 1843.

- [Pea88] J. Pearl. *Probabilistic Reasoning in Intelligence Systems*. Morgan Kaufmann, San Mateo, CA, 1988.
- [Pea93] J. Pearl. “Comment: Graphical Models, Causality, and Intervention.” *Statistical Science*, 8:266–269, 1993.
- [Pea94] J. Pearl. “A probabilistic calculus of actions.” In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pp. 454–462. Morgan Kaufmann, San Mateo, CA, 1994.
- [Pea95a] J. Pearl. “Causal diagrams for experimental research.” *Biometrika*, 82:669–710, December 1995.
- [Pea95b] J. Pearl. “On the Testability of Causal Models with Latent and Instrumental Variables.” In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pp. 435–443. Morgan Kaufmann, 1995.
- [Pea99] J. Pearl. “Probabilities of causation: three counterfactual interpretations and their identification.” *Synthese*, 121(1-2):93–149, November 1999.
- [Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, 2000.
- [PGV90] J. Pearl, D. Geiger, and T. Verma. “The logic of influence diagrams.” In R.M. Oliver and J.Q. Smith, editors, *Influence Diagrams, Belief Nets and Decision Analysis*, pp. 67–87. John Wiley and Sons, Inc., New York, NY, 1990.
- [PR95] J. Pearl and J.M. Robins. “Probabilistic evaluation of sequential plans from causal models with hidden variables.” In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pp. 444–453. Morgan Kaufmann, San Francisco, 1995.
- [PV91] J. Pearl and T. Verma. “A Theory of Inferred Causation.” In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452. Morgan Kaufmann, San Mateo, CA, 1991.
- [Rei56] H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.
- [RG89] J.M. Robins and S. Greenland. “The probability of causation under a stochastic model for individual risk.” *Biometrics*, 45:1125–1138, 1989.

- [Rob86] J.M. Robins. “A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect.” *Mathematical Modeling*, **7**:1393–1512, 1986.
- [Rob87] J.M. Robins. “A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods.” *Journal of Chronic Diseases*, **40**(Suppl 2):139S–161S, 1987.
- [Rob97a] D.W. Robertson. “The common sense of cause in fact.” *Texas Law Review*, **75**(7):1765–1800, 1997.
- [Rob97b] J.M. Robins. “Causal inference from complex longitudinal data.” In *Latent Variable Modeling with Applications to Causality*, pp. 69–117. Springer-Verlag, New York, 1997.
- [RR83] P. Rosenbaum and D. Rubin. “The central role of propensity score in observational studies for causal effects.” *Biometrika*, **70**:41–55, 1983.
- [RW97] James M. Robins and Larry A. Wasserman. “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)*, pp. 409–420, San Francisco, CA, 1997. Morgan Kaufmann Publishers.
- [SBM00] C. Silverstein, S. Brin, R. Motwani, and J. Ullman. “Scalable Techniques for Mining Causal Structures.” *Data Mining and Knowledge Discovery*, **4**(2/3):163–192, 2000.
- [Sch82] J.J. Schlesselman. *Case-Control Studies: Design Conduct Analysis*. Oxford University Press, New York, 1982.
- [SGS93] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.
- [She58] M.C. Shep. “Shall we count the living or the dead?” *New England Journal of Medicine*, **259**:1210–1214, 1958.
- [Sim53] H.A. Simon. “Causal ordering and identifiability.” In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pp. 49–74. Wiley and Sons, Inc., 1953.
- [Sob90] M.E. Sobel. “Effect Analysis and Causation in Linear Structural Equation Models.” *Psychometrika*, **55**(3):495–515, 1990.

- [SR66] H.A. Simon and N. Rescher. “Cause and Counterfactual.” *Philosophy and Science*, **33**:323–340, 1966.
- [SW60] R.H. Strotz and H.O.A. Wold. “Recursive versus nonrecursive systems: An attempt at synthesis.” *Econometrica*, **28**:417–427, 1960.
- [Ver93] T. S. Verma. “Graphical aspects of causal models.” Technical Report R-191, Computer Science Department, University of California, Los Angles, 1993.
- [VP90] T. Verma and J. Pearl. “Equivalence and Synthesis of Causal Models.” In P. Bonissone et al., editor, *Uncertainty in Artificial Intelligence 6*, pp. 220–227. Elsevier Science, Cambridge, MA, 1990.