CAVEATS FOR CAUSAL REASONING WITH EQUILIBRIUM MODELS

A Dissertation by

DENVER DASH

B.Sc., CASE WESTERN RESERVE UNIVERSITY

M.Sc., UNIVERSITY OF PITTSBURGH

Submitted to the Graduate Faculty of
Arts and Sciences in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 2003

Intelligent Systems
University of Pittsburgh

CAVEATS FOR CAUSAL REASONING WITH EQUILIBRIUM MODELS

A Dissertation by

DENVER DASH

Approved as to style and content by:

Marek J. Druzdzel, Chair

Gregory F. Cooper, Member

Richard Scheines, Member

Milos Hauskrecht, Member

Vanathi Gopalakrishnan, Member

ABSTRACT

CAVEATS FOR CAUSAL REASONING WITH EQUILIBRIUM MODELS

Denver Dash, Ph.D.

University of Pittsburgh, April 2003

This thesis raises objections to the use of causal reasoning with equilibrium models. I consider two operators that are used to transform models: the *Do* operator for modeling manipulation and the *Equilibration* operator for modeling a system that has achieved equilibrium. I introduce a property of a causal model called the *EMC Property* that is true iff the *Do* operator commutes with the *Equilibration* operator. I prove that not all models obey the EMC property, and I demonstrate empirically that, when inferring a causal model from data, the learned model will not support causal reasoning if the EMC property is not obeyed. I find sufficient conditions for models to violate and not to violate the EMC property. In addition, I show that there exists a class of models that violate EMC and possess a set of variables whose manipulation will cause an instability in the system. All dynamic models in this class possess feedback, although I do not prove that feedback is a necessary or a sufficient condition for EMC violation. I define the *Structural Stability Principle* which provides a necessary graphical criterion for stability in causal models. Finally, I will argue that the models in this class are quite common given typical assumptions about causal relations.

To mom, Erica, & Googl

ACKNOWLEDGMENTS

First and most I would like to thank my advisor Marek Druzdzel: for taking me into the Decision Systems Laboratory; for suggesting to me a thought-provoking research topic; for arranging my “sabbatical” semester in Poland which deepened my knowledge of machine learning and broadened my knowledge about the world; for encouraging me to write my NASA fellowship proposal; for giving me much insight into many different ways to be a good academic that are too numerous to mention; for Alaska; in short, for rescuing me from a life of endless, thankless physics postdocs, and turning me into a world traveller.

Thanks to the NASA Graduate Student Research program for three years of funding which gave me the freedom to explore many research ideas that I would certainly not have had otherwise.

Thanks to all my committee for helpful suggestions. Thanks to Richard Scheines, Nanny Wermuth and Greg Cooper for their encouragement for this research, and thanks to Peter Spirtes, Clark Glymour, and the late Herb Simon for fruitful discussions. Thanks to Andy Connolly in Physics for help in proofreading and sponsoring me for my NASA fellowship.

Thanks to all DSL members whom I have had the pleasure to interact with, especially Tsai-Ching Lu, Hans van Leijen, Jeroen Bogers and Roger Flynn for great and fun discussions about causality, and Yan Lin and Dani Garcia for helping to enlighten me about program design and programming in general. Thanks also to Adam Zagorecki, who was always ready to help me out and Haiqin Wang for our always interesting debates. Thanks to Leon Bobrowski and Agnieszka Oniśko for their generous hospitality in Poland.

Thanks to my family for their (cough) patience and support, and all my friends for much-needed diversions. Of course, thanks to Erica for giving me the best times of my life so far.

I would like to thank several graduate and undergraduate professors whom have influenced and helped me. I would like to thank Cyrus Taylor, Philip Taylor and Don Schuele from CWRU who were invaluable in getting me into graduate school in the first place. Thanks to Gary Chottiner from CWRU, Xiao-Lun Wu and Maartin Rutgers from Pitt for helping to shape my outlook on research design. Thanks to Anthony Duncan from Physics and Kirk Pruhs from Computer Science for their outstanding and inspirational teaching that more than any other people have influenced my own teaching style.

Finally, thanks to Keena Walker for all her invaluable help. Thanks to all the staff of the Department of Physics and Astronomy for helping me manage my NASA grant and for their outstanding competence, and to Susan Strauss Williams at SIS for her patience and grace.

Pittsburgh, Pennsylvania Denver Dash

April 27, 2003

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

"We find ourselves confronted with this paradox: in order for the comparative-statics analysis to yield fruitful results, we must first develop a theory of dynamics." —Paul A. Samuelson, Foundations of Economic Analysis

INTRODUCTION

A precise definition of causality has proven elusive to philosophers ever since David Hume. Hume argued that the belief in causality was not an act of reason, but rather an act of habit. That habit has served the human race well, giving us a powerful tool for organizing phenomena, to form and test scientific hypotheses, to predict the effects of actions on our environment, and to reason about counterfactual claims. It could be argued that no human characteristic has done more to advance human progress than this habit. The power of causal reasoning thus provides a justification for seeking a mathematical definition of causality even if there is no purely logical reason to accept a given causal model.

In traditional scientific disciplines, causal explanations are not typically embedded explicitly into the models being studied. Mathematical effort in these disciplines has instead concentrated on the ability to compile the systems into simplified forms that allow for the easiest analytical or numerical solutions, and causal knowledge is used only implicitly by the scientist to construct these reduced-form relationships.

There are, however, good reasons to include causality explicitly within a model's representation. First, when one is studying extremely complex systems, such as those found in biological or social science applications, sorting out the network of cause and effect can be extremely difficult to accomplish without an explicit representation. In the study of artificial intelligence, an explicit representation of causality creates the potential for developing an agent that can perform extremely sophisticated reasoning tasks. Constructing a causal model provides an agent with a robust means to diagnose symptoms, to perform prediction given a current observed state of the system, and most importantly, a causal model releases an agent from the need to store a combinatorially large set of pairs $\{action \Rightarrow effect\}$, allowing the result of external manipulation on various system components to be predicted directly from the model using the *Do* operator [Wold, 1954; Goldszmidt and Pearl, 1992]. By accepting the assumption of *causal faithfulness* [Pearl, 1988; Pearl and Verma, 1991; Spirtes *et al.*, 1993], it is possible in principle to recover causal models from data using constraint-based [Spirtes *et al.*, 1993; Verma and Pearl, 1991; Cheng *et al.*, 2002] or Bayesian [Cooper and Herskovits, 1992; Heckerman *et al.*, 1995; Bouckaert, 1995] causal discovery methods. Causal reasoning plus the ability to learn causal models from data could potentially enable an intelligent agent to build and test hypotheses about its environment and could help automate the process of scientific discovery from data. These are topics that sit on the forefront of artificial intelligence research.

It has been shown by Iwasaki and Simon [1994] that, given assumptions about the form of the causal model, the causal relations governing a dynamic system can change as the time-scale of observation of the system is increased. In particular, they introduce the *Equilibration* operator that produces the causal relations of a system in *equilibrium* given the dynamic (non-equilibrium) causal system.

Informally, the *Do* operator, $Do(M, \mathbf{U} = \mathbf{u})$, transforms a causal model M to a new causal model M' where a subset of variables $\mathbf{U}$ in M' are fixed to specific values independent of the causes of $\mathbf{U}$. On the other hand, the *Equilibration* operator, $Equilibrate(M, X)$, transforms the model M with a dynamic (time-varying) variable X to a new causal model M' where X is static. This thesis will address the relationship between these two operators. In particular I am interested in the following property:

Definition 1 (Equilibration-Manipulation Commutability) *Let $M(\mathbf{V})$ be a causal model over variables $\mathbf{V}$. M satisfies the Equilibration-Manipulation Commutability (EMC) property iff $Equilibrate(Do(M, \mathbf{U} = \mathbf{u}), X) = Do(Equilibrate(M, X), \mathbf{U} = \mathbf{u})$, for all $\mathbf{U} \subseteq \mathbf{V}$ and all $X \in \mathbf{V}$.*

I use the shorthand EMC to denote Equilibration-Manipulation Commutability.

In this thesis, I ask the following question, which I refer to as the *Equilibrium-Causality Question*:

Question 1 (Equilibration-Causation Question) Does the EMC property hold for all causal models?

This question is important for at least the following reason: Very often in practice a causal model is first built from equilibrium relationships, and then causal reasoning is performed on that model. This common approach takes path A in Figure 1. When a manipulation is performed on a system, however, the state of the system in general becomes "shocked" taking the system out of equilibrium, a situation which is modeled by path B in Figure 1. The validity of the common approach of taking path A thus hinges on the answer to the Equilibrium-Causality Question.

$$\begin{tikzcd}
S \arrow[rr, "Equilibration"] \arrow[d, "Manipulation"] \arrow[drr, dotted, bend left=60, "A"] \arrow[dr, dotted, bend right=60, "B"'] & & \tilde{S} \arrow[d, "Manipulation"] \\
\hat{S} \arrow[r, "Equilibration"] & \tilde{\hat{S}} \arrow[r, phantom, "\overset{?}{=}"] & \hat{\tilde{S}}
\end{tikzcd}$$

Figure 1. The Equilibrium-Causality Question asks whether or not the *Do* operator commutes with the Equilibration operator operating on a dynamic causal model S.

The Equilibrium-Causation Question has implications for causal discovery from data. A very similar question can be posed in terms of the causal faithfulness condition as follows:

Question 2 (Equilibration-Causation Question 2) Does the *Equilibration* operator preserve causal faithfulness?

In other words, given a causally faithful dynamic model S, does the new model $\tilde{S}$ resulting from some equilibration of S obey causal faithfulness? This question can be viewed in terms of Figure 1: if path $S \rightarrow \tilde{S}$ leads to the only graph that is faithful to the equilibrium probability distribution, and if the manipulated equilibrium graph $\hat{\tilde{S}}$ is not equal to the true causal graph defined by $\tilde{\hat{S}}$, then $\tilde{S}$ does not obey the causal faithfulness assumption.

My hypotheses, which will be proven correct, is that the answer to both Equilibrium-Causation Questions is "No".

I will characterize sufficient conditions for when these questions are answered in the negative. I will also present examples and numerical simulations that illustrate the practical implications of these questions.

The thesis will be organized as follows: Chapter 1 will present the current state-of-the-art for modeling causality, for learning causality, and some background in dynamic systems that will be required in the remaining chapters. Chapter 2 answers both questions, presenting both empirical measurements and theorems that characterize EMC-violating systems, and Chapter 3 pulls together all results and analyzes their implications for future research in causal modeling.

CHAPTER 1
BACKGROUND CONCEPTS

In this chapter I present a review of the existing research relevant to the topics discussed in this thesis, and I define explicitly the background concepts needed. In Section 1.1, I present some general technical notation that will be used throughout this document. In Section 1.2, I present some background ideas in modeling causality, including the *Do* operator. Finally, in Section 1.3, I discuss concepts relating to temporal abstraction with causal models, including the definition of the *Equilibration* operator.

1.1 General Notation

This section summarizes some general conventions that I will be using to provide an index for the reader to refer back to. Also note that an extensive index has been provided following the Bibliography.

Sets and vectors of objects will be written in boldface type, e.g., **V** and **E**. Non-set variables will be written as capital letters such as $V \in \mathbf{V}$ and $E \in \mathbf{E}$. A specific constant value obtained by a random variable will be written in lower case and by sets will be lowercase bold, such as $V = v_0$ and $\mathbf{V} = \mathbf{v_0}$. If $\mathbf{V}$ is some set, I will use $|\mathbf{V}|$ to denote the number of elements of $\mathbf{V}$. Throughout this document I will frequently refer to both equation systems and to directed graphs. Equation systems involve sets of **E**quations while directed graphs involve sets of **E**dges. To avoid confusion, I will reserve the symbol $\{E/\mathbf{E}\}$ for equations and sets of equations, and I will refer to undirected edges with the character $\{U/\mathbf{U}\}$ and directed arcs with $\{A/\mathbf{A}\}$. Unordered sets will be denoted by curly braces $\{\ldots\}$, ordered tuples will be denoted by angle braces: $\langle\ldots\rangle$.

I will discuss several types of graphical objects throughout this document. The following notation will be prevalent:

Definition 2 (undirected edge) *If* $\mathbf{V}$ *is a set of objects, then an undirected edge* U *over* $\mathbf{V}$ *is an unordered pair* $\{X, Y\}$ *such that* $X, Y \in \mathbf{V}$*.*

Definition 3 (directed arc) *If* $\mathbf{V}$ *is a set of objects, then a directed arc* A *over* $\mathbf{V}$ *is an ordered pair* $\langle X, Y\rangle$*, usually denoted as* $(X \rightarrow Y)$*, such that* $X, Y \in \mathbf{V}$*.*

Definition 4 (directed graph) *A directed graph* G *is a pair* $G = \langle \mathbf{V}, \mathbf{A}\rangle$ *where* $\mathbf{V}$ *is a set of vertices and* $\mathbf{A}$ *is a set of directed arcs over* $\mathbf{V}$*.*

Definition 5 (partially directed graph) *A partially directed graph* G_p *is a triple* $G_p = \langle \mathbf{V}, \mathbf{U}, \mathbf{A}\rangle$*, where* $\mathbf{V}$ *is a set of vertices,* $\mathbf{U}$ *is a set of undirected edges over* $\mathbf{V}$ *and* $\mathbf{A}$ *is a set of directed arcs over* $\mathbf{V}$*.*

I will use the shorthand DiG (DAG) for "directed (acyclic) graph" and PDiG (pDAG) for"partially directed (acyclic) graph". If $G(\mathbf{V})$ is a graph, and $V \in \mathbf{V}$ then I use the following notation to indicate relationships between variables in G: $\boldsymbol{Pa}(V)$, $\boldsymbol{Ch}(V)$, $\boldsymbol{Anc}(V)$ and $\boldsymbol{Des}(V)$ denote the set of children, parents, ancestors and descendants of V in G, respectively. If E is an equation, then let $\boldsymbol{Params}(E)$ denote the set of free parameters of E. If $\mathbf{E}$ is a set of equations, let $\boldsymbol{Params}(\mathbf{E})$ denote $\bigcup_{E\in\mathbf{E}} \boldsymbol{Params}(E)$. If V is a random variable then I use $Rng(V)$ to denote the set of possible outcomes of V. If $\mathbf{V}$ is a set of random variables then I use $Rng(\mathbf{V})$ to denote the Cartesian products of the ranges of each variable $V_i \in \mathbf{V}$:

$$Rng(\mathbf{V}) = \bigotimes_{V_i \in \mathbf{V}} Rng(V_i)$$

I use the notation $(X \perp Y \mid \mathbf{Z})$ to denote the fact that a variable X is independent of a variable Y given a set of variables $\mathbf{Z}$.

1.2 Representing Causality

There are several ways that causality has been modeled in artificial intelligence, econometrics and the social sciences. These representations are closely related: all describe a set of relationships between variables in some set $\mathbf{V}$. All utilize the concept of a causal relation between variables $A, B \in \mathbf{V}$ as a directed arc $A \rightarrow B$; thus, indirectly at least, all conceptions define directed graphs of some sort. The conceptions differ as to what information constitutes the most fundamental description of a causal model; and these differences in turn impact the types of directed graphs that can be generated by the models. Here I define the representation that I will be using throughout this thesis. In Appendix A, I present an overview of the other conceptions, and I show that, given the assumptions I take for the theorems in this work, these other representations will be consistent with the definitions that I use here.

1.2.1 Causality

From a philosophical viewpoint, most would agree that causality is intertwined with the concept of *manipulation*. In fact, as I define it, *manipulation* requires the concept of *causation* and *causation* requires the concept of *manipulation*. Manipulation is the act of forcing a set of variables $\mathbf{U} \subseteq \mathbf{V}$ to a particular configuration $\mathbf{u}$ independent of the state of other variables in the system.:

Definition 6 (manipulation) *Let* $P(\mathbf{V})$ *be a joint probability distribution over a set of variables* $\mathbf{V}$*, let* $\mathbf{U}, \mathbf{W} \subseteq \mathbf{V}$ *be arbitrary disjoint subsets of* $\mathbf{V}$*. A manipulation* $\mathit{Manip}(P, \mathbf{U}, \mathbf{u})$ *of* P *is a new probability distribution* $\hat{P}(\mathbf{V})$ *such that:* $\hat{P}(\mathbf{U} = \mathbf{u}) = \hat{P}(\mathbf{U} = \mathbf{u} \mid \mathbf{W} = \mathbf{w}) = 1 \quad$ *for all* $\mathbf{w} \in \mathit{Rng}(\mathbf{W})$.

I use the notation $\mathbf{U} \hat{=} \mathbf{u}$ to indicate that $\mathbf{U}$ is being manipulated into the configuration $\mathbf{u}$ (as opposed to being observed), and I use $P(\mathbf{V} \mid \mathbf{U} \hat{=} \mathbf{u})$ as shorthand for $\mathit{Manip}(P, \mathbf{U}, \mathbf{u})$.

This definition of manipulation does not require the definition of a cause; however, it is also too general to be useful as it includes what I call *fat-hand manipulations.* A fat-hand manipulation is a manipulation that alters the conditional probability of a non-manipulated variable given its direct causes:

Definition 7 (fat-hand manipulation) *If* $\mathbf{V}$ *is a set of variables,* $\mathbf{U} \subset \mathbf{V}$*,* $\mathbf{V}' = \mathbf{V} \setminus \mathbf{U}$ *and* P *is a probability distribution over* $\mathbf{V}$*, then a manipulation* $P(\mathbf{V} \mid \mathbf{U} \hat{=} \mathbf{u})$ *is a fat-hand manipulation iff there exists a variable* $V \in \mathbf{V}'$ *and* $v \in \mathit{Rng}(V)$ *such that:*

$$P(V = v \mid \boldsymbol{Pa}(V) = \mathbf{p}) \neq P(V = v \mid \boldsymbol{Pa}(V) = \mathbf{p}, \mathbf{U} \hat{=} \mathbf{u}),$$

where $\boldsymbol{Pa}(V) \subset \mathbf{V}$ *is the set of direct causes of* V *with respect to* P *and* $\mathbf{V}$*.*

If a manipulation M is not a fat-hand manipulation, then I say that M is a *modular* manipulation:

Definition 8 (modular manipulation) *A manipulation* M *is a modular manipulation iff it is not a fat-hand manipulation.*

I define X to be a *cause* of Y iff Y depends on X when all other variables are modularly manipulated to constant values:

Definition 9 (cause) *Let* $\mathbf{V}$ *be a set of variables with* $X, Y \in \mathbf{V}$*, let* $\mathbf{U} = \mathbf{V} \setminus \{X, Y\}$*, and let* $P(\mathbf{V})$ *be a probability distribution over* $\mathbf{V}$*.* X *is a cause of* Y *with respect to* P *and* $\mathbf{V}$ *iff there exists a state* $y \in \mathit{Rng}(Y)$ *and modular manipulations* $\mathbf{U} \hat{=} \mathbf{u}$*,* $X \hat{=} x_0$ *and* $X \hat{=} x_1$ *such that:*

$$P(Y = y \mid \mathbf{U} \hat{=} \mathbf{u}, X \hat{=} x_0) \neq P(Y = y \mid \mathbf{U} \hat{=} \mathbf{u}, X \hat{=} x_1).$$

Given this definition of a cause, the concept of a *causal graph* can be defined:

Definition 10 (causal graph) *Let $\mathbf{V}$ be a set of variables, and let $P(\mathbf{V})$ be a probability distribution over $\mathbf{V}$. A causal graph with respect to P and $\mathbf{V}$ is a directed graph $G = \langle \mathbf{V}, \mathbf{A} \rangle$ such that $\{X \rightarrow Y\} \in \mathbf{A}$ iff X is a cause of Y with respect to P and $\mathbf{V}$.*

Again, these definitions of *modular manipulation* and *cause* are cyclic—each requiring the other. That does not, however, mean that these two definitions are meaningless. They provide consistency constraints on any working definitions of these two concepts: Given a system that is *a priori* determined to be causal, Definitions 6 and 8 can be used to define a modular manipulation on that system; likewise, given an operation that is *a priori* defined to be a modular manipulation, we can define a causal system using Definitions 9 and 10.

1.2.2 Causal Models

One of the first representations of causality was developed in econometrics over half a century ago. A structural equation model (SEM) is defined as a pair $\langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{E}$ is a set of simultaneous equations and $\mathbf{V}$ is a set of variables constrained by $\mathbf{E}$. In order for a set of equations and a set of variables to be causally meaningful, they must satisfy both syntactical as well as semantical constraints. By syntactical constraints, I mean conditions the equation set must satisfy that can be directly verified by examining the equations themselves; for example, the equations must specify a solution for each variable in $\mathbf{V}$. By semantic constraints, I mean that each equation is meant to represent some fundamental, invariant mechanism in the real world.

I make the following two syntactical assumptions about SEMs, letting $S = \langle \mathbf{V}, \mathbf{E} \rangle$ denote an arbitrary SEM:

Assumption 1 (fully bijective mappings) *Every $E \in \mathbf{E}$ can be written as $V = f(\mathbf{V}')$ where $V \in \mathbf{V}$, $\mathbf{V}' \subseteq \mathbf{V} \setminus \{V\}$ and f is a bijection, i.e., for any $Y \in \mathbf{V}'$ the* *function* $f(Y, \mathbf{V}'')|_{\mathbf{V}''=\mathbf{v}''}$ *is a bijection, where* $\mathbf{V}'' = \mathbf{V}' \setminus \{Y\}$ *and* $\mathbf{v}''$ *is a constant vector* $\mathbf{v}'' \in Rng(\mathbf{V}'')$.

There are two immediate consequences to this assumption. First, every equation can be solved for every variable appearing in the equation in terms of the remaining variables, e.g., $Y = f(X, Z)$ can be rewritten as $X = f^{-1}(Y, Z)$. Second, since all the functions are injections, all variables appearing in a function will be mapped to unique values, e.g. $f(x_1, z) \neq f(x_2, z)$. This class of equations includes all strictly monotonic functions (when the variables are continuous) and some discrete functions (when the variables are multinomial).

Next I assume that there are no latent confounding causes, an assumption labelled *causal sufficiency* by Spirtes *et al.* [1993]:

Assumption 2 (causal sufficiency) *Let* $\mathbf{V}$ *be a set of variables and let* P *denote a probability distribution over* $\mathbf{V}$. $\mathbf{V}$ *is causally sufficient with respect to* P *iff every common cause of any two or more variables is either in* $\mathbf{V}$ *or has a constant value.*

Finally, Simon [1953] makes the syntactical assumption that all SEMS are *self-contained*:[1]

Assumption 3 (self-contained structure) *Let* $S = \langle \mathbf{V}, \mathbf{E} \rangle$ *be an arbitrary SEM; let* $S' = \langle \mathbf{V}', \mathbf{E}' \rangle$ *be an arbitrary pair such that* $E' \subseteq E$ *and* $V' = \boldsymbol{Params}(E')$, *and let* $k \equiv |\mathbf{E}'|$ *and* $m \equiv |\mathbf{V}'|$. *The following conditions are true:*

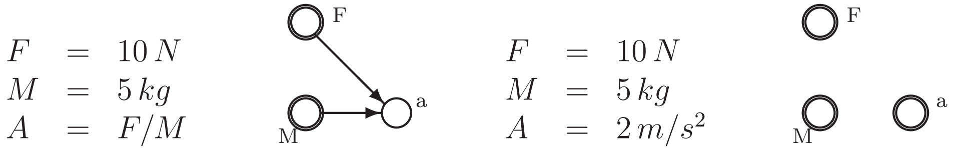
1. $m \geq k$, *and*

2. *If the values of any* $m - k$ *variables in* $\mathbf{V}'$ *are instantiated to constant values, the remaining* k *variables can be solved for unique values in terms of constants.*

[1] I use here the definition put forward by Iwasaki and Simon [1994], which is a generalization of that used by Simon [1953].

I will also refer to a set of equations $\mathbf{E}$ as being self-contained (with respect to a set of variables $\mathbf{V}$) if the SEM $S = \langle \mathbf{V}, \mathbf{E} \rangle$ is self-contained. The condition of being self-contained assures us not only that the equation set defines unique values for all variables, but that every subset of equations of size k that have been reduced to constraining k variables will also define a unique solution to those variables. I will assume that all SEMs in this thesis are self-contained.

Semantically, the equations in a SEM are meant to represent *fundamental, stable mechanisms*, or *physical laws*. As an example, the system of equations in Figure 1.1b does not represent stable mechanisms for the system of a five kilogram mass M being influenced by a ten Newton force F and possessing an acceleration A. Although both sets of equations are algebraically equivalent and entail that the value of the acceleration is two meters per second squared, the equations in Figure 1.1b are not invariant for this system because the equation $A = 2\,m/s^2$ does not in general hold for this system when either of the other two equations are changed. The property of equation invariance or fundamentality is obviously a semantical distinction rather than a syntactical property of a SEM; i.e., Figure 1.1b is a syntactically valid SEM, so when I state that the equations are not structural, I am making an assertion about the ability of the equations to explain the real-world interactions of this system.


$$\begin{aligned} F &= 10\,N \\ M &= 5\,kg \\ A &= F/M \end{aligned}$$

(a)

$$\begin{aligned} F &= 10\,N \\ M &= 5\,kg \\ A &= 2\,m/s^2 \end{aligned}$$

(b)

Figure 1.1. Two structural equation systems and the corresponding causal graphs. Although both sets of equations are algebraically equivalent, the structures obtained are different.

How to identify a stable mechanism in reality is not a straightforward matter. A debate that is of interest to this thesis occurred in the econometrics literature between

Wold [1954, 1955] and Simon [1955]. Simon and Wold debated the differences in their respective definitions of causality. In that debate, Wold seems especially adamant about allowing an "equilibrium" relationship to represent causality, although he does not precisely define what an "equilibrium" relationship is. He does not object to the use of a relationship that is independent of time, so by "equilibrium" he certainly does not mean "stationary". The objection of Wold is relevant to this thesis because I will also raise objections to the use of causal reasoning with equilibrium relationships; however, I prove that some equilibrium models support causal reasoning while others do not, and I characterize these different types of equilibrium relations.

In a SEM associating each variable with one equation in which that variable appears defines a directed graph. An example of such a mapping between equations and variables is shown in Figure 1.2. Such a graph possesses a one-to-one correspondence between vertices and variables, and may contain cycles. I term the one-to-one correspondence a *total causal mapping*:

$$\begin{array}{ll} (E_1, W): & f_1(X, V, Z, W) = 0 \\ (E_2, X): & f_2(X) = 0 \\ (E_3, Z): & f_3(X, Y, Z) = 0 \\ (E_4, Y): & f_4(Y, Z) = 0 \\ (E_5, V): & f_5(V) = 0 \end{array}$$

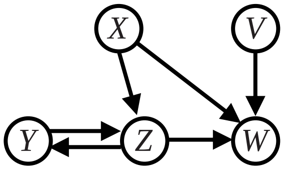


Figure 1.2. Establishing a one-to-one correspondence between equations $\mathbf{E}$ and variables $\mathbf{V}$ in an SEM defines a directed (possibly cyclic) graph where each node corresponds to a variable in $\mathbf{V}$.

Definition 11 (total causal mapping) *If* $\mathbf{E}$ *is a set of equations with* $\mathbf{V} \equiv \boldsymbol{Params}(E)$*, then a total causal mapping over* $\mathbf{E}$ *is an onto mapping* $\phi : \mathbf{V} \rightarrow \mathbf{E}$*.*

It was proven in [Nayak, 1994] that a set of independent equations is self-contained iff it possesses a total causal mapping. Again, a total causal mapping defines a directed (possibly cyclic) graph (DiG). This DiG can be constructed by the following procedure: For each association $\langle X, E\rangle$ direct an arc from $X' \rightarrow X$ for each $X' \in \boldsymbol{Params}(E)$ such that $X' \neq X$.

I define a *causal model* as follows:

Definition 12 (causal model) *A causal model is a pair $\langle S, \phi\rangle$, where $S = \langle \mathbf{V}, \mathbf{E}\rangle$ is a structural equation model and ϕ is a total causal mapping.*

A causal model is an explicit hypothesis about the detailed causal interactions between variables in the system: for each variable $V \in \mathbf{V}$, a causal model hypothesizes which variables directly affect V and the precise functional form of that affectation. In addition to the set of equations, a causal model requires additional semantic knowledge about the system, expressed as the mapping ϕ. Some researchers have developed automated methods to generate matchings between variables and equations [Serrano and Gossard, 1987; Nayak, 1994].

Causal models have been used in econometrics and the social sciences for half a decade (see Wright [1934]; Haavelmo [1943]; Strotz and Wold [1960], for example) and have strongly influenced the work of modern researchers in artificial intelligence [Pearl and Verma, 1991; Spirtes *et al.*, 2000; Pearl, 1995]. In the econometrics and social sciences these models are referred to as simply "structural equation models", and the equations are typically linear.

I define *aggregation* as an operation on a causal model that is used to reduce the *causal resolution* of the model:

Definition 13 (*Aggregation* operator) *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model with a corresponding directed graph G. An aggregation $Agg(M, X)$ of a variable $X \in \mathbf{V}$ is a new causal model $M' = \langle\langle \mathbf{V}', \mathbf{E}'\rangle, \phi'\rangle$ such that the following conditions hold:*

1. $\mathbf{V}' = \mathbf{V} \setminus \{X\}$,

2. $\phi'(X') = \phi(X')$, *for all* $X' \in \mathbf{V}'$ *such that* $X' \notin \boldsymbol{Ch}(X)|_G$.

3. *Let $\phi(X)$ be written as $X = f(\mathbf{P}_X)$, where $\mathbf{P}_X = \boldsymbol{Pa}(X)|_G$, let $X_{ch} \in \boldsymbol{Ch}(X)$ be an arbitrary child of X and let $\phi(X_{ch})$ be written as $X_{ch} = g(X, \mathbf{P})$ where $\mathbf{P} = \boldsymbol{Pa}(X_{ch}) \setminus \{X\}$. $\phi'(X_{ch})$ is the equation that results from substituting f into g for X: $X_{ch} = g(f(\mathbf{P}_X), \mathbf{P})$.*

Aggregation is a way of removing some variables from the model while preserving all causal links between the remaining variables.

If the directed graph corresponding to a causal model is acyclic then the model is called *recursive*:

Definition 14 (recursive causal model) *A causal model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ with a graph G is recursive if and only if G is acyclic.*

Different causal mappings will in general produce different directed graphs. In Figure 1.2, if we were to instead associate E_3 with Y and E_4 with Z, then X would be a parent of Y in the graph instead of a parent of Z. However, the following theorem states that if a causal model is recursive, there exists exactly one causal mapping. This in turn implies that the graph is unique:

Lemma 1 *If $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ is a recursive structural equation model then any causal mapping $\phi' : \mathbf{V} \rightarrow \mathbf{E}$ must be identical to ϕ: i.e., $\phi(V) = \phi'(V)$ for all $V \in \mathbf{V}$.*

Proof: This proof follows as a corollary to the correctness of the Causal Ordering algorithm of Simon [1953], which I prove in Appendix D. □

Uncertainty is usually modeled in causal models by allowing each equation E_i to constrain a single random variable γ_i called a *noise term*. I assume that the noise terms are independently distributed: γ_i is independent of γ_j when $i \neq j$. However, correlated error terms are also used to model a system with latent confounding variables.

I will sometimes make the assumption that the noise terms are *non-uniformly distributed*, which I define as follows:

Definition 15 (non-uniform distribution) *Let X be a random variable and let $P(X)$ be the probability distribution (or density if X is continuous) function. P is non-uniform iff there exist two values $x_1, x_2 \in Rng(X)$ such that $P(x_1) \neq P(x_2)$.*

The *Markov condition* relates a probability distribution to a directed (causal) graph:

Definition 16 (causal Markov condition) *A causal graph $G(\mathbf{V})$ over variables $\mathbf{V}$ obeys the Markov condition with respect to a probability distribution $P(\mathbf{V})$ if all variables $X \in \mathbf{V}$ are independent of their causal non-descendants given their causal parents: $(X \perp Y \mid \boldsymbol{Pa}(X)) \in \boldsymbol{Indep}(P)$ for all $Y \in \{\mathbf{V} \setminus \boldsymbol{Des}(X)_G\}$.*

A graph that obeys the Markov condition is also called an *I-map* [Pearl, 1988]. Pearl [2000] proves that the graph associated with a recursive causal model with independent error terms always satisfies the Markov condition.

The concept of *d-separation* was defined by Pearl [1988] as a graphical condition applied to a directed graph:

Definition 17 (d-separation) *Let $G = \langle \mathbf{V}, \mathbf{A} \rangle$ be a directed graph. Two variables $X, Y \in \mathbf{V}$ are d-separated given a set $\mathbf{Z} \subset \mathbf{V}$ iff, for every (undirected) path P between X and Y:*

1. *P contains a chain $A \to B \to C$ or a structure $A \leftarrow B \to C$ with $B \in \mathbf{Z}$, or*

2. *P contains a structure $A \to B \leftarrow C$ such that neither B nor a descendant of B is in $\mathbf{Z}$.*

I use the notation $(X \perp\!\!\!\perp Y \mid \mathbf{Z})|_G$ to indicate that X and Y are d-separated by $\mathbf{Z}$ in G.

The following theorem relating d-separation to conditional independence was proven also by Pearl [1988]:

Theorem 1 *Let $P(\mathbf{V})$ be a probability distribution over a set of variables $\mathbf{V}$ and $G(\mathbf{V})$ be a directed acyclic graph over $\mathbf{V}$ that obeys the Markov condition with respect to P. Then a d-separation condition $(X \perp\!\!\!\perp Y \mid \mathbf{Z})$ in G implies a corresponding conditional independence condition $(X \perp Y \mid \mathbf{Z})$ in P.*

The Markov condition between a graph G and a PDF P does not guarantee that an independence relation in P corresponds to a d-separation condition in G. This correspondence requires the *causal faithfulness* condition:

Definition 18 (causal faithfulness condition) *A directed graph $G = \langle \mathbf{V}, \mathbf{A} \rangle$ obeys the causal faithfulness condition with respect to a probability distribution $P(\mathbf{V})$ if a conditional independence relation in P implies a d-separation in G: $(X \perp Y \mid \mathbf{Z})_P \Rightarrow (X \perp\!\!\!\perp Y \mid \mathbf{Z})_G$.*

A graph that obeys faithfulness was called a *D-map* by Pearl [1988]. The term “faithfulness” was coined by Spirtes *et al.* [1993], in the context of inference of causal structure from data; the identical notion of “stability” was used in [Pearl and Verma, 1991].

1.2.3 Causal Reasoning

In the arc-cutting account of manipulation, the fundamental knowledge of a causal system consists of causal parent-child relationships: these are expressed in terms of a causal model $S = \langle \langle \mathbf{V}, \mathbf{E} \rangle, \phi \rangle$, where the function determining each variable $X \in \mathbf{V}$ is explicated by the mapping ϕ. Manipulating X is defined by replacing the equation $\phi(X)$ with a new equation, $X = x_0$, specifying the manipulated value of X. This striking of mapped equations goes at least as far back as Wold [1954] and was emphasized by Strotz and Wold [1960]. This operation corresponds to the *Do* operator[2] [Goldszmidt and Pearl, 1992; Spirtes *et al.*, 2000]:

Definition 19 (Do-operator) *If $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ is a causal model, and $\mathbf{U} \subseteq \mathbf{V}$, then $Do(M, \mathbf{U} = \mathbf{u})$ is a causal model $\hat{M} = \langle\langle \mathbf{V}, \mathbf{E}'\rangle, \phi'\rangle$, such that:*

1. *For all $V \notin \mathbf{U}$, $\phi'(V) = \phi(V)$.*
2. *If $U \in \mathbf{U}$ then $\phi'(U)$ takes the form $U = u$, where u is the component of $\mathbf{u}$ assigned to U in the manipulation.*

Spirtes *et al.* [2000] prove a theorem called the *Manipulation Theorem*, which states that given the Markov condition, it is possible to calculate the new distribution resulting from applying the *Do* operator to a recursive causal model. The axiomatizations of causal reasoning developed by Galles and Pearl [1997] and extended by Halpern [2000], treat the *Do* operator as the definition of a manipulation. However, whether or not the *Do* operator correctly corresponds to a given manipulation depends on the manipulation being considered.

The arc-cutting account of manipulation is not the only possible one. Simon [1953] argued for a model of manipulation based on his Causal Ordering Algorithm. For completeness I contrast the two approaches in Appendix B.

It is an interesting question to ask under what conditions a causal model and the *Do* operator correspond to the definitions of a *causal graph* (Definition 10) and a *modular manipulation* (Definitions 6 and 8). The following theorems show that if all equations are fully bijective and the error terms are independent and non-uniform, then the concepts of a causal model and the *Do* operator are consistent with the definition of a causal graph and a modular manipulation. To my knowledged these theorems have not been proven elsewhere:

[2] Spirtes *et al.* define a more general concept of manipulation in which the conditional distribution is changed, but the causal parents may still have an effect on the manipulated variable. The concept that I consider in this thesis corresponds to their concept of a *perfect* manipulation.

Theorem 2 *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model with independent error terms, let P be the probability distribution defined by M, and let $G = \langle \mathbf{V}, \mathbf{A}\rangle$ be the directed graph associated with M. If G is a causal graph with respect to P and $\mathbf{V}$, then the probability distribution defined by $Do(M, \mathbf{U} = \mathbf{u})$ is a modular manipulation of P.*

Proof: I need to show first that the *Do* operation is a manipulation and second that it is not a fat-hand. First, in the model $\hat{M} \equiv Do(M, \mathbf{U} = \mathbf{u})$, all variables $U \in \mathbf{U}$ are specified by an equation of the form $U = u$, where u is a constant; so $P(U = u) = 1$ independent of the state of any other variable in $\mathbf{V}$. Second, if $V \in \mathbf{V}$ and $V \notin \mathbf{U}$, then both before and after the manipulation V is specified by an equation of the form $V = f(\boldsymbol{Pa}(V), \gamma_V)$, where $\boldsymbol{Pa}(V)$ are the set of parents of V in G and γ_V is the error term. Thus, before and after the manipulation $P(V = v \mid \boldsymbol{Pa}(V) = \mathbf{p}) = P(\gamma_V = f^{-1}(v, \mathbf{p}))$ which depends only on the distribution of γ_V. Finally, if G is a causal graph then $\boldsymbol{Pa}(V)$ is the set of causal parents of V and the *Do* operator is thus not a fat-hand manipulation. □

Theorem 3 *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model with non-uniform independent error terms, $P(\mathbf{V})$ be the probability distribution defined by M and $G = \langle \mathbf{V}, \mathbf{A}\rangle$ be the directed graph associated with M. If, for all subsets $\mathbf{U} \subseteq \mathbf{V}$, the probability distribution defined by $Do(M, \mathbf{U} = \mathbf{u})$ is a modular manipulation of P, then G is a causal graph with respect to P and $\mathbf{V}$.*

Proof: $\Rightarrow$ If $\{X \to Y\} \in A$ then the equation determining Y takes the form $Y = f(X, \mathbf{P}, \gamma_Y)$, where $\mathbf{P} \equiv \boldsymbol{Pa}(Y) \setminus \{X\}$ and γ_Y is an error term. Let $\mathbf{V}' \equiv \mathbf{V} \setminus \{X, Y, \mathbf{P}\}$. Let $y \in Rng(Y)$, $x_i \in Rng(X)$, $\mathbf{p} \in Rng(\mathbf{P})$ and $\mathbf{v}' \in Rng(\mathbf{V}')$ be arbitrary values, then $Do(M, \{X = x_i, \mathbf{P} = \mathbf{p}, \mathbf{V}' = \mathbf{v}'\})$ defines a distribution $P(Y = y \mid X \hat{=} x_i, \mathbf{P} \hat{=} \mathbf{p}) = P(\gamma_Y = f^{-1}(y, x_i, \mathbf{p}))$. Since γ_Y is non-uniform, there exists two values g_1, g_2 such that $P(\gamma_Y = g_1) \neq P(\gamma_Y = g_2)$. Define values $x_1, x_2 \in Rng(X)$ such that $x_1 = f^{-1}(y, g_1, \mathbf{p})$ and $x_2 = f^{-1}(y, g_2, \mathbf{p})$. Since f is a bijection x_1 and x_2 exist and $x_1 \neq x_2$. Then $P(Y = y \mid X \hat{=} x_1, \mathbf{P} \hat{=} \mathbf{p}) \neq P(Y = y \mid X \hat{=} x_2, \mathbf{P} \hat{=} \mathbf{p})$

and X is a cause of Y with respect to P and $\mathbf{V}$.

$\Leftarrow$ Assume X is a cause of Y with respect to P and $\mathbf{V}$. Let $\mathbf{P} \equiv \boldsymbol{Pa}(Y) \setminus \{X\}$ and $\mathbf{V}' \equiv \mathbf{V} \setminus \{X, Y, \mathbf{P}\}$. The equation determining Y either takes the form $Y = f(\mathbf{P}, \gamma_Y)$ or $Y = f(X, \mathbf{P}, \gamma_Y)$; I need to show that it in fact takes the form $Y = f(X, \mathbf{P}, \gamma_Y)$. Since X is a cause of Y, there exist values $y \in Rng(Y)$, $x_i \in Rng(X)$, $\mathbf{p} \in Rng(\mathbf{P})$ and $\mathbf{v}' \in Rng(\mathbf{V}')$ such that $P(Y = y \mid X \hat{=} x_1, \mathbf{P} \hat{=} \mathbf{p}, \mathbf{V}' \hat{=} \mathbf{v}') \neq P(Y = y \mid X \hat{=} x_2, \mathbf{P} \hat{=} \mathbf{p}, \mathbf{V}' \hat{=} \mathbf{v}')$. Assuming $Y = f(\mathbf{P}, \gamma_Y)$, this equation can be rewritten in terms of γ_Y: $P(\gamma_Y = f^{-1}(y, \mathbf{p})) \neq P(\gamma_Y = f^{-1}(y, \mathbf{p}))$, which is a contradiction. Therefore it must be the case that $Y = f(X, \mathbf{P}, \gamma_Y)$ and X is a parent of Y in G. □

1.3 Temporal Abstraction

The formalism presented in Section 1.2 allows the concept of causality to be discussed at all levels of abstraction and is thus quite general. In particular, it is trivial to approximate a time-dependent variable X simply by adding one variable for X at each discrete time slice:

$$X \rightsquigarrow \{X^{(0)}, X^{(1)}, \ldots, X^{(n)}\}.$$

The model itself must of course specify the functional relations and probability distributions between all variables in the model, including those at various time slices. In this way, a time-dependent causal model is just a special-case of a causal model in general, and need not be treated any differently.

It has long been recognized that the causal graph of a system can depend on the time-scale at which the system is observed. For example, it is well established that a non-recursive system, when modeled over a shorter time scale, can be transformed into a recursive one [Bentzel and Hansen, 1954]. Strotz and Wold [1960] provide an example of such a system: an aquarium with multiple populations of fish competing for resources. In this toy model there are two types of fish, big and small, occurring in quantities Y_b and Y_s, respectively, and two populations of weeds W_b and W_s, occurring in quantities X_b and X_s, respectively. The big fish feed only on the small fish and on weed W_b; while the small fish feed only on weed W_s. The linear structural equation model that they use to describe this system on short time-scales is as follows:

$$\begin{aligned} Y_b(t) &= \alpha_1 + \beta_1 Y_s(t - \Delta t) + \gamma_1 X_b(t) + u_1(t) && (1.1) \\ Y_s(t) &= \alpha_2 + \beta_2 Y_b(t - \Delta t) + \gamma_2 X_s(t) + u_2(t) && (1.2) \end{aligned}$$

where Δt is a constant time lag, the u_i are independent random variables, and all other variables are constant non-zero coefficients. This model is a causal model with Equation 1.1 being associated with variable Y_b and Equation 1.2 being associated with Y_s, and it can be represented as an acyclic time-dependent graph with Y_b and Y_s varying in time on a time scale of Δt (assumed to be short).

If we make the assumption that the fish populations in the aquarium are in equilibrium (thus looking at the system on a longer time-scale), then $Y_b(t - \Delta t) = Y_b(t)$ and $Y_s(t - \Delta t) = Y_s(t)$, thus Equations 1.1 and 1.2 can be written as:

$$\begin{aligned} Y_b(t) &= \alpha_1 + \beta_1 Y_s(t) + \gamma_1 X_b(t) + u_1(t) && (1.3) \\ Y_s(t) &= \alpha_2 + \beta_2 Y_b(t) + \gamma_2 X_s(t) + u_2(t) && (1.4) \end{aligned}$$

which is a strongly coupled set and can only be represented by a cyclic graph. Thus a cyclic equilibrium causal graph is typically interpreted as arising from a system possessing feedback when viewed at shorter time-scales; although there is no syntactic reason why a cyclic causal graph must be based on such a system, i.e., the syntax of a nonrecursive model does not give us any indication of what a cycle means or how it came about in a system.

There has been other work relating dynamic models to non-dynamic models in general: Fisher [1970] discusses the implications of the fact that in SEMs, causality is assumed to be instantaneously occurring, when in reality a cause always requires an implicit time-lag to produce effects. He derives necessary conditions for an SEM that a set of variables, being defined as the averages over a short time-lag, will obey the same relationships as their instantaneous counterparts. Kuipers [1987] discusses temporal abstraction in dynamic qualitative models with widely varying time-scales; and Richardson [1996] discusses the signatures that dynamic models with feedback display in their non-recursive equilibrium counterparts.

1.3.1 Dynamic Causal Ordering

Iwasaki and Simon [1994] discussed dynamic causal models, defined as sets of differential equations, and introduced various operations that can be used to convert those dynamic models into static models or models with mixed equilibrium and dynamic relationships. In so doing, they presented examples which show that it is possible for an *acyclic* equilibrium causal graph, defined by an SEM, to be produced by a dynamic system with feedback. Their formulation used the representation of SEMs along with the COA, and they considered differential equation systems whereby if a variable X is changing in time, then there exists some mechanism that includes the *differential* of X: dX/dt (or the difference ΔX for a discretized time interval). Models of this form are convenient because they allow for simple analysis of the systems as some dynamic variables achieve equilibrium. I will use the standard notation $\dot{V}$ to denote dV/dt and $V^{(i)}$ to denote the ith derivative of V: $V^{(i)} \equiv d^iV/dT^i$.

I will illustrate these types of models by presenting the equilibrium and dynamic causal models of the non-damped simple-harmonic oscillator system (i.e., a mass dangling from a spring), depicted schematically in Figure 1.3. In the equilibrium system there are two physical laws at play. First, the gravitational force F_g depends

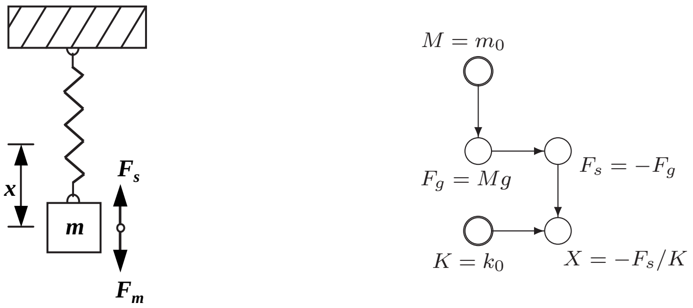


Figure 1.3. A non-damped simple-harmonic oscillator and its equilibrium causal model.

on the mass M and the gravitational constant g; and second, the spring force is proportional and opposite to the displacement X:

$$\begin{aligned} F_g =& Mg \qquad (1.5) \\ F_s =& -KX. \qquad (1.6) \end{aligned}$$

The mass M and spring constant (stiffness) K are set to constant values:

$$\begin{aligned} M =& m_0 \qquad (1.7) \\ K =& k_0. \qquad (1.8) \end{aligned}$$

Assuming the block is in equilibrium, the force of gravity must be equal and opposite to the force of the spring:

$$F_s = -F_g, \tag{1.9}$$

yielding the recursive equilibrium causal model shown in Figure 1.3 (with independent error terms being left implicit). Since I have not explicitly modelled a dissipative force in this system, oscillations will never "damp out" causing the system to go from a non-equilibrium state to an equilibrium state. Thus if this system is in an equilibrium state, it must have been put in that state from the start.

The dynamic causal model of this system has the same mechanisms at play, but it does not make the assumption that the block is in equilibrium; thus, Equation 1.9

is replaced with Newton's second law for the acceleration A of a mass M under forces in one dimension:

$$\Sigma_i F_i = MA, \tag{1.10}$$

where the sum is over the set $\{F_g,\ F_s\}$. In addition, the definitions of acceleration and velocity must be added to make the system self-contained (expressed in discrete form):

$$A_{(t)} \approx \frac{V_{(t+1)} - V_{(t)}}{\Delta t} \tag{1.11}$$

$$V_{(t)} \approx \frac{X_{(t+1)} - X_{(t)}}{\Delta t} \tag{1.12}$$

where $X_{(t)}$ refers to the value of variable X at time slice t, and Δt is the (constant) time between slices. These can be rewritten as the recurrence relations:

$$V_{(t+1)} = V_{(t)} + A_{(t)}\Delta t \tag{1.13}$$

$$X_{(t+1)} = X_{(t)} + V_{(t)}\Delta t \tag{1.14}$$

In order to specify a particular solution to these difference equations, we need to specify initial conditions for X: $X_{(0)} = x_0$ and for V: $V_{(0)} = v_0$, for some constants x_0 and v_0. Finally, we need to state which variables are exogenous, in this case: $M_{(t)} = m_0$, $g_{(t)} = g_0$, and $K_{(t)} = k_0$, for all t.

The recursive graph for this set of equations is shown in Figure 1.4 (a). This graph relates all the variables in our model at $t = 0$ with each other and with V and X at $t = 1$. Since $X_{(1)}$ and $V_{(1)}$ are now determined at $t = 1$, we can recursively iterate this procedure to generate causal graphs for arbitrary values of t.

Since this graph is based on continuous differential equations, all causation across time slices will only occur to a variable from its derivative, a proposal made by Simon

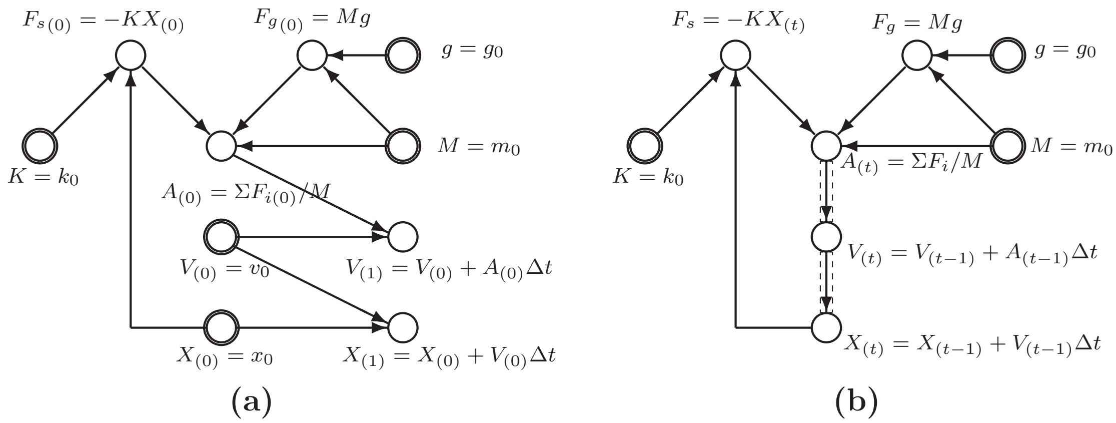

$F_{s(0)} = -KX_{(0)}$
$F_{g(0)} = Mg$
$g = g_0$
$K = k_0$
$M = m_0$
$A_{(0)} = \Sigma F_{i(0)}/M$
$V_{(0)} = v_0$
$V_{(1)} = V_{(0)} + A_{(0)}\Delta t$
$X_{(0)} = x_0$
$X_{(1)} = X_{(0)} + V_{(0)}\Delta t$
(a)
$F_s = -KX_{(t)}$
$F_g = Mg$
$g = g_0$
$K = k_0$
$A_{(t)} = \Sigma F_i/M$
$M = m_0$
$V_{(t)} = V_{(t-1)} + A_{(t-1)}\Delta t$
$X_{(t)} = X_{(t-1)} + V_{(t-1)}\Delta t$
(b)

Figure 1.4. (a) The first two time-slices of the dynamic causal model for the simple harmonic oscillator system. (b) A shorthand graph for the same system.

and Rescher [1966]. Thus, this graph is Markovian through time i.e., the variables in the future are d-separated from variables in the past by variables in the present, and it can be represented by a convenient shorthand graph for an infinite sequence of time steps. In this shorthand graph, depicted in Figure 1.4 (b), temporal subscripts have been dropped and special dashed links, labelled by Iwasaki and Simon as *integration links*, have been created to denote that a causal relationship is really occurring through a time slice.

The shorthand dynamic graph in Figure 1.4 adds some confusion to the concept of recursivity, since it possesses cycles itself although it really is meant to represent an acyclic graph. The following theorem relates an unrolled recursive graph to the shorthand graph:

Theorem 4 (recursive causal model) *A causal model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ with a shorthand causal graph G is recursive if and only if the causal graph $G_x^{(0)}$, obtained by removing all integration links from G, is acyclic.*

Proof: $\Rightarrow$ Assume that $G_x^{(0)}$ is acyclic. Then for a fixed time slice the causal graph is acyclic. Assuming there are n links from slice i to slice $i+1$ and that no variable at slice $i+1$ is an ancestor of a variable at slice i, then we can add another arc from slice i to $i+1$ without creating any cycles, thus the rolled out graph must be acyclic. Finally when there are no arcs from slice i to $i+1$ then there exist no ancestors in slice $i+1$ of any variables in slice i. Thus, by induction, the rolled-out graph will be acyclic.

$\Leftarrow$ Assume the unrolled graph is acyclic, then dropping all arcs across time slices will leave an acyclic graph at each time slice. $\square$

A *self-contained dynamic structure* was defined by Iwasaki and Simon [1994] essentially as a set of well-defined n first-order differential equations for n variables:

Definition 20 (self-contained dynamic structure) *Let S be a pair $S = \langle \mathbf{V}, \mathbf{E} \rangle$ where $\mathbf{V}$ is a set of variables and $\mathbf{E}$ is a set of equations such that $|\mathbf{V}| = |\mathbf{E}| = n$. Let $\mathbf{E}' \subseteq \mathbf{E}$ be an arbitrary subset with $k = |\mathbf{E}'|$ and r equal to the number of first derivatives contained in $\boldsymbol{Params}(\mathbf{E}')$. Then S is a self-contained dynamic structure iff:*

1. *$r \geq k$*
2. *If the values of any $(r-k)$ first derivatives are chosen arbitrarily, then the remaining k are determined uniquely as a function of the chosen variables.*

Each equation in a dynamic model so defined is a differential equation for some $V \in \mathbf{V}$. The restriction to first-order equations is general because a single nth-order differential equation can be converted into a set of n first-order equations by defining each first-derivative as a new variable.

Iwasaki and Simon [1994] generalize dynamic models to *mixed structures* by allowing both differential and equilibrium equations. Let $M = \langle \mathbf{V}, \mathbf{E} \rangle$ be a pair where $\mathbf{E}$ is a set of equations (some differential, some non-differential) constraining the set of variables $\mathbf{V}$. Then define *Inst*(M) to be the superset of equations consisting of E plus one constant equation of the form $V = v_0$, where v_0 is a constant value for each dynamic variable V. A *self-contained mixed structure* is then defined as (in verbatim from Iwasaki and Simon [1994]):

Definition 21 (self-contained mixed structure) *The set E of n equations for n variables is a self-contained mixed structure iff:*

1. *Zero or more of the n equations are first-order differential equations and the rest are equilibrium equations.*
2. *Inst(M) is a self-contained equilibrium structure when the variables and their derivatives are treated as distinct variables.*

Thus the model illustrated in Figure 1.4 is in fact a mixed-model according to Iwasaki and Simon.

The "mixed-model" definition makes an unnatural distinction between a first-order differential equation and an "equilibrium" equation. The distinction is not wholly consistent with their treatments of derivatives-as-variables. The restriction of a dynamic model to containing only first-order differential equations was justified on the grounds that any n-th order differential equation could be readily replaced by n first-order equations over n variables by treating derivatives as normal variables. Implicit in this argument is that derivatives should be treated at the same level as regular variables. If an "equilibrium" equation is defined as an equation that is mapped to a non-dynamic variable (a variable with no derivative in the model), then a first-order differential equation for variable X is equivalent to an equilibrium equation for $\dot{X}$. I therefore simplify this notation and refer to any causal model as a dynamic model if it contains a dynamic variable:

Definition 22 (dynamic variable) *Given a causal model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$, a variable $V \in \mathbf{V}$ is a dynamic variable if and only if $\dot{V} \in \mathbf{V}$.*

Definition 23 (dynamic causal model) *A causal model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ is a dynamic causal model with respect to $V \in \mathbf{V}$ if and only if V is a dynamic variable.*

Using this definition, the equations in a dynamic model must include the integration equations relating a variable to its derivative as an integral over time in order for the differential equations to specify a self-contained system. These integration equations also include the initial conditions required for the $Inst(M)$ model defined by Iwasaki and Simon [1994]. I will use the term *differential model* to emphasize that a model is meant to represent relations that hold on an infinitesimal time-interval, i.e., a model that is made up wholly of either continuous (not discretized) differential equations and instantaneous relations. I use the term *differential graph* to denote the directed graph of a dynamic model with all integration links removed.

Applying the $Do(M, \mathbf{U} = \mathbf{u})$ operator to a dynamic model M means setting the variables $\mathbf{U}$ to the configuration $\mathbf{u}$ for all time. An example of manipulating a dynamic variable Y in a causal graph is shown in Figure 1.5. Since Y is a dynamic variable,

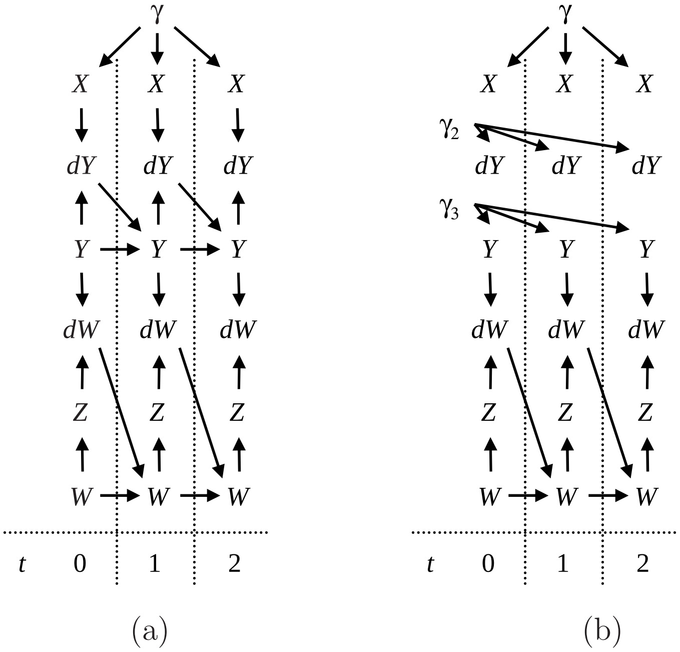


Figure 1.5. Applying the *Do* operator to a dynamic variable Y in a causal model. Y and all of Y's derivatives must be set to the same value for all time, represented in the graph as latent variables γ_2 and γ_3.

in order to set its value for all time, we must also set Y's derivatives to zero for all time. Manipulating dY and Y for all time is modelled with common causes γ_2 and γ_3 affecting these variables in each time slice. This action on the unrolled graph corresponds to simple arc-cutting on the shorthand dynamic graph.

1.3.2 Equilibrium Models from Dynamic Models

Iwasaki and Simon [1994] studied the ability to model dynamic systems on many different time scales and the ability to derive different causal models as the observation time-scale changes. The dynamic graph in Figure 1.4 represents the causal graph for the simple harmonic oscillator system modeled over an infinitesimal time scale. Alternatively, this system could be modeled over the time-scale necessary for the system to achieve equilibrium. The equilibrium model can be derived from the dynamic model in Figure 1.4 by assuming that the mass has come to equilibrium, which implies that both the velocity and the acceleration of the mass are zero: $A = 0$ and $V = 0$. Substituting these two constraints into the equations and eliminating A and V from the model yields a new set of equations which corresponds to the original equilibrium causal model shown in Figure 1.3.

By saying that the causal graph can be different for different time scales, I mean two things: (1) the COA of Simon (discussed in Appendices A and D) produces different causal structures, and (2) if uncertainty is added to the model in the form of independent error terms, then the Markov condition can be violated when the probability distribution of the system in equilibrium is compared to the dynamic causal graph. For example, in Figure 1.4, the Markov condition entails that F_g and F_s are marginally independent; however, if independent error terms γ_a are added to each equation, then in the equilibrium probability distribution this independence relation does not hold, since $F_s = -F_g + \gamma_a$ in equilibrium.

The operation of *equilibration* was presented in Iwasaki and Simon [1994], whereby the derivatives of a dynamic variable X are eliminated from a model by assuming that X has achieved equilibrium. Intuitively this operation can be sketched as follows:

Definition 24 (equilibration sketch) *Given a causal model with dynamic variable X, do the following:*

1. *Assume all derivatives of X are zero, and remove them from the model.*

2. *Remove all integration equations for X or derivatives of X from the model.*

3. *Alter all remaining equations by substituting zero for all derivatives.*

4. *Construct a new causal mapping ϕ.*

Before defining equilibration formally I will introduce some definitions. I use the notation $\boldsymbol{V_{del}}(X)$ and $\boldsymbol{E_{del}}(X)$ to denote the variables and equations that are deleted from a SEM due to equilibration:

Definition 25 ($\boldsymbol{V_{del}}(X)$, $\boldsymbol{E_{del}}(X)$) *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model with $X \in \mathbf{V}$ and with $X^{(n)} \in \mathbf{V}$ the highest order derivative of X in $\mathbf{V}$, then:*

$$\boldsymbol{V_{del}}(X) = \{X^{(i)} \mid 0 < i \leq n\} \text{ and}$$

$$\boldsymbol{E_{del}}(X) = \{\phi(X^{(i)}) \mid 0 \leq i < n\}$$

Note that $X \notin \boldsymbol{V_{del}}(X)$ and $\phi(X^{(n)}) \notin \boldsymbol{E_{del}}(X)$. I define equilibration as follows:[3]

Definition 26 (equilibration) *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model with $X \in \mathbf{V}$ and with $X^{(n)} \in \mathbf{V}$ the highest order derivative of $X \in \mathbf{V}$. The model $M_{\tilde{x}} = \langle\langle \mathbf{V}_{\tilde{\mathbf{x}}}, \mathbf{E}_{\tilde{\mathbf{x}}}\rangle, \phi_{\tilde{x}}\rangle$ due to the equilibration of X is obtained by the following procedure:*

[3] This definition differs slightly from Iwasaki and Simon [1994] in how it handles higher-order derivatives. Their version required n equilibration operations to equilibrate X when $X^{(n)}$ is the highest order derivative present in the model; whereas my definition allows X to be equilibrated with a single operation.

1. *Let* $\mathbf{V}_{\tilde{\mathbf{x}}} = \mathbf{V} \setminus \boldsymbol{V_{del}}(X)$*,*

2. *Let* $\mathbf{E}_{\tilde{\mathbf{x}}} = \mathbf{E} \setminus \boldsymbol{E_{del}}(X)$*,*

3. *For each* $E \in \mathbf{E}_{\tilde{\mathbf{x}}}$ *set* $V = 0$ *for all* $V \in \boldsymbol{V_{del}}(X)$*.*

4. *Construct a new mapping* $\phi_{\tilde{x}} : \mathbf{V}_{\tilde{\mathbf{x}}} \rightarrow \mathbf{E}_{\tilde{\mathbf{x}}}$*.*

An example of applying this operator to the damped simple-harmonic oscillator dynamic graph is shown in Figure 1.6. This system is identical to the non-damped simple-harmonic oscillator of Figure 1.4, except here a damping force F_d which is proportional to the negative of the velocity V has been added (this term is necessary to ensure that the system will have a stable equilibrium; this point is discussed in Section 1.3.3). The resulting equilibrium model is identical to the equilibrium model of the simple-harmonic system of Figure 1.3, and is the only mapping possible by Lemma 1 (Page 15). Note that after an equilibration operation, the new mapping $\phi_{\tilde{x}}$ may be completely different from the original mapping ϕ. In Figure 1.6 variables F_s and X are both mapped to different equations than they were mapped to in the dynamic model.

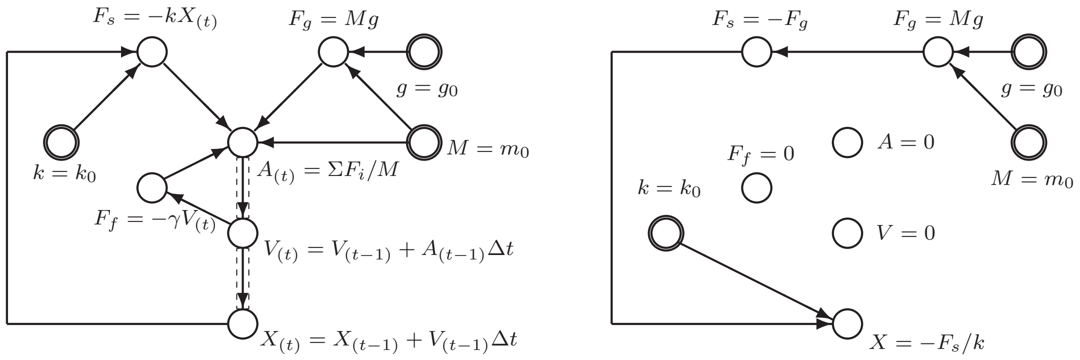


Figure 1.6. Applying the Equilibration operator to the damped simple harmonic oscillator system.

Iwasaki and Simon [1994] define a *self-regulating mechanism* as one through which a variable causes its derivative:[4]

Definition 27 (self-regulating mechanism) *Let $S = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model. An equation $E \in \mathbf{E}$ is a self-regulating mechanism (with respect to S) for variable $V \in \mathbf{V}$ if $V \in$ **Params**(E) and $\phi(V^{(n)}) = E$, where $V^{(n)}$ is the highest-order derivative of V in $\mathbf{V}$.*

If E is a self-regulating mechanism for V then I say that V is a *self-regulating variable.* A mechanism that is self-regulating for V causes an arc $V \rightarrow \mathbf{V}^{(n)}$ to be present in the dynamic causal graph. To eliminate non-general problems that may arise due to particular characteristics of equations in an SEM, Iwasaki and Simon also define a A *qualitative* self-contained structure as a self-contained SEM where each equation is only qualitatively specified. That is, each equation only specifies which variables are constrained by an equation E, but does not specify the actual functional form of the relationship. They show that if S is a qualitative SEM and E is a self-regulating mechanism for V, then equilibrating V will always result in a qualitative self-contained structure.

The *Equilibration* operator from Definition 26 can cause the remaining set of equations to be non-self-contained. For example, setting variables to zero could cause two equations that were initially independent to become dependent, or it could cause the system to be over-constrained if some variable drops out of an equation. I call equilibration *well-defined* if these do not happen:

Definition 28 (well-defined equilibration) *If $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ is a causal model and $\mathbf{V}_{\tilde{\mathbf{x}}}$ and $\mathbf{E}_{\tilde{\mathbf{x}}}$ are the respective variables and equations that result when variable*

[4] This is slightly different from the definition of Iwasaki and Simon [1994] to account for the difference in my definition of equilibration.

$X \in \mathbf{V}$ is equilibrated, then I say that equilibrating X is well-defined iff $\langle \mathbf{V}_{\tilde{\mathbf{x}}}, \mathbf{E}_{\tilde{\mathbf{x}}} \rangle$ is a self-contained SEM.

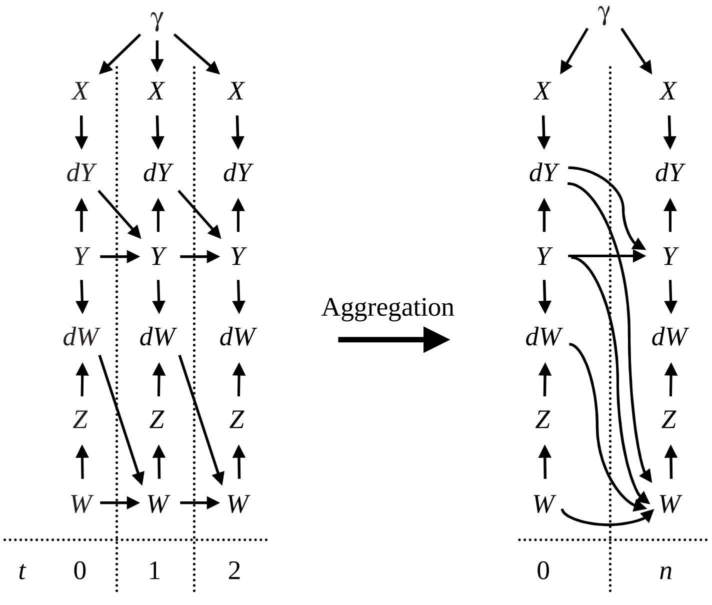
I define an equilibrium model any model that is not a dynamic model:

Definition 29 (equilibrium model) *A causal model $M = \langle \langle \mathbf{V}, \mathbf{E} \rangle, \phi \rangle$ is an equilibrium model with respect to X for some $X \in \mathbf{V}$ if and only if X is not a dynamic variable in M.*

An SEM $M = \langle \mathbf{V}, \mathbf{E} \rangle$ is an equilibrium model if and only if it is not a dynamic model.

I also define an *equilibrated model* as an equilibrium model that is derived from a dynamic model by equilibrating one or more variables in a dynamic model:

Definition 30 (equilibrated model) *Let $M = \langle \langle \mathbf{V}, \mathbf{E} \rangle, \phi \rangle$ be a causal model and let $X \in \mathbf{V}$ be a dynamic variable. A causal model $M_{\tilde{x}} = \langle \langle \mathbf{V}_{\tilde{\mathbf{x}}}, \mathbf{E}_{\tilde{\mathbf{x}}} \rangle, \phi_{\tilde{x}} \rangle$ is an equilibrated model with respect to X and M if and only if $M_{\tilde{x}}$ is the model that results by performing a well-defined equilibration on X in M.*



(a)

(b)

Figure 1.7. The dynamic graph rolled out to the n-th time slice with all intermediate time-slices aggregated out.

It should be emphasized that the causal structure obtained by performing an equilibration, although it is meant to represent the relations of a system observed over a long time scale, is not necessarily the same as the unrolled dynamic graph aggregated out to an arbitrary time scale n. An example of the latter graph is shown in Figure 1.7. Figure 1.7-a shows the first three time slices of the full rolled-out dynamic graph. Note that X is assumed to be manipulated to a constant value at all time. This is reflected by including a common latent cause γ into the model. Figure 1.7-b shows the same graph with the intermediate time slices aggregated out to reveal causal relations between the initial time slice and the final time slice. If we were to further aggregate out all derivative variables we get the structure of Figure 1.8-a. Finally, by using the independence relations between just the variables at slice n of this graph (taking into account the latent common cause γ), and using the background knowledge that X is exogenous, we would recover the independence graph shown in Figure 1.8-b. This structure is different from that obtained by equilibrating Y and W (Figure 1.9-a). The fact that these two graphs are different reflect the fact that the occurrence of equilibrium adds a new relation to the causal model that can possibly

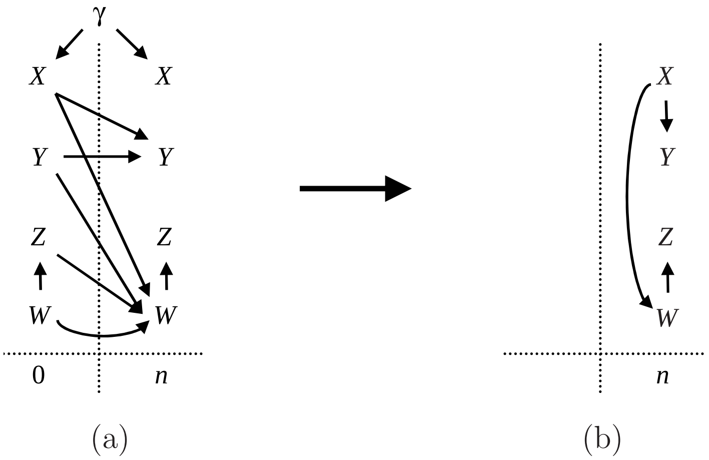


Figure 1.8. (a) The rolled-out aggregated graph with all derivatives further aggregated, and (b) the independence structure between variables just at the n-th time slice, assuming that X is exogenous.

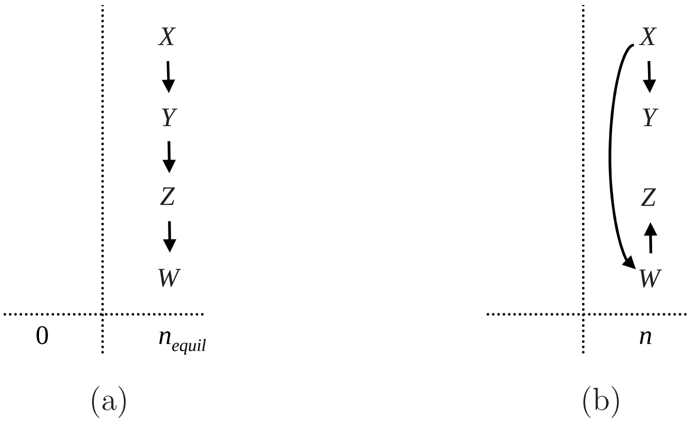


Figure 1.9. The equilibrium graph (a) is not the same as the rolled-out graph with intermediate and differential nodes aggregated out (b).

dramatically alter the causal ordering. Consider the following method of generating long-time-scale data using the dynamic model:

1. Starting with random initial conditions for the dynamic model in Figure 1.7-a. Simulate the temporal evolution of the system for n steps using the dynamic model.

2. After n steps, take a snapshot of all variables and save that as one record in a database of values.

3. Repeat 1 and 2 until N records have been generated.

If the database so generated were input into a causal discovery program, for arbitrary n we would expect to recover the graph in Figure 1.9-b. On the other hand if n were long enough for both Y and W to achieve equilibrium, then a shift in the dependence relations will occur, and the graph of Figure 1.9-a will be discovered. This fact is verified empirically in Section 2.2.

As already defined, in an equilibrium SEM, under the assumption of no latent variables, noise is modeled by allowing each mechanism E_i to include an independent error term γ_i: random variables whose distributions in turn define the joint probability distribution over all observed variables. When considering a non-deterministic dynamic causal model, however, there is an issue of how to model noise. One point is that an integral equation (e.g., $X_t = X_{t-1} + \dot{X}_t \cdot \Delta t$), essentially the definition of a derivative $\dot{X}$, should not include a noise term because it is a mathematical identity. Another, less trivial issue, concerns the assumption about time-dependence of the noise terms themselves. If a dynamic model is to be interpreted literally as a static model with a copy of each variable made at each time slice, then in general one would expect to specify a noise term for each variable in the model, and thus a complete set for each time slice. Such a noise model assumes that the noise terms themselves are changing in time. An alternative model would give all variables in time-slice $t = 0$ a noise term, and at subsequent times allow the system to evolve deterministically.

Which model is most appropriate depends on the system and on how fast the noise term is changing. Consider for example the equation for the value of the mass M in the simple-harmonic oscillator system of Figure 1.4 (Page 25). A possible source of noise for the mass is that we have a selection of different bodies that have been weighed and sorted into bins. When we pick a certain mass from a bin, it may only be approximately equal to the value assigned to the bin, displaying some random variation which is specified by the distribution of the noise term for M. However, once a mass is chosen, the noise term will be constant from one time-slice to another. On the other hand, the value of the spring constant K may be temperature dependent, so if the system is submerged in a rapidly varying temperature environment, the noise term for K may be best modeled as time-dependent.

When discussing issues of learning causal models of a dynamic system, I will make the simplifying assumption that the noise terms are very slowly changing compared to the time-step Δt. This limit is best modeled with noise terms at $t = 0$ and deterministic propagation through time.

1.3.3 Dynamic Stability

The subject of dynamic stability is an old one and is widely-studied. In this section I introduce some terminology and state some standard results which will be relevant to causal reasoning in equilibrium models. A first-order differential equation $\dot{X} = f(X)$, denoted as E, defines a family of solutions for the path of X for all time and for all initial conditions of X. Given a particular initial condition $X = x_0$, where x_0 is a constant, E defines a particular trajectory of X through time. A *fixed point* of E is a solution $X = x_f$ such that X will remain fixed at for all time; thus it is a point at which $\dot{X} = 0$ for all time. A fixed point is *stable* if all sufficiently small disturbances away from equilibrium damp out over time and cause the variable to return to the fixed point.

A common geometrical interpretation of stability and fixed points is obtained by viewing the derivative $\dot{X}$ as a function of X, defined by the differential equation E (assuming E can be written in the form $\dot{X} = f(X)$). Figure 1.10 illustrates the use of this geometrical technique. In Figure 1.10a, a single solution path is shown for which there exist two fixed-points, labelled X_s and X_u. The point X_s is stable because when $X = X_s$, a slight perturbation δ in the positive (negative) direction will cause

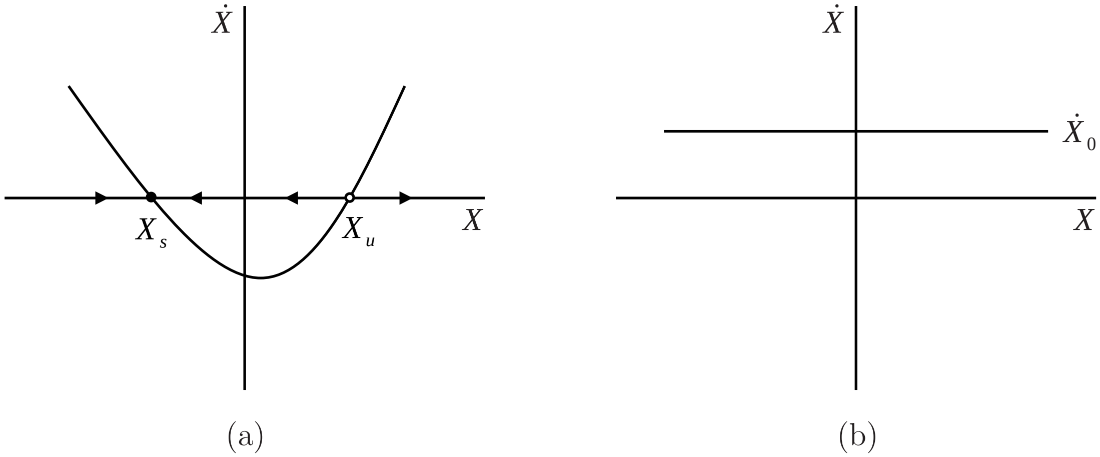


Figure 1.10. A geometrical interpretation of fixed-points and stability.

a negative (positive) velocity to arise which tends to push X back to X_s. On the contrary, the point X_u is an unstable fixed-point, because, although $\dot{X}$ is zero exactly at $X = X_u$, a perturbation in the positive (negative) direction will cause a positive (negative) velocity to arise, pushing X even further away from X_u.

There exist many classifications of stability. A fixed-point can be either *locally* or *globally* stable (meaning stable under small or large perturbations, respectively), or *bistable* (meaning stable in one direction but not the other). This thesis is interested in all of these types of stability. Figure 1.10a suggests the following sufficient conditions for local stability in a dynamic system:[5]

Theorem 5 (stability condition) *Let X be a dynamic variable having a fixed-point solution at $X = x_0$. Then x_0 is locally stable if*

$$\left.\frac{\partial \dot{X}}{\partial X}\right|_{x_0} < 0$$

where $\dot{X}$ is the time-derivative of X.

For the case when $\left.\partial \dot{X}/\partial X\right|_{x_0} = 0$, in general nothing can be said about stability of a fixed point at x_0; these situations must be handled on a case-by-case basis. However, when $\partial \dot{X}/\partial X = 0$ for all values of X as in Figure 1.10b, it is obvious that no fixed point exists unless $\dot{X} = 0$ for all time, in which case all values of X represent a fixed point that is neither stable nor unstable. Thus, it is clear that a necessary condition for a stable fixed-point to exist is that $\dot{X}$ be a nonzero function of X. The function $f(X)$ also defines a characteristic time scale for X to reach stability. The quantity $f' \equiv \partial f/\partial X$ has units of inverse time, and $1/f'$ defines the characteristic time-scale for X to converge.

[5] See [Strogatz, 1991] for example.

Samuelson [1947] made the connection between comparative statics and dynamics in the context of econometric models by deriving the *Correspondence Principle*. Although this analysis was not made on causal models, it nonetheless made arguments similar to those presented in this thesis. Namely, Samuelson argued that in order to draw meaningful conclusions from a static model it is necessary to have knowledge about dynamics. In practice, invoking the Correspondence Principle in econometrics amounts to assuming that the dynamic system is in equilibrium so that a static analysis can be brought to bear.

CHAPTER 2

RESULTS

In this chapter I discuss empirical and theoretical results which answer the Equilibrium-Causation Questions. First, in Section 2.1, I present an example of a real-world system that does not obey the EMC property, thus proving my hypothesis correct. In Section 2.2 I present simulation studies that illustrate that the Equilibrium-Causation Question 2 can also be answered in the negative. Finally, in Section 2.3 I characterize certain models that are guaranteed to violate and those that are guaranteed to obey the EMC property.

2.1 Motivating Example: the Ideal Gas System

Here I provide a real-world example showing that the *causal resolution* of a model can determine whether or not the *Do* operator commutes with the *Equilibration* operator. Consider in Figure 2.1 the example of an ideal-gas trapped in a chamber with a movable piston, on top of which sits a mass, M. The temperature, T, of the gas is controlled externally by a temperature reservoir placed in contact with the chamber. Therefore, M and T can be controlled directly and so will be exogenous variables in our model of this system. When the values of either M or T are altered, the height of the piston will change: If M is increased then the height will decrease; whereas if T is increased then H will increase.

In Section 2.1.1 I show that, for this system, the model $Do(Equilibrate(M))$ differs from the model expected by physical intuition. In Section 2.1.2 I show that the models derived from physical intuition correspond to the models obtained by the *Equilibrate*($Do(M)$) operator.

2.1.1 Manipulating the Equilibrium Model

Assuming that the piston in the ideal-gas system is in equilibrium, the precise expression of H in terms of T and M is a combination the ideal-gas law together with the equilibrium assumption, as given in Figure 2.1 (g, k, m_0, and t_0 are constants.).

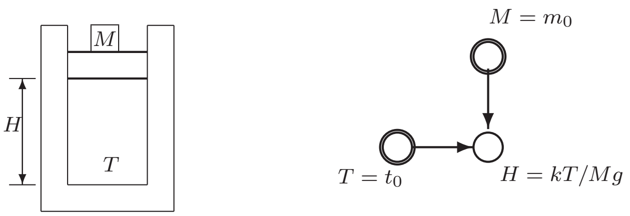


Figure 2.1. Causal model of the ideal-gas in equilibrium.

I could increase the resolution of the ideal-gas model in order to explain in more detail how the equation for H in Figure 2.1 comes about. In Figure 2.2 I have added two intermediate variables: the total force on the bottom of the piston, F_b, and the pressure of the gas, P (a is a constant). The model of Figure 2.1 can be derived from the model of Figure 2.2 by applying the *Aggregation* operator to F_b and P.

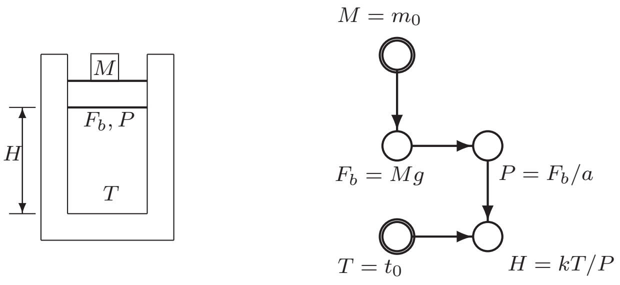


Figure 2.2. The equilibrium ideal-gas model with increased resolution.

In words the causal ordering can be described as follows: *"In equilibrium, the force* *applied to the bottom of the piston must equal the weight of the mass on top of the piston. Given the force on the bottom of the piston, the pressure of the gas must be determined, which together with the temperature determines the height of the piston through the ideal-gas law."* I now consider what happens when various variables in this model are manipulated.

2.1.1.1 Manipulating the Height of the Piston

Consider what happens when the height of the piston is set to a constant value: $H = h_0$. Physically this can be achieved by inserting pins into the walls of the chamber at the desired height, as shown in Figure 2.3a. Applying the *Do* operator to the models in Figures 2.1 and 2.2 yields the graphs depicted in Figure 2.3b and Figure 2.3d, respectively.

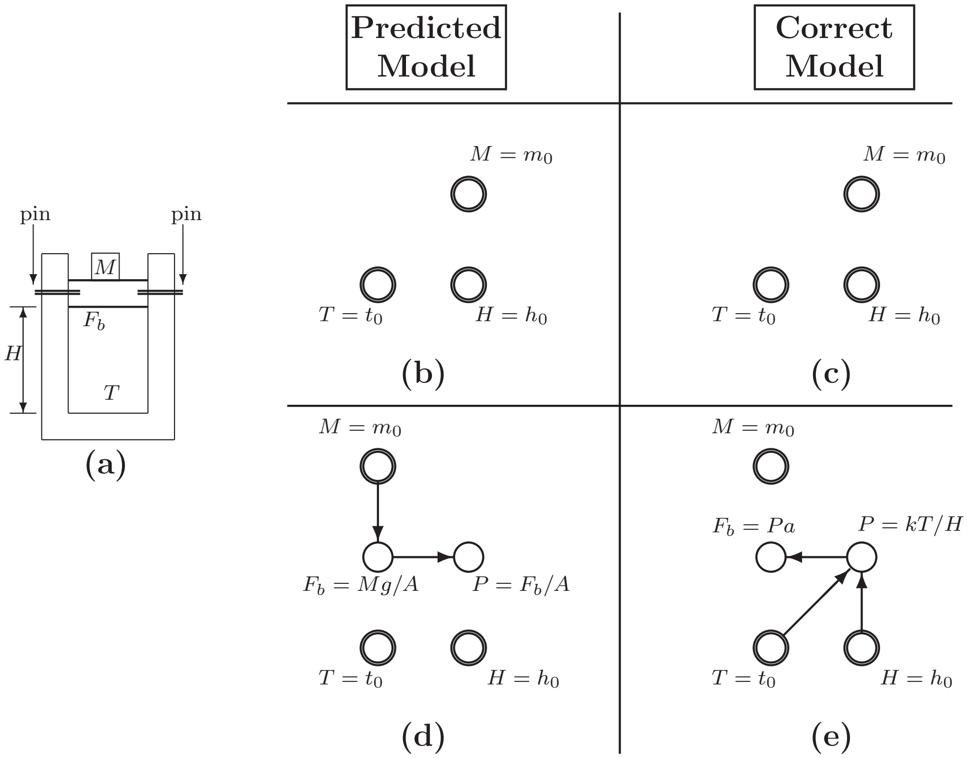


Figure 2.3. Whether or not the ideal-gas model possesses the EMC property when H is manipulated depends on the resolution of the model.

Now let us consider what the "true" causal graph for these models should look like. For the non-resolved model of Figure 2.3b, all variables in the model are being manipulated so obviously the *Do* operator will produce the correct manipulated model. The resolved model of Figure 2.3d is a simple system which we understand well, so we are able to write down the true governing equations for the manipulated version of this system, given in Figure 2.3e. Constructing the causal mapping (unique by Lemma 1) for these equations yields the graph shown. In words: *Since H and T are both fixed, P is determined by the ideal-gas law, $P = kT/H$. Since the gas is the only source of force on the bottom of the piston, F_b is determined by P: $F_b = Pa$.* Thus, P is no longer determined by F_b, and F_b is independent of M. It is clear that the true causal model shown in Figure 2.3e differs from that predicted by the *Do* operator shown in Figure 2.3d. I will show shortly that what I call the "true" model is in fact the model that results when the *Do* operator is applied to the dynamic model and then the resulting model is equilibrated.

The fact that changing the resolution of a model can cause it to violate the EMC property is a disturbing conclusion. Causal modelers are accustomed to being able to switch back and forth between different levels of abstraction for ease of model construction and explanation. Considered from the standpoint of causal discovery these results are also disheartening. Using data from the equation system of Figure 2.2 with independent error terms, the causal graph shown there would be learned by a constraint-based discovery algorithm such as the PC algorithm. On the other hand, using data from the equations governing the manipulated system would yield the causal graph in Figure 2.3e. Both of these facts can readily be verified by calculating the independencies given by the respective equation systems with independent error terms. This fact was also verified empirically using simulation in Section 2.2. The end result is clear: a causal graph learned based on the equilibrium ideal-gas system and altered with the *Do* operator will yield the incorrect causal graph of Figure 2.3d.

2.1.1.2 Manipulating the Force on the Bottom of the Piston

There are other, even more dramatic problems with manipulating variables in the expanded-resolution model. Referring back to Figure 2.2, imagine that for some reason we want to minimize the value of H. It would not be unreasonable, given the graph and the equations in Figure 2.2, to set H by applying a manipulation to F_b, since F_b is a causal ancestor of H. In particular, in order to make H as small as possible, we would want to make F_b as large as possible according to the equations in Figure 2.2.

Consider what happens when F_b is manipulated in this way: Again, the *Do* operator predicts that a manipulation will cause the arc from M to F_b to be removed from the model, but otherwise the model will be unchanged. This model is depicted in Figure 2.4b. In the real system, the force on the bottom of the piston can be set independently of the pressure of the gas by raising a movable stage up through the chamber and directly applying the desired force to the piston with the stage, as shown in Figure 2.4a. Something very unexpected happens under this manipulation. Unless by coincidence the force applied exactly balances the force due to the mass, the piston will continually be accelerated out of the cylinder, and H, which we in-

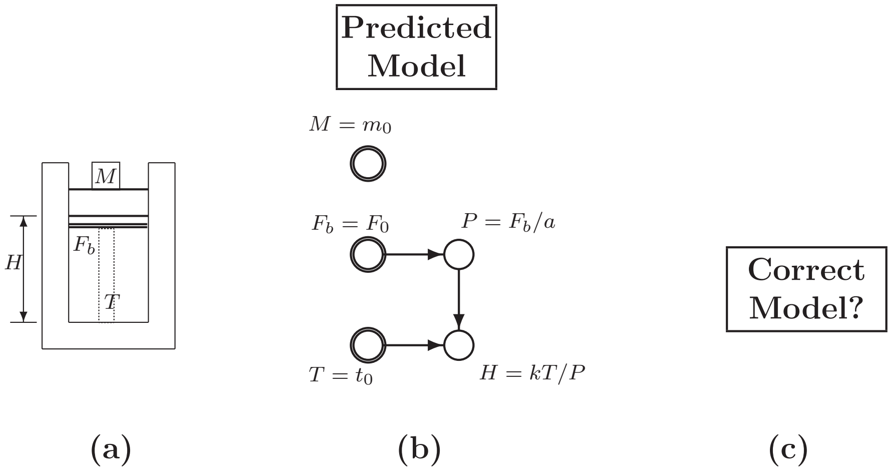


Figure 2.4. A more severe violation of the EMC property: no equilibrium model exists after manipulating F_b.

tended to minimize, instead grows without bound. Not only does this manipulation violate EMC, but even worse, we have discovered a *dynamic instability* in the system, i.e., there *is no* equilibrium model; a fact which an equilibrium causal graph alone provides no indication of.

The most disturbing fact about this example is that the instability caused by our manipulation created exactly the opposite effect we were attempting to achieve. Imagine for instance that, instead of the height of a piston, H represented the cancer level in a population of patients. If this example seems exaggerated it is only because we have some concrete understanding about the equations underlying the ideal-gas. However, imagine applying manipulations to automatically learned models of complex socio-economic or medical systems, where our basic knowledge is typically much less.

2.1.2 Equilibrating the Manipulated Dynamic Model

Manipulating the force in the ideal-gas system led to an instability. This effect gives us a clue as to what is happening; namely, underlying the equilibrium ideal-gas system is a dynamic system. When certain manipulations are made, this dynamic system may not possess an equilibrium point; the result is the hidden instability discovered in the ideal-gas system. Thus it makes sense to model this system on a shorter time-scale, using the dynamic model formalism reviewed in Section 1.3.

Imagine dropping a mass M on the piston, simultaneously altering the temperature of the gas, and shortly after measuring the values of all the remaining variables. The physics of this system is comprised of a few fundamental equations: The force on the top of the piston F_t is given by the weight of the mass M:

$$F_t = Mg. \tag{2.1}$$

The acceleration A of the piston is given by Newton's second law:

$$\Sigma_i F_i = MA. \tag{2.2}$$

The pressure of the gas P is related to the temperature T and the height of the piston H through the ideal gas law:

$$P = kT/H, \tag{2.3}$$

where k is a constant. The force on the bottom of the piston is determined by the pressure and the cross-sectional area a of the cylinder:

$$P = F_b/a \tag{2.4}$$

The height H and the velocity V are determined by recurrence relations (integrals):

$$V_{(t)} = V_{(t-1)} + A_{(t-1)}\Delta t \tag{2.5}$$

$$H_{(t)} = H_{(t-1)} + V_{(t-1)}\Delta t \tag{2.6}$$

The causal graph of this system is shown in Figure 2.5. Using this model, it is possible to show what is happening when we apply the *Do* operator to the equilibrium ideal-gas system.

2.1.2.1 Manipulating the Height of the Piston.

Let us again fix the height of the piston, using the dynamic model of Figure 2.5 to describe the ideal-gas system. To fix the piston, we must set H to some constant value for all time, $H_{(t)} = h_0$. We also must stop the piston from moving because we want to set H for all time, so we must set $V_{(t)} = 0$ and $A_{(t)} = 0$. Thus, in the dynamic graph with integration links, we can think of this one action of setting the height of the piston as three separate actions. Applying the *Do* operator to these three variables results in the causal graph shown in Figure 2.6 (b). Since H is being held constant, the graph in Figure 2.6 (b) is already an equilibrium graph with respect to H, so

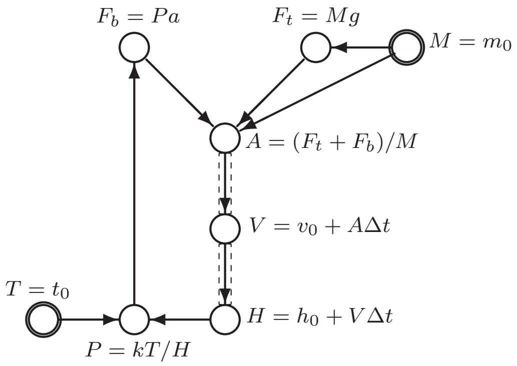


Figure 2.5. The dynamic causal model for the ideal-gas system. Three new variables have been added: A, V, and F_t, the acceleration of the piston, velocity of the piston, and the total force on the top of the piston, respectively.

applying the *Equilibration* operator results in no change to the graph. By comparing Figure 2.6 (b) to the "true" model of Figure 2.3 (c), we can see that aside from the extra variables that were added to the differential model for clarity (F_t, A and V), Figure 2.6 (b) (the *Equilibrate*$(Do(M, H), H)$ model) is the model we expected to get based on our physical knowledge when inserting pins into the walls of the cylinder to fix the height of the piston. Clearly the EMC property is not obeyed for this system, and the *Equilibrate*$(Do(M, H), H)$ model is the more useful.

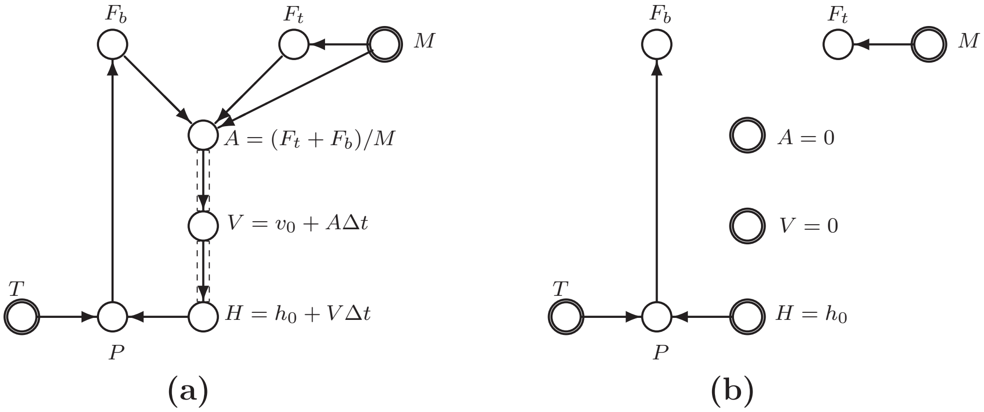


Figure 2.6. The graph corresponding to the *Equilibrate*$(Do(M, H), H)$ operation on the ideal-gas dynamic model is identical to the expected graph from Figure 2.3 (c).

2.1.2.2 Detecting Instabilities

Differential models can also be used to predict when a manipulation will cause an instability. According to the discussion in Section 1.3.3, the variable $\dot{X}$ must somehow be a function of X for stability to occur. What does this imply about dynamic causal models? In order for stability to occur, at the very least, $X_{(t)}$ must be an ancestor of $\dot{X}_{(t)}$. When there exist higher order derivatives in the model, by construction this ancestry can only occur through the highest-order derivative in the model. These observations thus suggests a structural condition for stability in a causal graph (as a reminder, a *differential graph* is a dynamic graph with all integration links removed):

Definition 31 (The Structural Stability Principle) *Let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be a causal model corresponding to a differential causal graph G_d, let $V \in \mathbf{V}$ be a dynamic variable and let $V^{(n)}$ denote the highest derivative of V in $\mathbf{V}$, the Structural Stability Principle states that V will possess a stable fixed-point only if $V \in \boldsymbol{Anc}(V^{(n)})$ in G_d.*

In fact, the structural stability principle immediately points out a flaw in the dynamic model of Figure 2.5: since V is a dynamic variable in this model, in order for it to stabilize, it needs to also be an ancestor of A. This observation is a restatement of the fact that a dissipative (frictional) force must be present in a second-order system in order for oscillations to dampen out. Thus, a better model of the ideal gas system is given in Figure 2.8.

Consider the effects of manipulating F_b in the dynamic model of the ideal-gas system. Applying the *Do* operator to F_b in Figure 2.7 (a) yields the model shown in Figure 2.7 (b). We can see immediately from the causal graph that this manipulation will break the only feedback loop for X in this system, and thus according to the Structural Stability criterion, there does not exist a stable equilibrium point for H. Therefore applying the *Do* operator to the non-equilibrated graph, and the addition

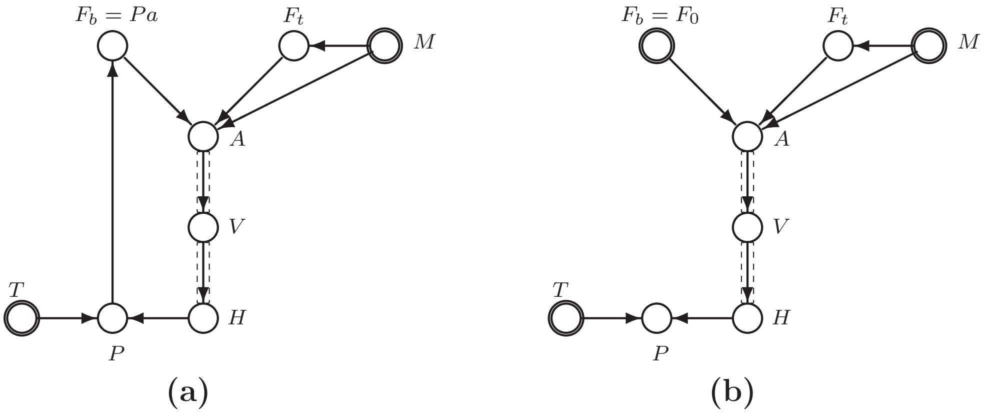


Figure 2.7. Applying the *Do* operator to the differential model when manipulating F_b allows the instability to be detected because we have broken the only feedback loop in the graph.

of the Structural Stability criterion shows that the $Equilibrate(Do(M, F_b), H)$ system will be unstable. Again, this model corresponds to path B in Figure 1, not path A.

2.2 Discovery from Data: Empirical Results

Section 2.1 demonstrated that the answer to the Equilibration-Causation Question 1 was “no”. This section addresses the Equilibrium-Causation Question 2 using empirical studies. I performed numerical simulations of some dynamic systems to demonstrate that as the time scale was increased enough so that an equilibration could occur, the causal structure that was learned from data corresponds to the structure obtained by applying the *Equilibration* operator to the dynamic model. This fact is significant because it indicates that whenever a causal structure that is learned from equilibrium data is used for causal reasoning, then Path A of Figure 1 is being taken: if the EMC property does not hold for the model being used then causal reasoning will produce incorrect results. These experiments provide an empirical answer to Question 2 because it has been proven [Spirtes *et al.*, 1993] that, in the absence of latent variables, assuming a faithful model to a distribution exists, then the PC algorithm will recover the graph that is faithful to the distribution that generated the data. Furthermore Spirtes *et al.* [1993] also prove that the probability of generating a non-faithful model by chance is zero.

In short, these experiments show that the empirically-determined faithful graph corresponds to the one given by the *Equilibration* operator. If the true causal graph (obtained by taking path B in Figure 1) is different from this faithful graph, then causal faithfulness is violated by definition.

2.2.1 Ideal-Gas System

The first experiment performed was to simulate the ideal-gas system. Preliminary simulations using the equation system presented in Figure 2.5 showed only non-stationary oscillatory solutions, i.e., the piston would oscillate about its fixed point but never converge. As already discussed, this oscillatory behavior was due to the absence of a dissipative force in the system of equations. Thus, to ensure a stable fixed-point, it was necessary to add a frictional force $F_f = -\gamma V$, where γ is the coefficient of friction.

The simulation assumed linear independent Gaussian noise terms in order to make possible the induction of structure from data using statistical significance tests. The complete system of equations used for the simulation is presented in Figure 2.8. The values of the constants were determined by trial-and-error to ensure that the velocity of the piston remained much less than H (to avoid numerically-induced instabilities) and that the height of the piston would never approach zero (which would cause a singularity in the ideal-gas law: $P = T/H$). The values that were used were: $h_0 = 6$, $v_0 = 1$, $m_0 = 6$, and $t_0 = 50$. Each γ_i term was assumed to be a Gaussian random variable with mean zero. It was observed that the ability to correctly recover the expected causal structures depended strongly on the relative noise levels of the variables. To illustrate this fact, I introduce an additional parameter ρ which links

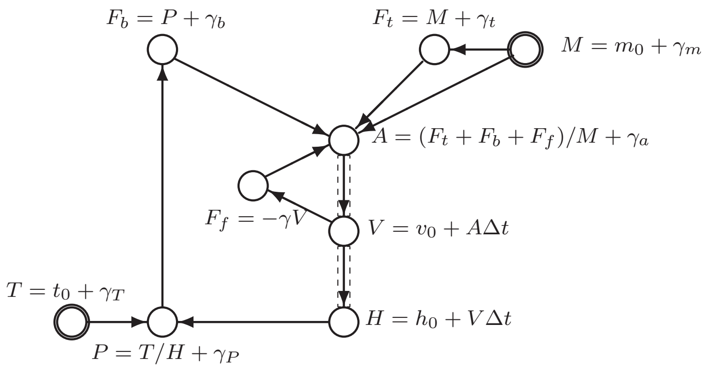


Figure 2.8. The system of equations used to simulate the learning of the ideal-gas. Independent Gaussian error terms (denoted with the symbol γ_i) have been added as well as a frictional force to ensure stability.

the standard deviations (denoted as σ_i) of the noise-terms. The following values were used: $\sigma_H = 0.75$, $\sigma_m = 0.5$, $\sigma_T = 5$, $\sigma_t = 0.5\rho$, $\sigma_a = 0.6\rho$, $\sigma_p = 0.9\rho$, $\sigma_b = 0.9\rho$. Since ρ has a constant value for all records in any given database, it will not violate causal sufficiency for this system. The frictional force was treated as a latent variable (no attempt was made to include it into the learning), and was treated as deterministic for simplicity—its only purpose was to damp out oscillations. The coefficient of friction γ was set to 0.25 to allow lightly damped oscillatory motion of the piston. A few typical equilibrations of the piston are illustrated in Figure 2.9.

Distinct runs were generated by repeatedly sampling the noise terms of each variable (i.e., “shocking” the system) and allowing the equation system to guide the evolution of the variables. In order for the system to converge, it was noted that an assumption of stationary noise terms was required. That is, all error terms are sampled once at time step $t = 0$, and thereafter the system was allowed to evolve deterministically until equilibrium, as opposed to sampling the noise terms anew at each time step. This was necessary because randomly shocking the system close to equilibrium will continuously bring it out of equilibrium again.

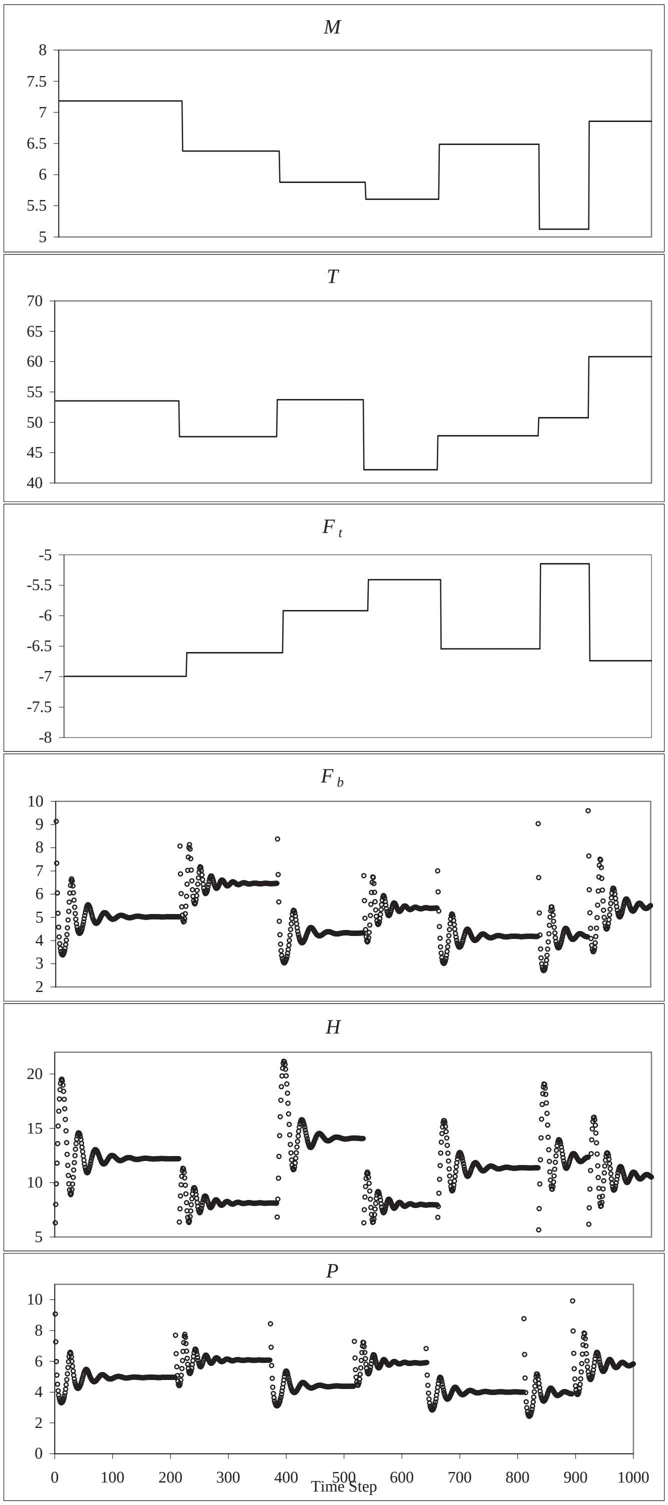


Figure 2.9. A few typical equilibrations of the ideal-gas system. M, T and F_t were exogenous for each run; the other variables evolved deterministically to equilibrium.

Each run was allowed to go up to 1000 time steps or until the system was determined to be in an equilibrium state, whichever came first. The system was deemed to be in the equilibrium state if the absolute difference in the change of H from one time step to the next was less than 0.0001. Given the mean value of H: $\langle H \rangle = \langle T \rangle / \langle M \rangle \simeq 10$, this amounts to a change of about 1/1000 of 1 percent. Thus, we can be confident that if the system was stopped prematurely, the values will be nearly identical to the those at time step $t = 1000$.

Using this procedure, two databases D_{dyn} and D_{equ} were generated. Each complete run to equilibrium corresponded to a single record in the databases: a snapshot of the system state at time step $t = 0$ produced a single record for D_{dyn}, and a snapshot at $t = 1000$ defined a record of D_{equ}. This was repeated until two databases of some size N were generated. These two databases were used with the PC algorithm (page 101) to learn the causal structures observed on short (D_{dyn}) and long (D_{equ}) time-scales. A modified version of PC was used which forbade cycles or bi-directional arrows and randomized the order in which independencies were checked [Dash and Druzdzel, 1999]. Data for each variable took on a continuous range of values, and in all cases the Fisher's-z statistic was used to test for conditional independence using a significance level of $\alpha = 0.05$.

I restricted structure learning to the variables $\{M, T, H, P, F_t, F_b\}$, namely the variables relevant to the static analysis of this system. Over this subset of variables we expect to recover the two structures S_1 and S_2 shown in Figure 2.10: S_1 when $t = 0$ and S_2 when $t = 1000$. N was systematically varied from the set {100, 500, 1000, 2000, 4000, 10000}, and ρ was varied from the set {0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4}. 100 measurements were taken for each (N, ρ) combination, and the following three statistics were averaged over the measurements: $N_{adj} = N^+_{adj} + N^-_{adj}$ is the number of extra adjacencies (N^+_{adj}) plus the number of missing adjacencies (N^-_{adj}); $N_v = N^+_v + N^-_v$ is the number of extra v-structures (N^+_v) plus the number of missing

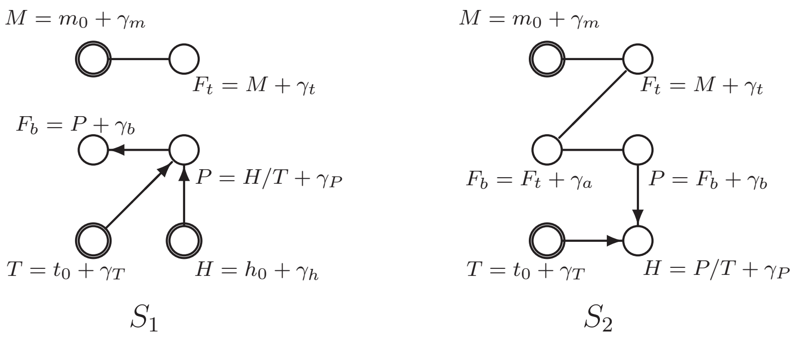


Figure 2.10. The two patterns expected to be recovered from the simulation of the ideal-gas system of Figure 2.8. S_1 is the expected pattern for $t = 0$ (D_{dyn}), and S_2 is the expected pattern for $t = 1000$ (D_{equ}).

v-structures (N_v^-); and $P_{correct}$ is the fraction of times that precisely the correct structure was learned. We expect that as N increases, the probability of recovering S_1 and S_2 should increase, ideally approaching unity. Figure 2.11 shows the measured probability of learning the correct graph, averaged over all values of N, as ρ is varied.

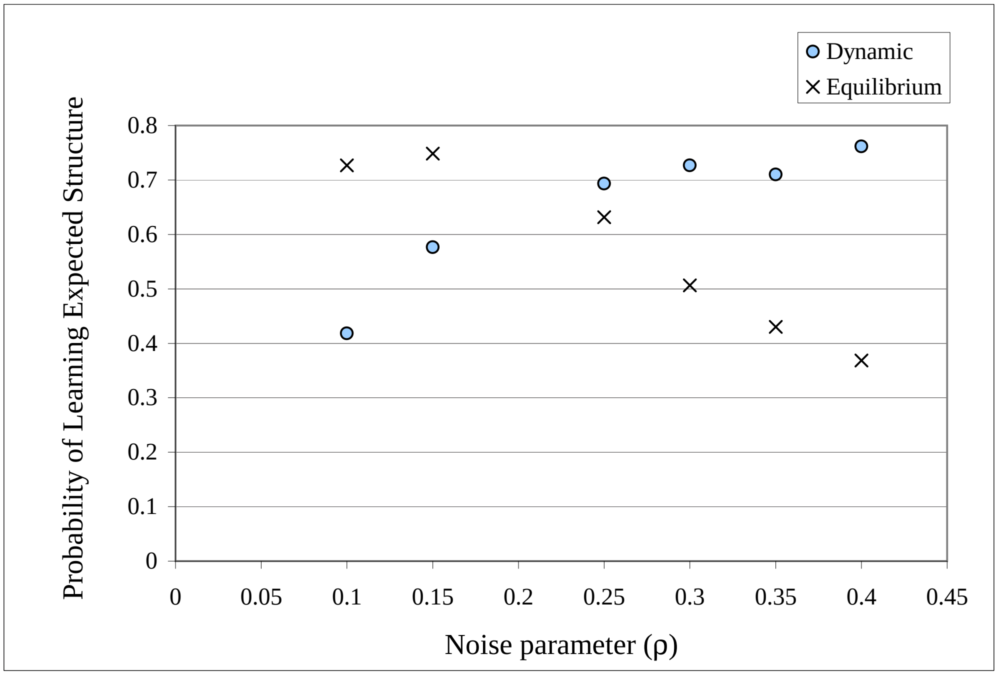


Figure 2.11. The probability of learning the expected dynamic (S_1) and equilibrium (S_2) graphs as the noise parameter ρ increases for the ideal-gas system.

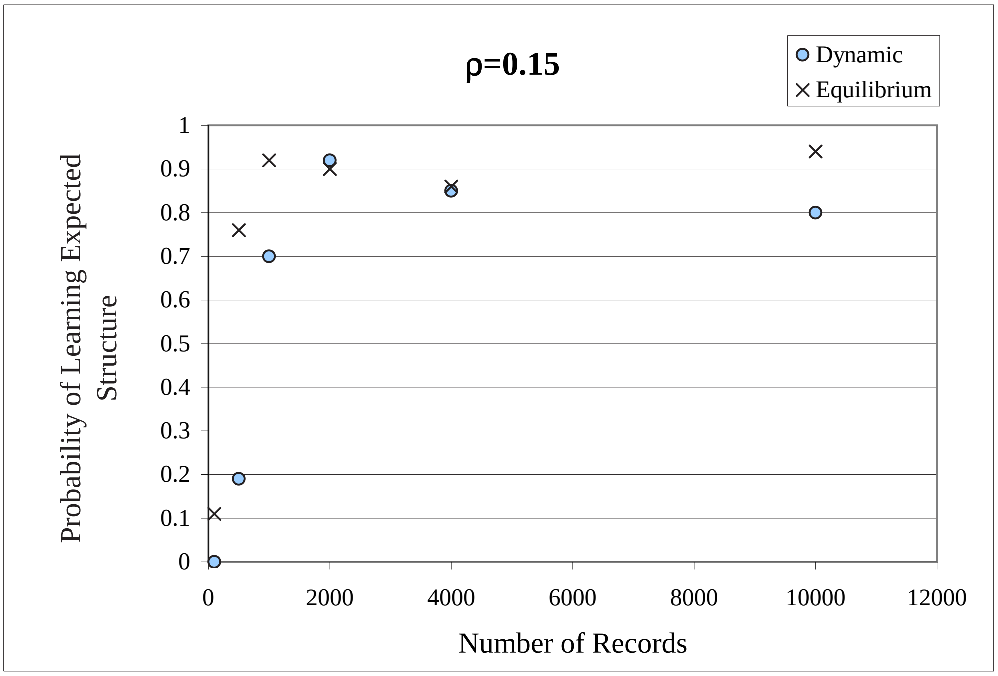


Figure 2.12. The probability of learning the expected dynamic (S_1) and equilibrium (S_2) graphs as the number of records increases for the ideal-gas system, for an optimum value of ρ.

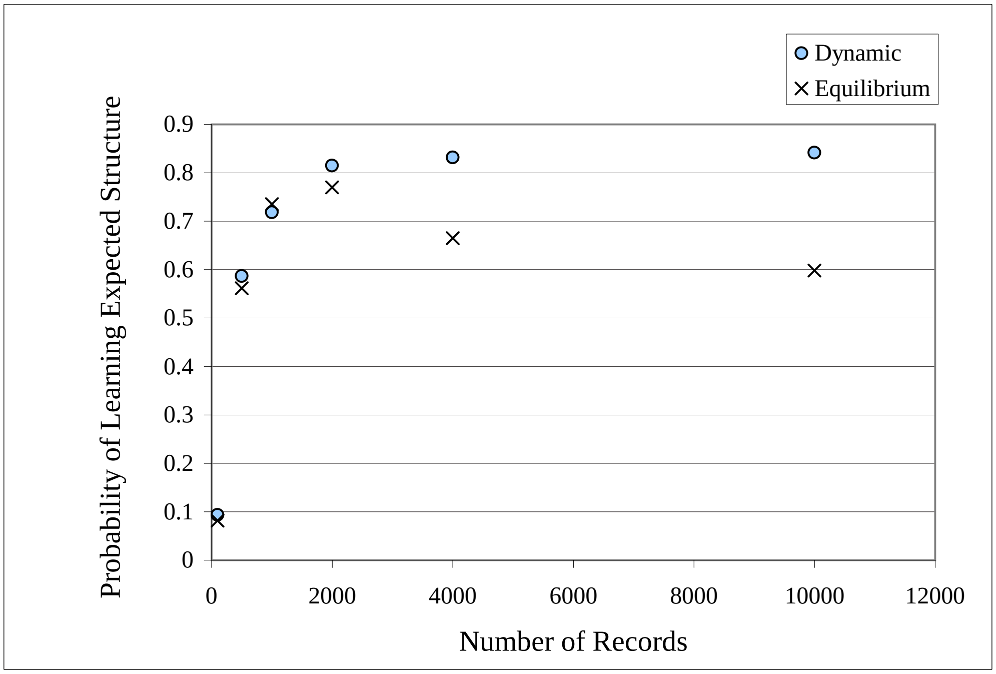


Figure 2.13. The probability of learning the expected dynamic (S_1) and equilibrium (S_2) graphs as the number of records increases for the ideal-gas system.

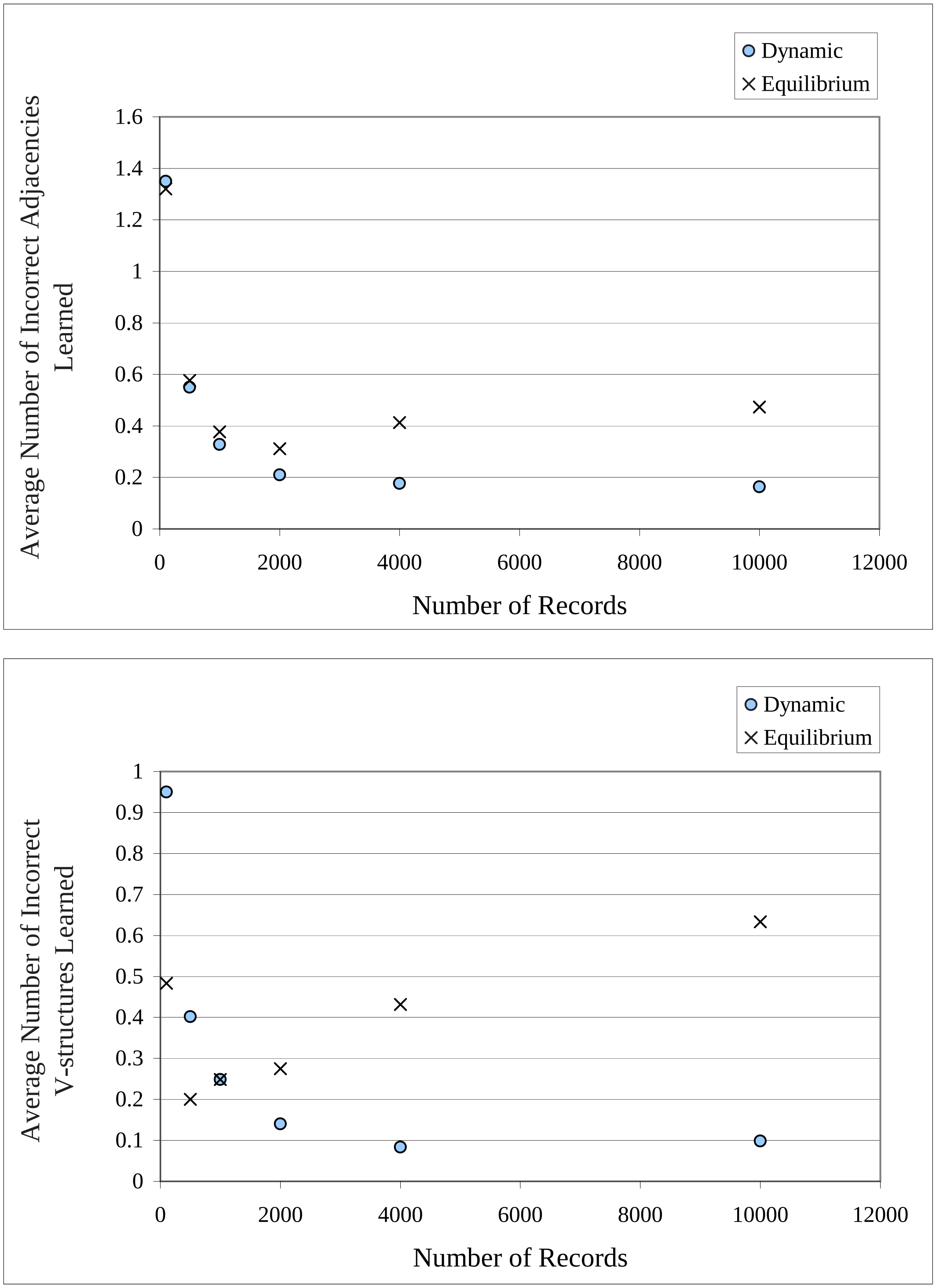


Figure 2.14. Incorrect structural features as a function of the number of records for the dynamic and equilibrium ideal-gas systems, averaged over all values of ρ.

It is apparent that there exists a complementarity between the ability to recover the dynamic structure (S_1) versus the ability to recover the equilibrium structure (S_2). Nonetheless, as Figure 2.12 shows, it was possible to find specific values of ρ for which this convergence was apparent for both structures. Figure 2.13 shows the probability of recovering the correct graphs averaged over all values of ρ. (Incidentally, the randomizing of the order in which independence relations were checked turned out to be a crucial factor in obtaining good results for these experiments. Apparently using a fixed ordering, PC would often get stuck at a local optimum significantly degrading the quality of the results.) Here, it is evident that, while the probability of recovering S_1 appears to be monotonically converging, the probability of recovering S_2 displays a local maximum at around $N = 2000$. Despite this fact, Figure 2.14 shows that on average fewer than one adjacency and one v-structure were learned incorrectly (either added or deleted).

2.2.2 Pseudo-Linear Ideal-Gas System

Although the measurements made with the ideal-gas system do illustrate the change in structure as the time-scale is changed, the lack of clear convergence with increasing N is not the optimal anticipated result. I hypothesized that the recovery rate of S_2 peaked at an intermediate number of records because of the use of the linear partial correlation for the testing of independence. The ideal gas law $H = P/T$ involves a non-linear relationship between T and H, and the presence of non-linear associations, together with the assumption of linearity and a large database of records, could allow the significance test to return low p-values if the relation is severely under-fit by a straight line.

To test this hypothesis, I performed a simulation on the linearized version of the ideal gas system, shown in Figure 2.15. This system is identical to the original ideal gas system, except the ideal gas law is replaced by the linear relationship $P =$

$-k(H - T - \hat{h})$. Physically, this change corresponds to replacing the ideal gas with a spring whose base can be adjusted with a constant offset T, and where the compression of the spring is given by $\hat{h} - H$ ($\hat{h}$ is the relaxed height of the piston when $M = 0$ and $T = 0$). One might argue that the equation for A in Figure 2.15 is non-linear because of the inverse dependence on M; however, this relation does not come into play when learning S_1 (because A is not included in the causal model), and the M drops out of the equation in equilibrium, leaving only a linear relationship between the forces in S_2. For this reason I refer to this system as the *pseudo-linear ideal gas system*. A typical time trace for this system is shown in Figure 2.16.

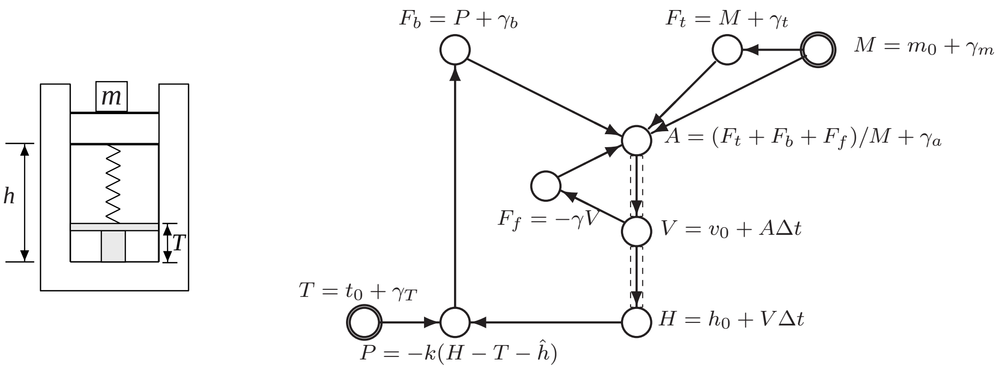


Figure 2.15. Using a linear ideal-gas law is equivalent to replacing the gas with a spring where T denotes a shift of the base of the spring and P denotes the force of the compressed spring.

Figure 2.17 shows the probability of recovering the correct structure as a function of N, averaged over values of ρ. When the linear equation system is used, the learned graphs converge neatly to S_1 and S_2. While the same complementarity between S_1 and S_2 (as in Figure 2.11) is evident as a function of ρ when averaging over N, for all values of ρ tested, the learned graphs converge to S_1 and S_2 as N was increased.

The important observation about these simulations is this: If we alter the system of Figure 2.15 by setting $A = V = 0$ for all time and setting $H = h_0$, we can simulate the ideal-gas system under the assumption that H is being manipulated to the value

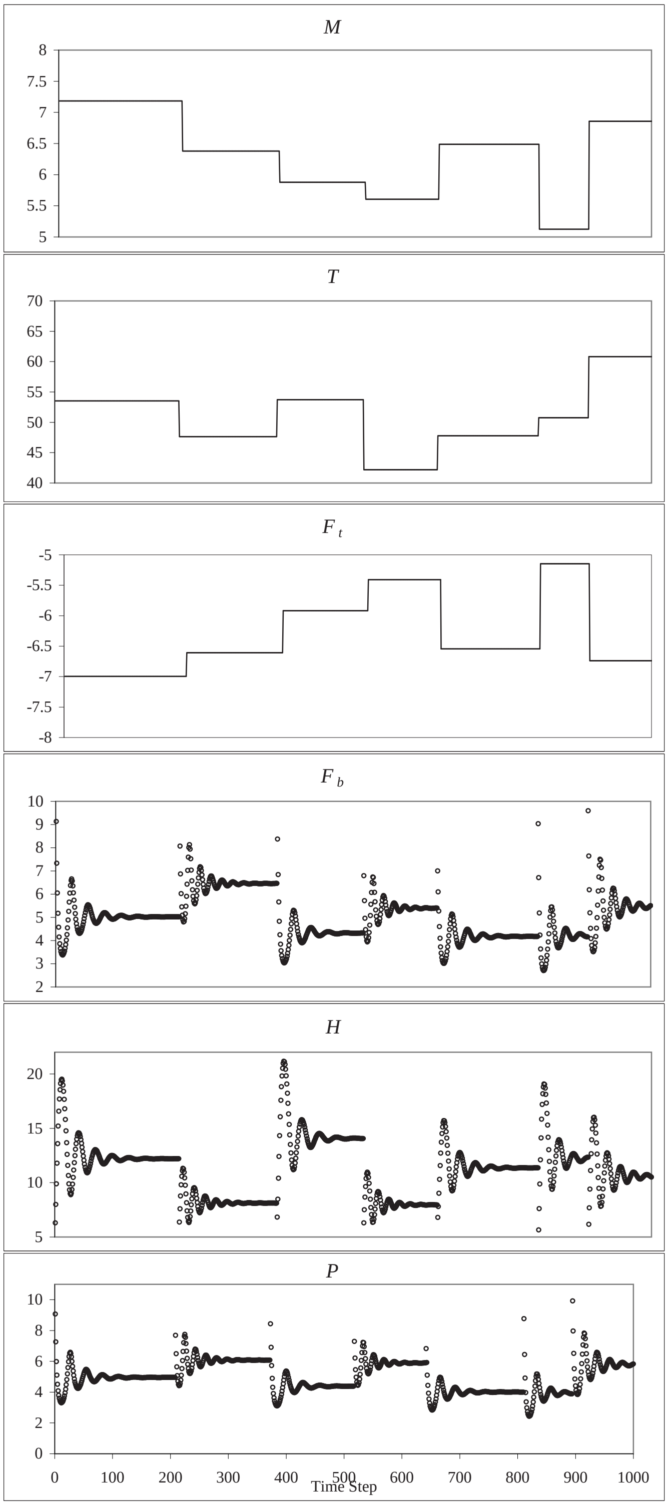


Figure 2.16. A few typical equilibrations of the pseudo-linear ideal-gas system.

h_0. However, this manipulation will produce data from a distribution identical to that of the model S_1, and therefore, we would learn S_1 from the data generated by manipulating H. This of course, is not the same graph that we would get by applying the Do operator to S_2, verifying exactly the observations of Section 2.1.

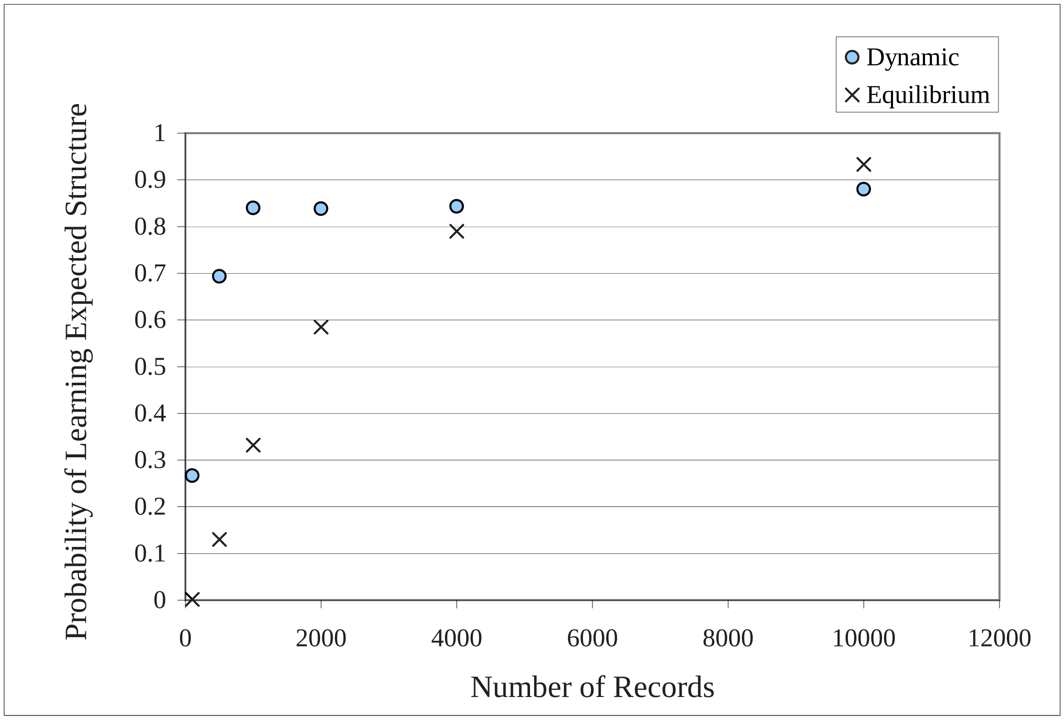


Figure 2.17. The probability of learning the expected dynamic (S_1) and equilibrium (S_2) graphs as the number of records increases for the pseudo-linear ideal-gas system, averaged over all values of ρ.

2.2.3 Filling Bathtub System

The failure of the ideal-gas system to obey the EMC property adds credence to the claim of Wold [1954, 1955] that "equilibrium" equations can not constitute causal mechanisms; for example, perhaps it was the equilibrium relation $F_b = F_t$ that was responsible for the ideal-gas violations. Here I discuss an example that shows that this conclusion is simplistic. The ideal-gas system possessed a single dynamic variable, and therefore only two patterns were expected to be observed as the observation time-scale was varied. If a system has several dynamic variables with widely-varying time-scales the situation is more complicated. One example is the filling-bathtub system introduced by Iwasaki and Simon [1994], and reproduced here.

In this system, water is entering the bathtub from the faucet at a rate Q_{in} liters per second and is exiting the drain at a rate Q_{out} liters per second. The pressure of the water at the base of the drain is P, the depth of the water is D, and the diameter of the drain is denoted as K. I take Q_{in} and K as being exogenous. If the outflow is not identical to the inflow, then the depth of the water will change in proportion to the difference:

$$\dot{D} = \alpha_0(Q_{in} - Q_{out}), \tag{2.7}$$

where α_0 is the inverse cross-sectional area of the tub. The pressure of the water at the base of the drain is proportional to the depth of the water:

$$P = \rho g D, \tag{2.8}$$

where ρ is the density of the water and g is the gravitational constant. Finally, the outflow of the water depends on the pressure at the base and the diameter of the drain:

$$Q_{out} = \alpha_1 K\sqrt{P}. \tag{2.9}$$

The causal ordering of this system is shown in Figure 2.18.

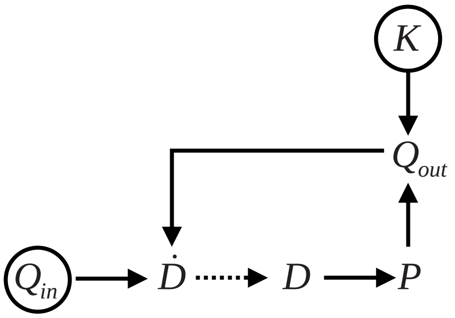


Figure 2.18. The intuitive or "mixed" causal ordering of the bathtub system.

If the depth D were shocked by very quickly adding a unit depth ΔD of water, it would take some time for the pressure at the base of the drain to change to its equilibrium value of $P = \rho g(D + \Delta D)$. Similarly, if we perturb P, it would take some short but nonzero time for the outflow rate Q_{out} to respond. If we assume for simplicity that the change in P and Q_{out} under these perturbations is proportional to the difference between the current values and the equilibrium values, we could derive the following differential equations:

$$\dot{P} = \alpha_2(\rho g D - P), \text{ and} \tag{2.10}$$

$$\dot{Q}_{out} = \alpha_3(\alpha_1 K\sqrt{P} - Q_{out}). \tag{2.11}$$

Replacing the equilibrium relations 2.8 and 2.9 with the dynamic relations 2.10 and 2.11, respectively, yields the causal ordering referred to as "dynamic" by Iwasaki and Simon, shown in Figure 2.19.

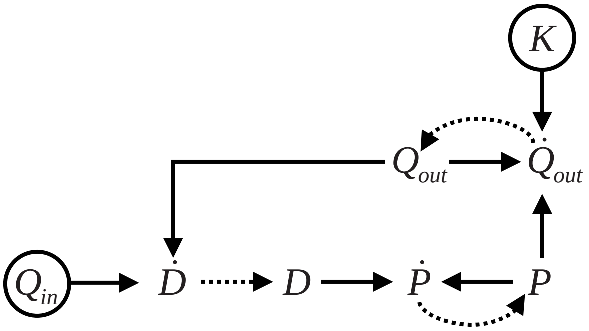


Figure 2.19. The dynamic causal ordering of the bathtub system.

Because the system of Figure 2.19 involves three dynamic variables, there exist three important time-scales for this system, controlled by the inverse of the coefficients: $\tau_D \propto 1/\alpha_0$, $\tau_P \propto 1/\alpha_2$, and $\tau_Q \propto 1/\alpha_3$, for D, P and Q_{out} respectively. Assuming the system is stable over all time-scales, if $\tau_P \ll \tau_Q \ll \tau_D$, then there will exist four possible equilibrium causal structures learned from data depending on the time, τ, at which the data was observed. These four structures (over variables $\mathbf{V} = \{Q_{in}, Q_{out}, P, T, K\}$) are shown in Figure 2.20. At $\tau = 0$ each of the five

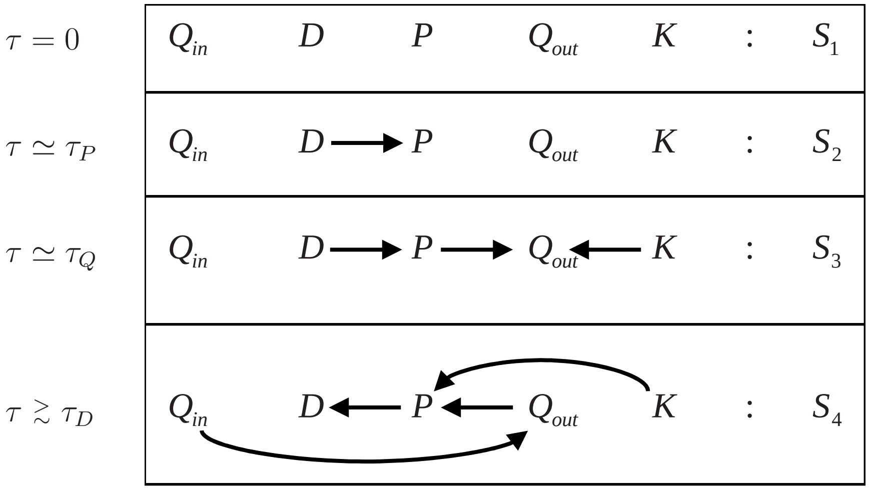

$\tau = 0$
Q_{in} D P Q_{out} K : S_1
$\tau \simeq \tau_P$
Q_{in} D P Q_{out} K : S_2
$\tau \simeq \tau_Q$
Q_{in} D P Q_{out} K : S_3
$\tau \gtrsim \tau_D$
Q_{in} D P Q_{out} K : S_4

Figure 2.20. The bathtub system has four correct equilibrium structures depending on the time scale at which the system is modelled.

variables in $\mathbf{V}$ are given by their initial conditions and so are exogenous; in this case S_1 will be the structure learned from data. After enough time has passed for P to equilibrate ($\tau \simeq \tau_P$), then Equation 2.10 reduces to Equation 2.8, and the structure S_2 will result. After $\tau \simeq \tau_Q$, enough time has passed for Q_{out} to equilibrate, and Equation 2.11 becomes Equation 2.9, resulting in the structure S_3. Finally, after $\tau > \tau_D$, enough time has passed for D to equilibrate and Equation 2.7 reduces to the Equation 2.12:

$$Q_{in} = Q_{out}, \tag{2.12}$$

leading to a drastic restructuring of equations and resulting in model S_4.

I simulated learning over several time-scales for the filling-bathtub system. The following values for constants were used: $\rho = g = \alpha_1 = 1$, $\alpha_0 = 0.005$, $\alpha_2 = 0.05$, and $\alpha_3 = 0.01$. All variables were initialized from the uniform distribution over the interval $(0, 1)$. Independent Gaussian error terms with mean 0 were added to each derivative variable. The error terms for $\dot{D}$ and $\dot{Q}_{out}$ had standard deviation equal to 0.01, and $\dot{P}$ had standard deviation equal to 0.5. The first fifty steps of a typical time-trace is shown in Figure 2.21.

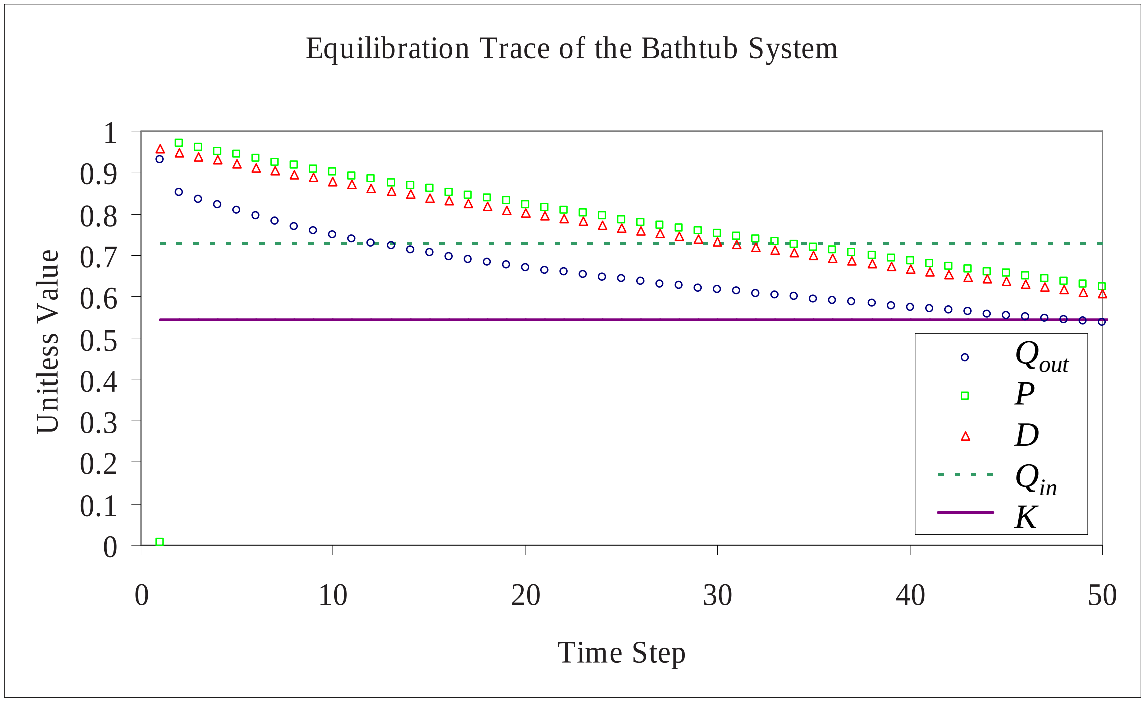


Figure 2.21. The first fifty steps of a typical bathtub simulation run.

A database of $N = 10000$ records was generated for each of the 29 time-scales given in the set $\mathcal{T} = \{0-10, 20, 30, 40, 50, 80, 100, 125, 150, 200, 250, 300, 500, 750, 1000, 1250, 1500, 1750, 2000\}$, and for each of these databases the PC algorithm was run as in the previous experiments presented in this chapter. This was performed 50 times for each time scale, and the number of times the pattern corresponding to the graphs in Figure 2.20 were exactly recovered was counted. The normalized results, showing the probability of retrieving the four structures as a function of the time scale, are shown in Figure 2.22. The time-scales $20 \leq \tau \leq 750$ are excluded from this figure—they produced empirical probabilities of 0 for all four structures. These

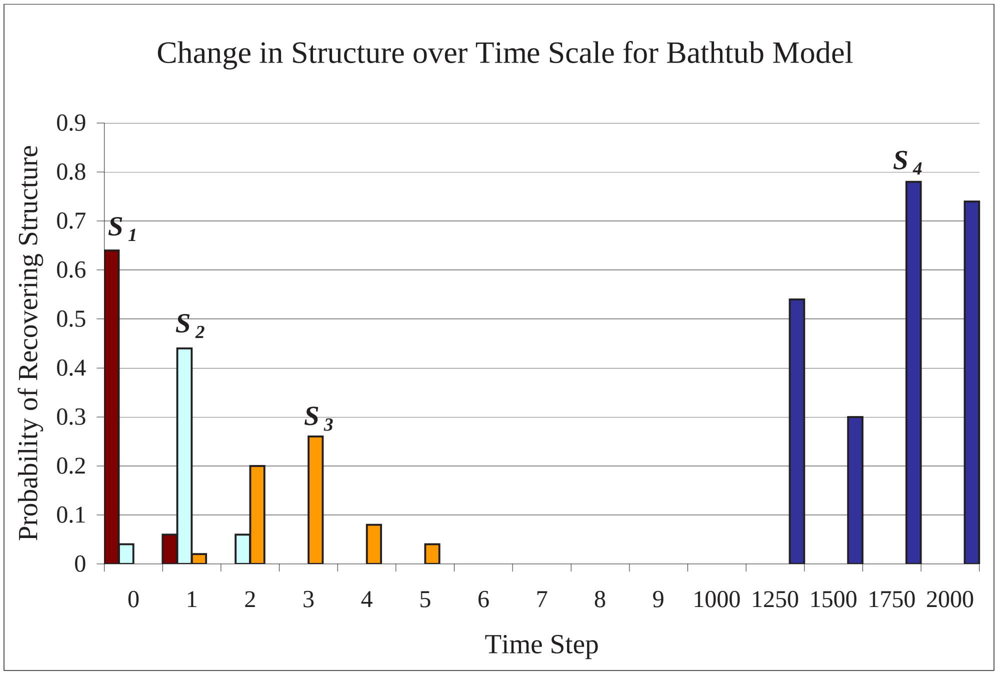


Figure 2.22. As the time step is varied, each of the four equilibrium structures can be recovered in sequence.

results show that as the time-scale increases each of the four causal structures will be learned, in the order predicted by the analysis above.

The overarching point to this section is that the causal structure learned from data depends very much on the time-scale of the data, and there might be several important time scales. Obviously, if no recourse is made to dynamics, it would be in general impossible to use a learned causal graph to predict the effects of manipulation. These results demonstrate that the equilibrated model is the one learned from data (and therefore the faithful one). If the EMC property is not obeyed, then the model that is faithful will not support causal inferences with the *Do* operator; therefore such a model violates the causal faithfulness condition.

In regard to Wold's suspicion that equilibration relations cannot be used for causal reasoning, this example shows that this suspicion is overly simplistic. In models S_2 and S_3, variables P and Q were equilibrated, respectively; yet these graphs obeyed

EMC. It was not until the equilibration of D that the equilibrium causal structure became inconsistent with the differential graph. Why the equilibration of D did not commute with the Do operator while that of P and Q_{out} did is addressed by Theorem 10 in Section 2.3.

2.3 Theoretical Results

In this chapter I treat the Equilibrium-Causation Question in a more formal way. This formalization allows me to more precisely characterize when a model will or will not obey the EMC property.

I now state several definitions and theorems that lead up to my main proofs. For the remainder of this section, let $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ be an arbitrary dynamic causal model, let $X \in \mathbf{V}$ be a dynamic variable in M and let $M_{\tilde{x}} = \langle\langle \mathbf{V}_{\tilde{\mathbf{x}}}, \mathbf{E}_{\tilde{\mathbf{x}}}\rangle, \phi_{\tilde{x}}\rangle$ be a causal model obtained by performing a well-defined equilibration operation on X. Let G and $G_{\tilde{x}}$ be the causal graphs for M and $M_{\tilde{x}}$, respectively and $G_x^{(0)}$ be the differential graph corresponding to G. I define $\boldsymbol{Fb}(X)$ to be the set of feedback variables: $\boldsymbol{Fb}(X) = \{\boldsymbol{Anc}(X)_G \cap \boldsymbol{Des}(X)_G\}$. Let $\boldsymbol{V_{del}}(X)$ and $\boldsymbol{E_{del}}(X)$ be defined as in Definition 25 (Page 30).

2.3.1 Guaranteed Violation of EMC

Definition 32 (RFRE Model, $\mathcal{F}$) *$M_{\tilde{x}}$ is a feedback-resolved equilibrated model with respect to M and X if and only if:*

1. **Equilibration:** *$M_{\mathcal{F}}$ is derived from a dynamic model M_d by equilibrating X in M_d,*

2. **Recursivity:** *$M_{\mathcal{F}}$ and M_d are both recursive, and*

3. **Feedback-resolution:**

   $\{\boldsymbol{Fb}(X) \setminus \mathbf{V}_{\mathbf{del}}(X)\} \cap \boldsymbol{Ch}(X)_{G_d} \neq \emptyset$.

I denote the class of all RFRE models as $\mathcal{F}$, and use $\mathcal{F}(M, X)$ to denote the set of RFRE models with respect to M and X.

Lemma 2 *If M is recursive, then there exists an ordering relation O on the associations of ϕ such that:*

1. *$O(\langle V_i, E_i\rangle) < O(\langle V_j, E_j\rangle)$ if $V_i \in \boldsymbol{Anc}(V_j)_{G_x^{(0)}}$, and*
2. *the pairs corresponding to $\boldsymbol{Fb}(X)$ form a contiguous sequence in O.*

Proof: In $G_x^{(0)}$, all $X^{(i)}$ such that $i \neq n$ are exogenous by construction (they are specified by the initial conditions in the model). Thus they can be ordered before all other $V \in \boldsymbol{Fb}(X)$. Define $\boldsymbol{Anc}(\boldsymbol{Fb}(X))_{G_x^{(0)}} \equiv \bigcup_{V \in \boldsymbol{Fb}(X)} \boldsymbol{Anc}(V)_{G_x^{(0)}}$ and $\boldsymbol{Des}(\boldsymbol{Fb}(X))_{G_x^{(0)}} \equiv \bigcup_{V \in \boldsymbol{Fb}(X)} \boldsymbol{Des}(V)_{G_x^{(0)}}$ to be the set of ancestors and descendants, respectively of $\boldsymbol{Fb}(X)$. By transitivity of the ancestor and descendant relationships, if there exists a $V \in \boldsymbol{Anc}(\boldsymbol{Fb}(X)) \cap \boldsymbol{Des}(\boldsymbol{Fb}(X))$ then $V \in \boldsymbol{Fb}(X)$. Thus an ordering can be defined such that $O(V_{anc}) < O(V_{fb}) < O(V_{des})$ for arbitrary variables $V_{anc} \in \boldsymbol{Anc}(\boldsymbol{Fb}(X)) \setminus \boldsymbol{Fb}(X)$, $V_{des} \in \boldsymbol{Des}(\boldsymbol{Fb}(X)) \setminus \boldsymbol{Fb}(X)$, and $V_{fb} \in \boldsymbol{Fb}(X)$. □

Throughout the remainder of the thesis, I use the notation that if there exists an ordering over associations O, then for an arbitrary association $\langle V, E\rangle$ I define $O(V) = O(E) = O(\langle V, E\rangle)$.

The following lemma shows that, during an equilibration, all non-feedback variables retain their original mappings:

Lemma 3 *Let $\overline{F}$ denote the set $V_{\tilde{x}} \setminus \{\boldsymbol{Fb}(X) \cup \{X\}\}$. If M and $M_{\tilde{x}}$ are recursive then $\phi_{\tilde{x}}(V) = \phi(V)$ for all $V \in \overline{F}$.*

Proof: Define an ordering O for ϕ and label the pairs $\langle X_i, E_i\rangle$ in ϕ as in the proof of Lemma 1. By Lemma 2, O can be defined such that all associations for variables in $\boldsymbol{Fb}(X)$ form a contiguous sequence in O with $O(X) < O(V)$ for all $V \in \boldsymbol{Fb}(X)$. Define $o_x = O(X)$ and $o_{fb} = O(V)$ where V is highest ordered variable in $\boldsymbol{Fb}(X)$.

Partition the variables and equations into three sets: $\mathbf{V} = \{\mathbf{V_{pre}}, \mathbf{V_{fb}}, \mathbf{V_{post}}\}$ and $\mathbf{E} = \{\mathbf{E_{pre}}, \mathbf{E_{fb}}, \mathbf{E_{post}}\}$, where

$$\mathbf{V_{fb}} \equiv \{V \mid V \in \boldsymbol{Fb(}X\boldsymbol{)} \cup \{X\}\},$$
$$\mathbf{V_{pre}} \equiv \{V \mid V \in \mathbf{V} \text{ and } O(V) < o_x\},$$
$$\mathbf{V_{post}} \equiv \{V \mid V \in \mathbf{V} \text{ and } O(V) > o_{fb}\},$$

and $\mathbf{E_{fb}}$, $\mathbf{E_{pre}}$, $\mathbf{E_{post}}$ are the sets of equations associated with $\mathbf{V_{fb}}$, $\mathbf{V_{pre}}$, and $\mathbf{V_{post}}$, respectively, in ϕ.

First I show that all $E \in \mathbf{E_{pre}}$ get assigned to some $V \in \mathbf{V_{pre}}$ and all $E \in \mathbf{E_{post}}$ get assigned to some $V \in \mathbf{V_{post}}$, then the result follows by the fact that within the post and pre sets ϕ already provides a causal mapping because none of the equations or variables in $\mathbf{E_{pre}}$ have changed and no dependency on the variables in $\mathbf{V_{post}}$ have changed in the equations in $\mathbf{E_{post}}$ by the assumption of a well-defined equilibration. Since $\phi_{\tilde{x}}$ is unique it must possess the same mapping between these sets.

By an argument identical to that of Lemma 1 it can be proven that the first pair $\langle V_j, E_i\rangle \in \phi_{\tilde{x}}$ such that $j \neq i$ will occur when $j = o_x$; thus all $V \in \mathbf{V_{pre}}$ get mapped only to equations in $\mathbf{E_{pre}}$. Finally, no $E_{fb} \in \mathbf{E_{fb}}$ can be assigned to some $V \in \mathbf{V_{post}}$ because by construction no equation of order i can be assigned to a variable of order $j > i$ since $O(V_E) \leq O(E)$ for all $V_E \in \boldsymbol{Params(}E\boldsymbol{)}$. Therefore, no $E_{post} \in \mathbf{E_{post}}$ can be assigned to a $V \in \boldsymbol{Fb(}X\boldsymbol{)}$ because there would be no equation to replace it in $\mathbf{E_{fb}}$. □

The next lemma says, informally, that all ancestors of X in $\boldsymbol{Fb(}X\boldsymbol{)}$ that are not dynamic variables in $G_x^{(0)}$ must pass through $X^{(n)}$:

Lemma 4 *The following relation holds:* $\boldsymbol{Fb(}X\boldsymbol{)} \setminus \mathbf{V_{del}}(x) \subseteq \boldsymbol{Anc(}X^{(n)}\boldsymbol{)}_{G_x^{(0)}}$.

Proof: First note that if V is a dynamic variable, then in $G_x^{(0)}$, by construction

V must be given by initial conditions and so must be exogenous. Therefore, in the chain of derivatives:

$$X^{(n)} \to X^{(n-1)} \to \cdots \to X$$

all $X^{(i)}$ such that $i \neq n$ must have a single parent which is connected by an integration link. Therefore, all $V \in \boldsymbol{Anc}(X)_G \setminus \mathbf{V_{del}}(X)$ must be ancestors of $X^{(n)}$, i.e., $\boldsymbol{Fb}(X) \setminus \mathbf{V_{del}}(X) \subseteq \boldsymbol{Anc}(X^{(n)})_{G_X^{(0)}}$. □

Lemma 5 *If $M_{\tilde{x}} \in \mathcal{F}(M, X)$ then there does not exist an $X^{(i)}$ such that $X^{(i)} \in \boldsymbol{Ch}(X)_G$.*

Proof: First note that the result follows for all $X^{(j)}$ such that $j < n$, because by construction $\boldsymbol{Pa}(X^{(j)}) = \{X^{(j+1)}\}$ in M. Thus I only need to prove that $X^{(n)} \notin \boldsymbol{Ch}(X)$. I prove this result by contradiction. Define $\{\mathbf{V_{pre}}, \mathbf{V_{fb}}, \mathbf{V_{post}}\}$ and $\{\mathbf{E_{pre}}, \mathbf{E_{fb}}, \mathbf{E_{post}}\}$ as in the proof of Lemma 3. $M_{\tilde{x}}$ is recursive by assumption; therefore there only exists one causal mapping, $\phi_{\tilde{x}}$, by Lemma 1. I show that if $X \in \boldsymbol{Pa}(X^{(n)})_G$ then $G_{\tilde{x}}$ will not be acyclic, which violates the assumption that $M_{\tilde{x}} \in \mathcal{F}(M, X)$.

According to Lemma 3, $\phi_{\tilde{x}}(V) = \phi(V)$ for all $V \in \mathbf{V_{pre}} \cup \mathbf{V_{post}}$. I therefore only need to construct a mapping over $\mathbf{V_{fb}}$. Assume $X \in \boldsymbol{Pa}(X^{(n)})_G$ and let $\langle X^{(n)}, E^{(n)}\rangle, \langle X, E_x\rangle \in \phi$. Then when $X^{(n)}$ is removed from the set of variables during equilibration, $E^{(n)}$ will be of the form $f(X, V'_{fb}, P) = 0$, where $V'_{fb} \subset \mathbf{V_{fb}}$ and P is a set of variables that are not in $\boldsymbol{Fb}(X)$. During equilibration E_x will be removed from the model, thus X can be assigned to $E^{(n)}$. Furthermore, all remaining variables in $\boldsymbol{Fb}(X)$ can be associated with their original equations in ϕ. Let $\phi' : \mathbf{V_{\tilde{x}}} \to \mathbf{E_{\tilde{x}}}$ be the causal mapping defined such that $\phi'(V) = \phi(V)$ for all $V \neq X$ and $\phi'(X) = E^{(n)}$. This forms a valid causal mapping and therefore by uniqueness $\phi' = \phi_{\tilde{x}}$. By Lemma 4, since $\boldsymbol{Fb}(X) \setminus \boldsymbol{V_{del}}(X) \subset \boldsymbol{Anc}(X^{(n)})_{G_{\tilde{x}}^{(0)}}$ it must be the case that $\boldsymbol{Fb}(X) \setminus \boldsymbol{V_{del}}(X) \subset \boldsymbol{Anc}(X)_{G_{\tilde{x}}}$, because the equation that was assigned to $X^{(n)}$ is now assigned to X and all the remaining associations are unchanged. However, since $M_{\tilde{x}} \in \mathcal{F}(M, X)$ it must be the case that $\{\boldsymbol{Fb}(X) \setminus \boldsymbol{V_{del}}(X)\} \cap \boldsymbol{Ch}(X) \neq \emptyset$, therefore, there exists an $f \in \boldsymbol{Fb}(X) \setminus \boldsymbol{V_{del}}(X)$ such that $X \in \boldsymbol{Pa}(f)$. Thus $G_{\tilde{x}}$ is cyclic, which is a contradiction. □

Lemma 6 *If $M_{\tilde{x}} \in \mathcal{F}(M, X)$, then there exists a $V \in \mathbf{V}_{\tilde{\mathbf{x}}}$ such that $V \in \boldsymbol{Pa}(X)|_{G_{\tilde{x}}}$ and such that $V \in \boldsymbol{Ch}(X)|_G$.*

Proof: Define an ordering O for ϕ and label the pairs $\langle V_l, E_l \rangle$ in ϕ according to O as in the proof of Lemmas 1 and 3. Let $\langle X, E_i \rangle$ be the association for X in $\phi_{\tilde{x}}$. By construction $X \neq V_i$, and by Lemma 3, $V_i \in \boldsymbol{Fb}(X)$. Since $X \in \boldsymbol{Params}(E_i)$ and since $\langle V_i, E_i \rangle \in \phi$ it must be the case that $V_i \in \boldsymbol{Ch}(X)_G$. Since $X^{(l)}$ is exogenous in $G_x^{(0)}$ for all $l \neq n$ and since, by Lemma 5, $V_i \neq V^{(n)}$, it follows that $V_i \notin \mathbf{V_{del}}(X)$. Therefore $V_i \in \boldsymbol{Fb}(X) \setminus \mathbf{V_{del}}(X)$, and since $V_i \in \boldsymbol{Params}(E_i)$ it must be the case that $V_i \in \boldsymbol{Pa}(X)_{G_{\tilde{x}}}$. □

Lemma 7 *If $M_{\tilde{x}} \in \mathcal{F}(M, X)$ and $M_{\hat{x}} = \langle \langle \mathbf{V}_{\hat{\mathbf{x}}}, \mathbf{E}_{\hat{\mathbf{x}}} \rangle, \phi_{\hat{x}} \rangle$, with causal graph $G_{\hat{x}}$, is the causal model resulting when X is manipulated in M, then in $G_{\hat{x}}$ there will exist an edge $X \rightarrow V$ for all $V \in \boldsymbol{Ch}(X)_G \cap \mathbf{V}_{\hat{\mathbf{x}}}$.*

Proof: When applying the *Do* operator to M, the only arcs that will be removed from G when X is manipulated will be the arcs coming into X and into X's derivatives $X^{(i)}$. Since by Lemma 5, X is not a parent of any $X^{(i)}$ the children of X must be preserved in $G_{\hat{x}}$. □

Finally, Theorem 6 presents conditions which are sufficient for $M_{\tilde{x}}$ to violate the EMC property.

Theorem 6 (change-of-structure) *If $M_{\tilde{x}} \in \mathcal{F}(M, X)$ then the EMC property does not hold for $M_{\tilde{x}}$.*

Proof: Manipulating X in M produces an equilibrium model with respect to X (since X will be manipulated to a constant value by definition of the *Do* operator), $M_{\hat{x}}$, so the subsequent applying of the *Equilibration* operator will not change the causal graph. Let $G_{\hat{x}}$ be the causal graph corresponding to $M_{\hat{x}}$. Since $M_{\tilde{x}} \in \mathcal{F}(M, X)$, by Lemma 6 there exists a $V \in \boldsymbol{Ch}(X)_G$ such that $V \rightarrow X$ in $G_{\tilde{x}}$; however, according to Lemma 7, the edge $X \rightarrow V$ must exist in $G_{\hat{x}}$. Thus, manipulating X in $G_{\tilde{x}}$ leads to a graph $G_{\tilde{x}\hat{x}} \neq G_{\hat{x}}$, because it will not contain an edge between V and X. □

The theorem's proof relies on the guaranteed reversal of an arc; nonetheless, it is clear by the examples given in Section 2.1 that there is more complex behavior being exhibited in these systems than mere reversibility.

The following theorem proves that hidden dynamic instabilities are a mathematical feature of some equilibrium causal models:

Theorem 7 (instability) *If $M_{\tilde{x}} \in \mathcal{F}(M, X)$ and the Structural Stability condition holds then there exists a set of variables $\mathbf{V}' \subset \mathbf{V}_{\tilde{\mathbf{x}}}$ such that if $\mathbf{V}'$ is manipulated in M, the variable X will become unstable.*

Proof: Define $\mathbf{V}' \equiv \boldsymbol{Fb}(X) \setminus \boldsymbol{V_{del}}(X)$. It must be the case that $\mathbf{V}' \neq \emptyset$ by definition of $\mathcal{F}(M, X)$. Manipulating $\mathbf{V}'$ in G will create a new graph $G_{\hat{V}'}$ with $\boldsymbol{Fb}(X)_{G_{\hat{V}'}} = \emptyset$. Therefore, according to the Structural Stability principle, X will be unstable in $G_{\hat{V}'}$.
□

2.3.2 Guaranteed Obeyance of EMC

It was observed in Section 2.2 that equilibrating some variables in the bathtub example did not result in an equilibrium structure that contradicted the dynamic structure. The following theorems formalize that observation by showing that if the variable being equilibrated is self-regulating (see Definition 27), then the EMC property will hold. The next three theorems temporarily suspend the common notation for the symbols X and M defined above in Section 2.3.

The following theorem shows that the *Do* operator commutes with itself:

Theorem 8 *The Do operator commutes with itself, i.e., for a causal model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$, $Do(Do(M, X), Y) = Do(Do(M, Y), X)$ for $X, Y \in \mathbf{V}$.*

Proof: Both $Do(Do(M, X), Y)$ and $Do(Do(M, Y), X)$ correspond to model M with $\phi(X)$ and $\phi(Y)$ replaced with equations of the form $X = x_0$ and $Y = y_0$, respectively, so they are identical. □

Theorem 8 justifies the use of the *Do* operator on sets of variables rather than just singletons.

Next I show that the *Aggregation* operator commutes with itself:

Theorem 9 *The Aggregation operator commutes with itself, i.e., for a causal model* $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$, $Agg(Agg(M, X), Y) = Agg(Agg(M, Y), X)$ *for* $X, Y \in \mathbf{V}$.

Proof: Let $B \in \mathbf{V}$ be an arbitrary variable such that $B \neq X$ and $B \neq Y$. Write $\phi(X)$, $\phi(Y)$ and $\phi(B)$ as $X = f_x(\mathbf{P_x})$, $Y = f_y(\mathbf{P_y})$ and $B = f_b(\mathbf{P_b})$ respectively. In general, if $\mathbf{P_w}$ is the parent set of a variable W, I use the notation $\mathbf{P_{w|z}}$ to denote the parent set of W after a variable Z has been aggregated. I thus need to show that $\mathbf{P_{b|xy}} = \mathbf{P_{b|yx}}$. There are four possibilities: $\mathbf{P_b} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{Y, \mathbf{P}\}$, or $\{X, Y, \mathbf{P}\}$, where $\mathbf{P} \subset \mathbf{V}$ is such that $X \notin \mathbf{P}$ and $Y \notin \mathbf{P}$. Thus, $\mathbf{P_{b|x}} = \{\mathbf{P}\}$, $\{\mathbf{P_x}, \mathbf{P}\}$, $\{Y, \mathbf{P}\}$, or $\{\mathbf{P_x}, Y, \mathbf{P}\}$, respectively, and $\mathbf{P_{b|xy}} = \{\mathbf{P}\}$, $\{\mathbf{P_{x|y}}, \mathbf{P}\}$, $\{\mathbf{P_{y|x}}, \mathbf{P}\}$, or $\{\mathbf{P_{x|y}}, \mathbf{P_{y|x}}, \mathbf{P}\}$, respectively. Repeating this argument for $\mathbf{P_{b|yx}}$ will yield the same four sets, so regardless of which set corresponds to $\mathbf{P_b}$, it will be the case that $\mathbf{P_{b|yx}} = \mathbf{P_{b|xy}}$. □

Again, this theorem allows us to unambiguously apply the *Aggregation* operator to sets of variables, rather than constraining it to singletons.

The next lemma shows that the *Aggregation* operator commutes with the *Do* operator as long as they are not applied to the same variable:

Lemma 8 *The Aggregation operator commutes with the Do operator, i.e., for a causal model* $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$ *and two variables* $X, Y \in \mathbf{V}$, *if* $X \neq Y$ *then* $Agg(Do(M, X), Y) = Do(Agg(M, Y), X)$.

Proof: Let $B \in \mathbf{V}$ be an arbitrary variable such that $B \neq X$ and $B \neq Y$. Write $\phi(Y)$ and $\phi(B)$ as $Y = f_y(\mathbf{P_y})$ and $B = f_b(\mathbf{P_b})$, respectively. For some set $\mathbf{P}$, I use

$\mathbf{P}_{\hat{x}}$ to denote $\mathbf{P}$ after X has been manipulated, and $\mathbf{P}_{\overline{y}}$ to denote $\mathbf{P}$ after Y has been aggregated out. I thus need to show that $\mathbf{P}_{\mathbf{b}|\hat{\mathbf{x}}\overline{\mathbf{y}}} = \mathbf{P}_{\mathbf{b}|\overline{\mathbf{y}}\hat{\mathbf{x}}}$.

Mirroring the proof of Theorem 9, there are four possibilities: $\mathbf{P}_{\mathbf{b}} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{Y, \mathbf{P}\}$, or $\{X, Y, \mathbf{P}\}$, where $\mathbf{P} \subset \mathbf{V}$ is such that $X \notin \mathbf{P}$ and $Y \notin \mathbf{P}$. Thus, $\mathbf{P}_{\mathbf{b}|\hat{\mathbf{x}}} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{Y, \mathbf{P}\}$, or $\{X, Y, \mathbf{P}\}$, respectively, and $\mathbf{P}_{\mathbf{b}|\hat{\mathbf{x}}\overline{\mathbf{y}}} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{\mathbf{P}_{\mathbf{y}|\hat{\mathbf{x}}}, \mathbf{P}\}$, or $\{X, \mathbf{P}_{\mathbf{y}|\hat{\mathbf{x}}}, \mathbf{P}\}$, respectively. On the other hand, $\mathbf{P}_{\mathbf{b}|\overline{\mathbf{y}}} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{\mathbf{P}_{\mathbf{y}}, \mathbf{P}\}$, or $\{X, \mathbf{P}_{\mathbf{y}}, \mathbf{P}\}$, respectively, but $\mathbf{P}_{\mathbf{b}|\overline{\mathbf{y}}\hat{\mathbf{x}}}$ is also equal to $\mathbf{P}_{\mathbf{b}|\overline{\mathbf{y}}} = \{\mathbf{P}\}$, $\{X, \mathbf{P}\}$, $\{\mathbf{P}_{\mathbf{y}}, \mathbf{P}\}$, or $\{X, \mathbf{P}_{\mathbf{y}}, \mathbf{P}\}$. However, if $Y \neq X$ then $\mathbf{P}_{\mathbf{y}|\hat{\mathbf{x}}} = \mathbf{P}_{\mathbf{y}}$, so $\mathbf{P}_{\mathbf{b}|\hat{\mathbf{x}}\overline{\mathbf{y}}} = \mathbf{P}_{\mathbf{b}|\overline{\mathbf{y}}\hat{\mathbf{x}}}$. □

Finally, returning to the notation defined in Section 2.3, the following theorem shows that equilibrating X will commute with the *Do* operator if X is a self-regulating variable:

Theorem 10 (self-regulation) *If X is a self-regulating variable, then there exists a mapping $\phi_{\tilde{x}}$ such that Equilibrate(Do(M, Y), X) = Do(Equilibrate(M, X), Y), for any $Y \in \mathbf{V}_{\tilde{x}}$.*

Proof: Let $\phi_{\tilde{x}}(Y) = \phi(Y)$ for all $Y \in \mathbf{V}_{\tilde{\mathbf{x}}} \setminus X$ and let $\phi_{\tilde{x}}(X) = \phi(X^{(n)})$. With this mapping all variables maintain the same equation except for X which has the equation originally mapped to $X^{(n)}$; thus, $M_{\tilde{x}} = \langle\langle \mathbf{V}_{\tilde{\mathbf{x}}}, \mathbf{E}_{\tilde{\mathbf{x}}}\rangle, \phi_{\tilde{x}}\rangle$ is equivalent to *Agg*(M, $\boldsymbol{V_{del}}(X)$). Since $Y \notin \boldsymbol{V_{del}}(X)$, the result follows by Lemma 8. □

CHAPTER 3

DISCUSSION

In this chapter I discuss the implications of the results presented in the previous chapters. I explore the question of how likely a model is to violate the EMC property, expanding the analysis to nonrecursive models and models that assume latent variables. I also discuss ways in which the results of this thesis can be mitigated, including ways to identify EMC-violating systems, and I discuss the assumptions necessary for learning dynamic models from time-series data.

3.1 The Ubiquity of EMC-Violating Systems

In this section I ask how common an EMC-violating system is to occur. I discuss not just the size of the RFRE class, but other classes of models that also can violate the EMC property.

3.1.1 The Size of the RFRE Class

The most significant restriction placed on a model M, to guarantee that M does not obey the EMC property, is that M be in $\mathcal{F}$. Thus all recursive dynamic models whose equilibrated model is also recursive and who possesses intermediate feedback variables are guaranteed to not obey the EMC property. The condition that the model possess feedback variables is trivial. According to the structural stability principle *all* dynamic systems require feedback in order for stability to occur. Furthermore, if one assumes that all causal interactions require time-lags to occur, then in the real-world there is no such thing as a "simultaneous equation", and all stationary mechanisms are thus equilibrium equations. Under these assumptions, whether or not feedback variables are present in the model has little to do with the system itself and more to do with the modeler.

The condition that the dynamic model be recursive is also very weak. Most, if not all, previous work on dynamic modeling maintains this assumption. The condition that the equilibrium model be recursive is less trivial, but is taken for convenience in theorem-proving only: there certainly exist non-recursive models that do not obey the EMC property (as discussed in Section 3.1.2). Nonetheless, this restriction is also modeler-dependent: any dynamic model such that set of variables $\boldsymbol{Fb}(X) \cap \mathbf{V}_{\mathbf{x}}$ forms a chain from X to $X^{(n)}$:

$$X \rightarrow X^1_{fb} \rightarrow X^2_{fb} \rightarrow \ldots \rightarrow X^{(n)},$$

i.e., such that X has exactly one child in $\boldsymbol{Fb}(X)$ and $X^{(n)}$ has exactly one parent in $\boldsymbol{Fb}(X)$, will produce a recursive model when X is equilibrated.

To give a feel for the scope of these conditions, Table 3.1 provides a sample of physical systems which can possess RFRE models. This table is a virtual laundry list of some of the simplest physical systems known. Significantly, it contains common prototype systems that are used as idealized versions of a host of many other systems which are mathematically isomorphic to it. For example the simple-harmonic oscillator system, which has already been discussed in detail in Section 1.3, is an extremely

System	X	$\{\boldsymbol{Fb}(X) \setminus \boldsymbol{V}_{\boldsymbol{del}}(X)\} \cap \boldsymbol{Ch}(X)$
Ideal-gas	height of piston	pressure
Body in viscous medium	velocity	damping force
Simple harm. oscillator	position of mass	spring force
RC circuit	charge on capacitor	current
Inverting amplifier	output voltage	voltage at neg term

Table 3.1. Examples of physical systems that can be modeled as RFRE models.

general model that is applicable to almost any system where second-order damped oscillations exist. Another general system is the inverting amplifier circuit, shown in Figure 3.1. This system can be an idealized representation for many systems-control problems.

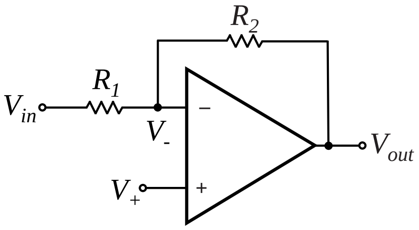


Figure 3.1. The inverting amplifier op-amp circuit.

The operational amplifier (op-amp) is essentially a black-box solid-state device with three terminals, the voltage at each is denoted as V_-, V_+, and V_{out} and two resisters of resistance R_1 and R_2. The device works based on two idealized rules:

1. The input terminals (associated with V_- and V_+) draw no current.
2. The output voltage is proportional (with high gain) to the difference $V_+ - V_-$.

Thus, when an op-amp possesses *negative feedback*, it tends to bring the voltage difference at the input terminals to zero, leading to a steady-state. The set of equations for this system are as follows: First, assume that V_{in}, V_+, R_1 and R_2 are exogenous: $V_{in} = v_{in}$, $V_+ = v_+$, $R_1 = r_1$, and $R_2 = r_2$. Using conservation of current and Ohm's law, the voltage at the negative terminal can be solved as a function of V_{in}, R_1 and R_2 and V_{out}.

$$V_- = f(V_{in}, V_{out}, R_1, R_2) \tag{3.1}$$

Based on rule number 2 of the op-amp operation, I assume that the op-amp analyzes the current differential in input voltage, and it adjusts the output current by a proportional amount in the next time-step:

$$\Delta V_{out} = \alpha(V_+ - V_-) \tag{3.2}$$

The dynamic causal graph for this system looks like that of Figure 3.2a. Obviously, the assumption of equilibrium yields the model in Figure 3.2b, and thus this system produces an RFRE model.

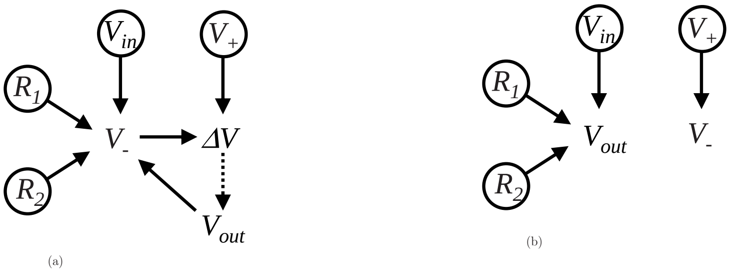


Figure 3.2. The dynamic and equilibrium causal models for the inverting amplifier.

3.1.2 Non-Recursive Models

Exacerbating the ubiquity of the RFRE class is the fact that membership in $\mathcal{F}$ is only sufficient, not necessary for violation of the EMC property. There exist feedback-resolved non-recursive models, for example, that also do not obey EMC. Consider the dynamic model shown in Figure 3.3. When this model is equilibrated there are two possible equilibrium causal structures, both of which include an arc from $F_1 \rightarrow X$ even though in the dynamic model X was a parent of F_1. Thus, any equilibrated model for this system that includes F_1 will display reversibility when X is manipulated.

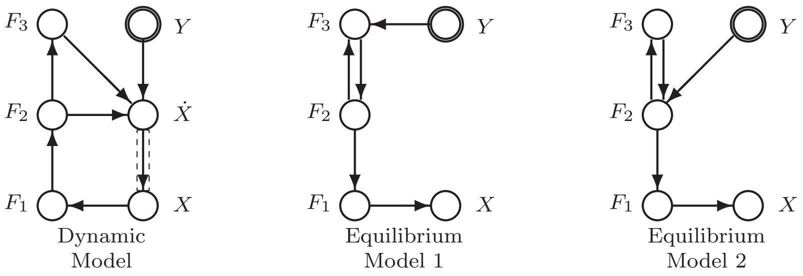


Figure 3.3. Some non-recursive feedback-resolved equilibrated models will also exhibit reversibility.

Again, the culprit in this system is the presence of feedback resolution in the equilibrium models. This observation might lead one to make the following generalization:

Conjecture 1 *Any feedback-resolved equilibrated model will violate the EMC property.*

This conjecture is certainly false, however, as can be seen by the counterexample of a self-regulating feedback-resolved equilibrated model, depicted in Figure 3.4a. The equilibrated structure of this example, shown in Figure 3.4b, is the aggregate graph of the dynamic structure. In fact, Figure 3.4 illustrates the essence of why self-regulating variables produce well-behaved equilibrated models. “Self-regulation” is another way of saying that there exists at least one feedback loop that is not resolved. That is not to

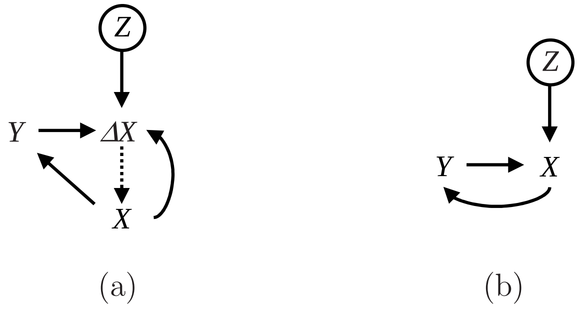


Figure 3.4. A self-regulating feedback-resolved equilibrated model will obey the EMC property.

say that the model in Figure 3.4-b is guaranteed to be stable when Y is manipulated, or Z for that matter; nonetheless, this model is not guaranteed to be unstable or to violate the EMC property because of the potential for the self-regulation path to maintain stability.

With this argument in mind, a plausible conjecture might be proposed. For a model $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$, I define a *feedback-resolved variable* $V \in \mathbf{V}$ as a dynamic variable with feedback variables present in $\mathbf{V}$.

Conjecture 2 *Given a dynamic causal model* $M = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$*, if* $V \in \mathbf{V}$ *is a feedback-resolved variable that is not self-regulating, then equilibrating* V *will result in a model that violates the EMC property.*

I have not yet found either a counter-example to, or a proof for, this conjecture.

3.1.3 Latent-Variable Models

Throughout this thesis, I have assumed that the models under discussion are not being influenced by the presence of latent variables. This assumption was made to show that even this simplistic case was plagued with problems under manipulation. One would expect that the presence of latent variables would only confound the issue more.

Nonetheless, it is a valid question to ask if it is possible to avoid violations of the EMC property by assuming the presence of latent variables when learning a causal model from data. Can the more stringent detection of causality given by a partial ancestral graph still be found to violate EMC based on underlying dynamics? The answer is "yes". As an example, consider the simple dynamic model given in Figure 3.5. Assuming Figure 3.5a to be the true causal system, then the equilibrated system is given by Figure 3.5b. Using the FCI algorithm for causal discovery (i.e., one that does not make the assumption of absence of latent variables), one would learn the partial ancestral graph shown in Figure 3.5c (see Scheines *et al.*, 1999, for

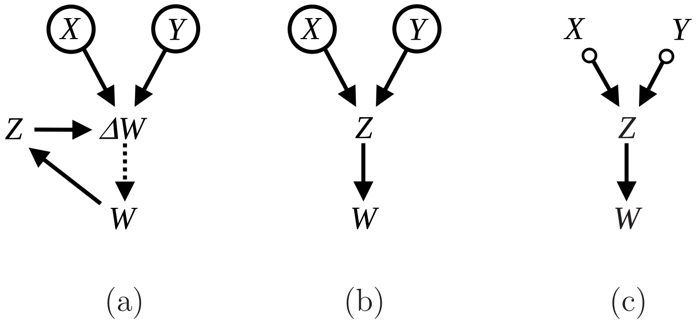


Figure 3.5. Allowing latent variables will not guarantee a learned model will obey the EMC property. If the true causal graph is given by (a), the equilibrated graph will be given by (b), and the FCI algorithm applied to data learned from this graph will return the PAG (c).

example). Thus, in this example, we would conclude that Z is an ancestor of W, and we would predict that manipulating W could not have an impact on Z, which is incorrect according to Figure 3.5a.

3.2 Learning Differential Models

The obvious conclusion to this work is that, rather than learning equilibrium causal models, we should attempt to learn dynamic models, differential models ideally. In this section, I explore the assumptions that are necessary to learn a dynamic causal model as it has been defined in this thesis.

The shorthand dynamic graph representation makes the assumptions that all causal structure between time-slices occurs through derivatives, and that the instantaneous causal structure at a fixed time slice is stationary.

Assumption 4 (stationary structure) *A variable X causes a variable Y at time slice $t = l$ if and only if X causes Y at time slice $t = 0$.*

Assumption 5 (differential causation) *All causation between time-slices occurs to a variable from its first-derivative.*

In addition, one assumption often made is that in reality causation always takes a time-lag to occur:

Assumption 6 (non-simultaneous causation) *If X is a cause of Y, (for example, expressed by some invertible mechanism $f(X, Y) = 0$) and if X is perturbed from value x_0 to some value x_1 at time $t = 0$, then Y will achieve its final value (e.g., $f^{-1}(x_1)$) at time $t' > t$.*

The non-simultaneous causation assumption adds bewilderment to the prospect of learning a differential model. If there exist *no* instantaneous mechanisms, then it would be impossible to learn any structure at all with instantaneously gathered data, because for the system to evolve in time, there must exist *some*-order derivative $V^{(n)}$ for each $V \in \mathbf{V}$ for which an instantaneous mechanism exists. Thus, given the three assumptions just stated, any model learned from truly instantaneous data would in principle result in a complete lack of dependence between the variables (as in the structure S_1 of the filling-bathtub system of Figure 2.20). This apparent paradox is resolved by the reality that we never in the real-world have data that is truly instantaneous. Thus, each set of data has an implicit time lag τ over which it will often be the case that very fast mechanisms will be able to equilibrate, even as τ is made to be as small as experimentally possible.

If this set of assumptions is accepted, then one must conclude that *all* learned models are, in fact, equilibrium models. In that case every learned mechanism would be either an equilibrium one or a transient one. If the mechanism is an equilibrium one, then the variable corresponding to it must be either self-regulating or feedback-resolved (or both); and we are thus at risk of violating EMC in the ways presented in this thesis. Transient mechanisms could be used to form a coherent causal model if the transience time is very long.

Assuming that all mechanisms are equilibrium mechanisms, is there then any benefit to using finely-spaced time-series data to learn a dynamic model? I propose that even in this case it is beneficial, based on the following argument: Let X be the only feedback resolved variable in a causal model, and let $\mathbf{Y}$ denote the set of (self-regulating) feedback variables of X. It is obviously not likely for X to equilibrate before any $Y \in \mathbf{Y}$. If some $Y \in \mathbf{Y}$ is changing, it will in general cause X to come out of equilibrium since it will cause a derivative of X to change as well. Thus, we expect feedback-resolved variables to take much longer to equilibrate than the self-regulating variables. Since we are only concerned with discovering the feedback-resolved relationships, then learning with the finest possible time data will be beneficial.

There are many open questions regarding how to best learn a dynamic graph from time-series data. There already exist many algorithms for learning causal structure and for learning dynamic models in general. However, the assumptions behind the models considered here may allow for more accurate or more efficient algorithms to be developed tailored to this representation.

One theoretical solution is to take a Bayesian approach to this problem. This would entail the following: (1) Establish a set of (maybe non-informative) priors on the space of all possible dynamic models, (2) Given some data, update beliefs about these dynamic models. Finally, (3) use these updated beliefs to calculate the expected effect of manipulation on the system by weighting the prediction of each dynamic model by that model's posterior probability. As with many Bayesian solutions, this one is both comprehensive and probably impossible to achieve in practice, but it could form the basis for a reasonable Bayesian approximate solution.

3.3 Future Directions

This work raises many new open research questions. Obviously these results raise the importance of being able to learn dynamic causal graphs from time-series data. However, another alternative to focusing efforts on constructing dynamic models would be instead to attempt to more precisely characterize the relationship between a dynamic model and its equilibrium counterpart. In this way it may be possible to extract general rules for when an equilibrium model will obey the EMC property. Among the questions to be answered are:

Question 1 *Is Conjecture 2 correct?*

More generally,

Question 2 *Do there exist general necessary and sufficient conditions for an equilibrium model to obey EMC?*

Constructing a dynamic model instead of a non-dynamic model requires the modeler to know more about the system being modeled; for example one might need detailed time-series data. However, maybe it is possible to ensure that a model will support manipulation with less than full time-series data, motivating the following question:

Question 3 *What is the minimal information needed to insure that a model will obey EMC?*

An example is knowledge of temporal disjunction, i.e., the knowledge that two variables do not completely co-vary in time (for example, if I can observe that one variable achieves its equilibrium value before another variable). Obviously all variables in $\boldsymbol{Fb}(X)$ must co-vary in time with X. Thus, it should be safe (assuming stability) to manipulate a variable that is known to be temporally disjoint from all other variables.

Instabilities seem to be a most serious problem with using learned models for manipulation. An important and hard question is:

Question 4 *Under what circumstances will a manipulation cause an instability?*

This question seems especially difficult to answer. It seems that even using a differential model, considering causal structure alone is insufficient to determine if a manipulation will cause an instability. For example, a feedback loop that provides positive (as opposed to negative) feedback will cause an instability to occur. Even manipulating non-feedback variables can cause instabilities. Imagine placing a large negative mass (e.g., a big helium balloon) on the piston in the ideal-gas system. Such a manipulation will obviously cause the height of the piston to increase indefinitely, with no available force to oppose the negative mass. Thus, detection of instabilities seems to require quantitative information about a dynamic system.

Although this thesis raises important objections to some uses of causal reasoning with models learned from data, I believe that the great potential of causal modeling and causal discovery in artificial intelligence make it all the more important for these questions to be explored further and answered as forcefully as possible. The fact that equilibrium causal models can not be proven to support causal inference should not deter us from building and using them in practice, any more than the fact that causality itself is not based in logical reason should dissuade us from employing it in our everyday organization of phenomena.

The fact that taking path A in Figure 1 produces predictions that differ from path B requires us, if we intend to perform causal reasoning with our model, to either ensure that we are taking path B or ensure that we are dealing with models that obey the EMC property. Currently, most work regarding the discovery or building of causal models takes path A and pays no regard to the EMC property. I hope that this thesis will bring attention to this fact and help to rectify it.

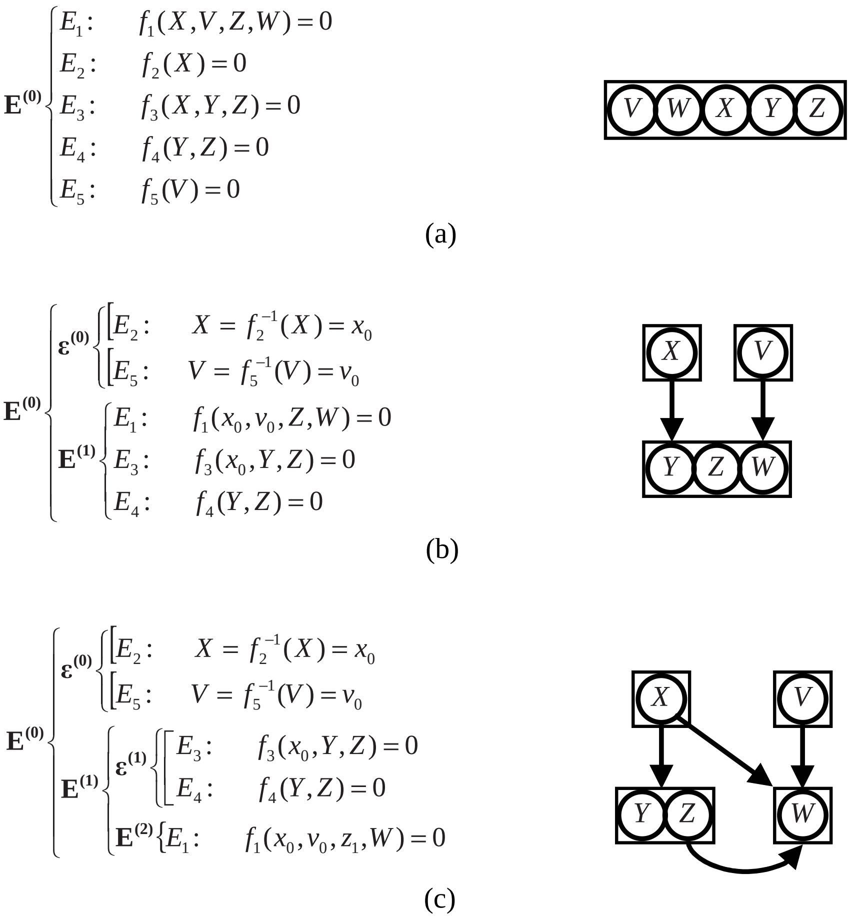
APPENDIX A
BACKGROUND CONCEPTS

A.1 The Causal Ordering Algorithm

Simon [1953] showed that an unmapped SEM $S = \langle \mathbf{E}, \mathbf{V} \rangle$ implies asymmetries among the variables in $\mathbf{V}$, and he proposed an algorithm for extracting these asymmetries in terms of a directed graph. This algorithm is known as the *causal ordering algorithm* (COA). In this section I illustrate the idea and terminology of the algorithm, first through an example and then formally.

Consider the set of equations $\mathbf{E^{(0)}}$ in Figure A.1a. $\mathbf{E^{(0)}}$ constrains the set of variables, $\mathbf{V} = \{X, Y, Z, W, V\}$ with five completely invertible functional relations. Informally, COA orders the variables in an SEM according to the order in which they can be solved in terms of constant parameters of the equations. In Figure A.1a, the single equation E_2 can be inverted to find a unique solution for variable X. This equation by itself thus forms a trivial SEM for the single variable X. Since it is the smallest set of equations that defines a solution for a subset of $\mathbf{V}$, it is a *minimal self-contained subset*. Likewise E_5 alone defines a minimal SEM for variable V, so for $\mathbf{E^{(0)}}$ there are two minimal self-contained subsets. I denote this set of subsets as $\boldsymbol{\mathcal{E}}^{\mathbf{(0)}} = \{\{E_2\}, \{E_5\}\}$.

The causal ordering algorithm proceeds by generating a new SEM, $\mathbf{E^{(1)}}$, by substituting the solutions for X and V (denoted as x_0 and v_0, respectively) into the remaining equations and removing equations E_2 and E_5. I use the terminology that X and V are *exogenous to the subset* $\{W, Y, Z\}$ meaning that X and V are determined before the set $\{W, Y, Z\}$. In $\mathbf{E^{(1)}}$, no variables can be determined within a



$$\mathbf{E}^{(0)}\begin{cases} E_1: & f_1(X,V,Z,W)=0 \\ E_2: & f_2(X)=0 \\ E_3: & f_3(X,Y,Z)=0 \\ E_4: & f_4(Y,Z)=0 \\ E_5: & f_5(V)=0 \end{cases}$$

(a)

$$\mathbf{E}^{(0)}\begin{cases} \boldsymbol{\varepsilon}^{(0)}\begin{cases} \big[E_2: & X=f_2^{-1}(X)=x_0 \\ \big[E_5: & V=f_5^{-1}(V)=v_0 \end{cases} \\ \mathbf{E}^{(1)}\begin{cases} E_1: & f_1(x_0,v_0,Z,W)=0 \\ E_3: & f_3(x_0,Y,Z)=0 \\ E_4: & f_4(Y,Z)=0 \end{cases} \end{cases}$$

(b)

$$\mathbf{E}^{(0)}\begin{cases} \boldsymbol{\varepsilon}^{(0)}\begin{cases} \big[E_2: & X=f_2^{-1}(X)=x_0 \\ \big[E_5: & V=f_5^{-1}(V)=v_0 \end{cases} \\ \mathbf{E}^{(1)}\begin{cases} \boldsymbol{\varepsilon}^{(1)}\begin{cases} \begin{bmatrix} E_3: & f_3(x_0,Y,Z)=0 \\ E_4: & f_4(Y,Z)=0 \end{bmatrix} \end{cases} \\ \mathbf{E}^{(2)}\{E_1: \quad f_1(x_0,v_0,z_1,W)=0 \end{cases} \end{cases}$$

(c)

Figure A.1. The causal ordering algorithm takes a structural equation model as input and outputs a directed partition graph (DPG).

single equation; however, the equations E_3 and E_4 taken together as a subsystem form an SEM for variables Y and Z, for which both can be simultaneously solved in terms of X. The variables Y and Z are said to be *strongly coupled* because they both reside in the same minimal SEM in which they were ultimately determined, and they are collectively represented by a single vertex in the graph, as in Figure A.1c. Since X appears in E_3, it is a parent of the entire subset $\{Y, Z\}$ in the graph. Finally, given X, V, and Z, the equation E_1 alone forms an SEM for W, so X, V and Z are parents of W in the graph.

All graphs produced by COA will be acyclic, but they may possess vertices which represent strongly-couple variables as $\{Y, Z\}$ in Figure A.1c. In fact, the graphs produced by the COA are somewhat unorthodox objects where parents in the graph are variables, but children in the graph are *partitions* of variables:

Definition 33 (partitioning) *A partitioning $\mathcal{V}$ of $\mathbf{V}$ is a set of disjoint sets $\mathcal{V} = \{\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_n}\}$ such that $\bigcup_{i=1}^{n} \mathbf{V_i} = \mathbf{V}$.*

If $\mathcal{V}$ is a partitioning, then we call an arbitrary element $\mathbf{V_i} \in \mathcal{V}$ a *partition* of $\mathcal{V}$. As an example, $\{\{X\}, \{V\}, \{W\}, \{Y, Z\}\}$ is a partitioning of $\{V, W, X, Y, Z\}$. I use the notation $\boldsymbol{Part}(X)_{\mathcal{V}}$ to denote the partition which contains the variable $X \in \mathbf{V}$, although I may drop the $\mathcal{V}$ subscript if the partitioning is clear by the context. I call the graphs produced by the COA "directed partition graphs" (DPGs):

Definition 34 (directed partition graph) *A directed partition graph over a set of variables $\mathbf{V}$ is an ordered pair $\langle \mathcal{V}, \mathbf{A} \rangle$, where the set of vertices $\mathcal{V}$ is a partitioning of $\mathbf{V}$ and $\mathbf{A}$ is a set of directed arcs $X \to \mathbf{V^{(i)}}$ where $X \in \mathbf{V}$ is a variable, $\mathbf{V^{(i)}} \in \mathcal{V}$ is a partition and $\mathbf{V^{(i)}} \neq \boldsymbol{Part}(X)$.*

The causal ordering recursively finds the *minimal self-contained subsets* (e.g., $\{E_2\}$ and $\{E_5\}$ in set $\mathbf{E^{(0)}}$, $\{E_3, E_4\}$ in $\mathbf{E^{(1)}}$ and $\{E_1\}$ in $\mathbf{E^{(2)}}$ in Figure A.1b) of equations and solves them for specific solutions:

Definition 35 (minimal self-contained structure) *If S is a self-contained structure, then S is minimal when there does not exist a subset $S' \subset S$ such that S' is also self-contained.*

Again, I will refer to a set of equations $\mathbf{E}$ as being a minimal self-contained set with respect to a set of variables $\mathbf{V}$ if the SEM $S = \langle \mathbf{V}, \mathbf{E} \rangle$ is a minimal self-contained set.

COA then substitutes the values of the variables determined by all minimal self-contained sets into the remaining equations to get a *derived subset* (e.g., both $\mathbf{E^{(1)}}$ and $\mathbf{E^{(2)}}$ in Figure A.1):

Definition 36 (derived subset) *Let $S = \langle \mathbf{V}, \mathbf{E} \rangle$ be a self-contained structure and let $\mathcal{E}$ be the set of all minimal self-contained subsets of $\mathbf{E}$. Let $\mathbf{E_{sc}} = \bigcup_{\mathbf{E_i} \in \mathcal{E}} \mathbf{E_i}$, and let $\mathbf{E}_{\hookleftarrow}$ denote the set of equations that are obtained when solutions of $\boldsymbol{Params}(\mathbf{E_{sc}})$ given by $\mathbf{E_{sc}}$ are substituted into $\mathbf{E} \setminus \mathbf{E_{sc}}$. $\mathbf{E}_{\hookleftarrow}$ is called the derived subset of $\mathbf{E}$.*

I use the general notation that $\mathbf{E^{(1)}} \equiv \mathbf{E}_{\hookleftarrow}$, $\mathbf{E^{(2)}} \equiv \mathbf{E}_{\hookleftarrow}^{\mathbf{(1)}}$, etc., where $\mathbf{E^{(i)}}$ is called the *derived subset of ith order*. If $\mathbf{E}$ is a set of equations with derived subset $\mathbf{E^{(i)}}$, and if $\mathbf{E}' \subseteq \mathbf{E^{(i)}}$ is some subset of $\mathbf{E^{(i)}}$, then I use $\hat{\mathbf{E}}'$ to denote the subset of $\mathbf{E}$ corresponding to the equations remaining in $\mathbf{E}'$, that is, the subset of original equations with no values substituted.

In Figure A.1, COA constructed a mapping between sets of variables and sets of equations. The mapping could be written as a list of association-pairs as follows:

$$\{\langle\{W\},\{E_1\}\rangle,\ \langle\{X\},\{E_2\}\rangle,\ \langle\{Y,Z\},\{E_3,E_4\}\rangle,\ \langle\{V\},\{E_5\}\rangle\}.$$

I call this a *partial causal mapping*:

Definition 37 (commensurate partitionings) *Let $\mathbf{A}$ and $\mathbf{B}$ be two sets such that $|\mathbf{A}| = |\mathbf{B}|$. A partitioning $\mathcal{P}_A$ over $\mathbf{A}$ is commensurate with a partitioning $\mathcal{P}_B$ over $\mathbf{B}$ iff there exists a one-to-one correspondence $\phi : \mathcal{P}_A \to \mathcal{P}_B$ such that for each set $\mathbf{S_A^{(i)}} \in \mathcal{P}_A$, $|\mathbf{S_A^{(i)}}| = |\phi(\mathbf{S_A^{(i)}})|$.*

Definition 38 (partial causal mapping) *If $\mathbf{E}$ is a set of equations with $\mathbf{V} = \boldsymbol{Params}(\mathbf{E})$, then a partial causal mapping Φ of $\mathbf{E}$ is a triple $\langle \mathcal{V}, \mathcal{E}, \phi \rangle$, where $\mathcal{V}$ is a partitioning of $\mathbf{V}$, $\mathcal{E}$ is a partitioning of $\mathbf{E}$, and ϕ is a bijection, $\phi : \mathcal{V} \to \mathcal{E}$, such that the following is true:*

1. *$\mathcal{V}$ is commensurate with $\mathcal{E}$, and*

2. *For all sets $\mathbf{V^{(i)}} \in \mathcal{V}$ we can match up each variable in $\mathbf{V^{(i)}}$ with a unique equation in $\phi(\mathbf{V^{(i)}})$. That is, there exists a bijection $\phi^{(i)} : \mathbf{V^{(i)}} \to \phi(\mathbf{V^{(i)}})$ such that $X \in \boldsymbol{Params}(\phi^{(i)}(X))$ for all $X \in \mathbf{V^{(i)}}$.*

A partial causal mapping $\Phi = \langle \mathcal{V}, \mathcal{E}, \phi \rangle$ can also be written as a list of ordered pairs or *associations*: $\Phi = \{\langle \mathbf{V^{(1)}}, \mathbf{E^{(1)}} \rangle, \langle \mathbf{V^{(2)}}, \mathbf{E^{(2)}} \rangle, \ldots, \langle \mathbf{V^{(n)}}, \mathbf{E^{(n)}} \rangle\}$, where $\mathbf{V^{(i)}} \in \mathcal{V}$ and $\mathbf{E^{(i)}} \in \mathcal{E}$ are sets for all i, and n is the number of partitions in $\mathcal{V}$. I will use these two representations of a partial causal mapping interchangeably. I define an elementary association as one that maps a single variable to a single equation:

Definition 39 (elementary association) *An association $\langle \mathbf{V^{(i)}}, \mathbf{E^{(i)}} \rangle$ is an elementary association if $|\mathbf{V^{(i)}}| = |\mathbf{E^{(i)}}| = 1$.*

If $\langle \mathbf{V^{(i)}}, \mathbf{E^{(i)}} \rangle$ is an elementary association where $\mathbf{V_p} = \{v\}$ and $E_p = \{E\}$, for clarity of notation I will often write this association as $\langle v, E \rangle$ rather than as $\langle \{v\}, \{E\} \rangle$. A partial causal mapping Φ is obviously isomorphic to a total causal mapping if each association in Φ is an elementary association.

The Causal Ordering Algorithm can now be formally defined:

Definition 40 (Causal Ordering Algorithm) *Given an SEM $S = \langle \mathbf{V^{(i)}}, \mathbf{E^{(i)}} \rangle$, the causal ordering algorithm produces a partial causal mapping $\Phi = \langle \mathcal{V}, \mathcal{E}, \phi \rangle$, where $\mathcal{V}$ is a partitioning of $\mathbf{V^{(i)}}$ and $\mathcal{E}$ is a partitioning of $\mathbf{E^{(i)}}$, through the following procedure:*

1. *Define $\mathcal{E}^{(i)}$ to be the set of all minimal self-contained subsets of $\mathbf{E^{(i)}}$.*

2. *For each set $\mathbf{E} \in \mathcal{E}^{(i)}$ add the association $\langle \boldsymbol{Params}(E), \hat{\mathbf{E}} \rangle$ to Φ.*

3. *Let $\mathbf{E^{(i+1)}} \equiv \mathbf{E}^{\mathbf{(i)}}_{\hookleftarrow}$ and recurse this procedure until $\mathbf{E^{(i+1)}} = \emptyset$.*

A partial causal mapping $\Phi = \langle \mathcal{V}, \mathcal{E}, \phi \rangle$ defines a DPG as follows: For each association $\langle \mathbf{V^{(i)}}, \mathbf{E^{(i)}} \rangle \in \phi$ and for each $E \in \mathbf{E^{(i)}}$ and each $V \in \boldsymbol{Params}(E) \setminus \mathbf{V^{(i)}}$, direct an edge from V to $\mathbf{V^{(i)}}$. This procedure will always produce an acyclic DPG.

A.2 Bayesian Networks

A Bayesian network (BN) is another explicit representation which models causality using a directed graph together with a set of conditional probability distributions.

Because I will often refer to the parent set of an enumerated node X_i, I will use the shorthand notation that $\mathbf{P}_i = \boldsymbol{Pa}(X_i)$. A Bayesian network is defined as:

Definition 41 (Bayesian network) *A Bayesian network is a pair* $\langle G, \mathbf{\Theta} \rangle$*, where* $G = \langle \mathbf{V}, \mathbf{A} \rangle$ *is a directed acyclic graph over nodes* $\mathbf{V}$ *possessing directed arcs* $\mathbf{A}$*, and* $\mathbf{\Theta}$ *is a set of conditional probability tables* $\mathbf{\Theta} = \{\theta_{ijk} = P(V_i = v_i^k \mid \mathbf{Pa_i} = \mathbf{pa_i^j}) : V_i \in \mathbf{V},\ \mathbf{Pa_i} = \mathbf{P}_i\}$.

The parameters $\mathbf{\Theta}$ are often written as a set of conditional probability tables θ_i, corresponding to each node $V_i \in \mathbf{V}$. Each CPT θ_i is in turn defined as a set of distribution functions θ_{ij}, one for each parent configuration $\mathbf{pa_i^j}$ of node V_i.

Like a causal model, a BN is an explicit hypothesis about the causal interactions present between variables; however, the quantification of that interaction takes the form of conditional probability distributions rather than deterministic functions.

If it is assumed that the Markov condition holds between nodes in the network, then a BN specifies the joint probability distribution over the nodes $P(\mathbf{V})$. This is evident based on the fact that any probability distribution over n variables can be factored according to the chain rule of probability:

$$P(\mathbf{X}) = P(X_1) \cdot P(X_2 \mid X_1) \cdot \ldots \cdot P(X_n \mid X_1, X_2, \ldots, X_{n-1}). \tag{A.1}$$

If we use the ordering of the nodes specified by a topological sort of the network, then Equation A.1 together with the Markov condition yields the equation:

$$P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i \mid \mathbf{P}_i), \tag{A.2}$$

each term of which is specified by the definition of a BN.

Unlike functional causal models, representing cyclic causal interactions with a Bayesian-network-like representation allowing cyclic graphs is not straightforward because although cyclic graphs may obey the Markov condition locally, they will not in general obey this condition in a global sense [Spirtes, 1995], which makes the factorization in Equation A.1 non-trivial to produce. It can be argued however, that this is not a drawback of Bayesian networks but rather a drawback of cyclic models. After all, no claim has been made that SEMS are guaranteed to specify the joint distribution for cyclic systems either. Nonetheless, most of the computational benefits of using Bayesian networks rely on the ability to factor the joint distribution according to Equation A.2, so in practice this representation has not been used to model cyclic systems.

A.3 Conditional Independence Models

If, rather than specifying the complete set of causal links directly, one only has information about independence relations that exist in a system, then a conditional independence (CI) model can be constructed:

Definition 42 (conditional independence model) *A conditional independence model is a pair $\langle \mathbf{V}, \mathbf{I} \rangle$, where $\mathbf{V}$ is a set of variables and $\mathbf{I}$ is a list of conditional independence statements of the form $(X \perp Y \mid \mathbf{Z})$ such that $X, Y \in \mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$.*

Even with the addition of the faithfulness condition, the mapping from CI models to BN models is not unique. If M_I is a CI model, then I use the notation $\boldsymbol{\mathcal{G}}(M_I)$ to denote the set of DAGs consistent with M_I according to the Markov and faithfulness conditions. If $G_1, G_2 \in \boldsymbol{\mathcal{G}}(M_I)$ then G_1 and G_2 are said to be *independence equivalent*. Conversely, if G is a DAG, then I use the notation $I(G)$ to indicate the set of independencies entailed by G given the Markov condition (and Theorem 1):

$$I(G) = \{(X \perp Y \mid \mathbf{Z}) \ : \ (X \perp\!\!\!\perp Y \mid \mathbf{Z})_G\}.$$

Obviously G_1 and G_2 are independence equivalent if and only if $I(G_1) = I(G_2)$. I also use the notation $\mathit{Skeleton}(G)$ to denote the set of undirected arcs in G:

$$\mathit{Skeleton}(G) \equiv \{\{A, B\} \ : \ \text{either } A \rightarrow B \text{ or } B \rightarrow A\},$$

and $\mathit{V\text{-}struct}(G)$ to denote the set of v-structures in G:

$$\mathit{V\text{-}struct}(G) \equiv \{\langle A, C, B\rangle \ : \ A \rightarrow C, \ B \rightarrow C, \text{ and } \{A, B\} \notin \mathit{Skeleton}(G)\}.$$

The following theorem was proven by Verma and Pearl [1991]; Spirtes *et al.* [1993]; Chickering [1995]:

Theorem 11 *If G_1 and G_2 are two Bayesian networks, then $I(G_1) = I(G_2)$ if and only if $\mathit{Skeleton}(G_1) = \mathit{Skeleton}(G_2)$ and $\mathit{V\text{-}struct}(G_1) = \mathit{V\text{-}struct}(G_2)$.*

A causal graph derived from a CI model M_I must be consistent with all possible BN models that are represented by M_I. This fact leads to the concept of *CI-based causality*:

Definition 43 (CI causal graph) *Given a CI model M_I over variables $\mathbf{V}$, then a CI causal graph is a partially directed graph $G_p = \langle \mathbf{V}, \mathbf{U}, \mathbf{A}\rangle$ where $\mathbf{U}$ is a set of undirected (ambiguous) edges:*

$$\mathbf{U} = \{\{A, B\} \ : \ (\exists\, G_1, G_2 \in \boldsymbol{\mathcal{G}}(M_I)) \wedge (A \rightarrow B)_{G_1} \wedge (B \rightarrow A)_{G_2}\},$$

and $\mathbf{A}$ is a set of directed (unambiguous) edges:

$$\mathbf{A} = \{(A \rightarrow B) \ : \ (A \rightarrow B)_G \text{ for all } G \in \boldsymbol{\mathcal{G}}(M_I)\}.$$

A CI causal graph is more commonly known as a *pattern* or the more descriptive but less common *essential graph*. A pattern provides a graphical representation of an equivalence class as all CI models in the same equivalence class will produce the same pattern. It is not clear that all patterns map to a unique equivalence class however because some patterns are cyclic and the analysis of cyclic probabilistic models is less mature.

In order to define the procedure used to construct a pattern, I need to introduce some terminology. If $M_I = \langle \mathbf{V}, \mathbf{I} \rangle$ is a CI model then I define the independency set $\boldsymbol{Indep}(M_I)$ to be the set of independent pairs defined by M_I, and I define the dependency set $\boldsymbol{Dep}(M_I)$ to be the set of dependencies implied by M_I and the faithfulness condition:

Definition 44 (independency set) *If $M_I = \langle \mathbf{V}, \mathbf{I} \rangle$ is a CI model then the independency set $\boldsymbol{Indep}(M_I)$ is the set of pairs of variables that can be made independent by a suitable conditioning event: $\boldsymbol{Indep}(M_I) \equiv \{\{X, Y\} \ : \ (X \perp Y \mid \mathbf{Z}) \in \mathbf{I}\}$.*

Definition 45 (dependency set) *If $M_I = \langle \mathbf{V}, \mathbf{I} \rangle$ is a CI model then the dependency set $\boldsymbol{Dep}(M_I)$ is the set of pairs of variables that are never independent in M_I: $\boldsymbol{Dep}(M_I) \equiv \boldsymbol{Pairs}(\mathbf{V}) \setminus \boldsymbol{Indep}(M_I)$.*

Definition 46 (v-structure signature) *In a CI model $M_I = \langle \mathbf{V}, \mathbf{I} \rangle$, a v-structure signature is a pair $\{\{A, C\}, \{B, C\}\}$ such that:*

1. $\{A, C\}, \{B, C\} \in \boldsymbol{Dep}(M_I)$ *and* $\{A, B\} \in \boldsymbol{Indep}(M_I)$,

2. $(A \perp B \mid \mathbf{Z}) \in \mathbf{I} \Rightarrow C \notin \mathbf{Z}$.

The following algorithm can be used to construct a pattern from a CI model M_I [Verma and Pearl, 1991; Spirtes *et al.*, 2000]:

Procedure 1 (pattern construction)

Input: *a CI model $M_I = \langle \mathbf{V}, \mathbf{I} \rangle$.*

Output: *a partially directed graph $G_p = \langle \mathbf{V}, \mathbf{U}, \mathbf{A} \rangle$.*

1. *Set* $\mathbf{U} = \boldsymbol{Dep}(M_I)$.

2. *For every v-structure signature* $\{\{X, Z\}, \{Y, Z\}\}$ *add edges* $(X \rightarrow Z), (Y \rightarrow Z)$ *to* $\mathbf{A}$ *and remove sets* $\{X, Z\}, \{Y, Z\}$ *from* $\mathbf{U}$.

3. *Orient all edges that are required to avoid additional v-structures and to avoid cycles.*

Step 3 can be performed by repeatedly applying the following four rules until no more edges can be added [Verma and Pearl, 1992; Meek, 1995]:

1. For every set $\{B, C\} \in \mathbf{U}$ such that $\exists\ (A \rightarrow B) \in \mathbf{A}$ and $\{A, C\} \notin \mathbf{U}$, orient the edge $B \rightarrow C$.

2. For every set $\{A, B\} \in \mathbf{U}$ such that $\exists\ (A \rightarrow C), (C \rightarrow B) \in \mathbf{A}$, orient the edge $A \rightarrow B$.

3. For every set $\{A, B\} \in \mathbf{U}$ such that $\exists\ \{A, C\}, \{A, D\} \in \mathbf{U}$ and $\exists\ (C \rightarrow B), (D \rightarrow B) \in \mathbf{A}$ such that C and D are non-adjacent, orient the edge $A \rightarrow B$.

4. For every set $\{A, B\} \in \mathbf{U}$ such that $\exists\ \{A, C\} \in \mathbf{U}$ and $\exists\ (C \rightarrow D), (D \rightarrow B) \in \mathbf{A}$ such that C and B are non-adjacent and A and D are adjacent, orient the edge $A \rightarrow B$.

The result of Procedure 1 is a partially directed graph that may or may not contain cycles.

In summary, a conditional independence model M_I (together with the Markov and faithfulness conditions) is a hypothesis about both the dependencies and the independencies among the variables in $\mathbf{V}$. From this model, a conservative definition of causality arises: an arc $A \rightarrow B$ is causal if and only if that arc is present in every BN model that is consistent with M_I. Some may interpret this as a practical definition of a causal arc, while others take the viewpoint that a CI model is incomplete, providing only a set of *necessary* arcs. Given a CI model, the causal graph can be constructed using the deterministic Procedure 1. CI models are especially relevant to the task of learning a causal model because some methods for learning use classical significance tests to search the data for independence relations so that a CI model can be constructed (see Section C.1).

APPENDIX B

TWO CONCEPTIONS OF MANIPULATION

After the manipulation $\mathbf{U}\hat{=}\mathbf{u}$, the marginal probability distribution $P(\mathbf{U})$ will have been altered so, even if $\mathbf{U}$ normally would depend on some set $\mathbf{P}$ of parents, $\mathbf{U}$ will become independent of $\mathbf{P}$:

$$P(\mathbf{U}\hat{=}\mathbf{u}) = P(\mathbf{U}\hat{=}\mathbf{u} \mid \mathbf{P} = \mathbf{p}) = 1 \quad \text{for all } \mathbf{p} \in \mathit{Rng}(\mathbf{P}).$$

Since a manipulation $\mathbf{U}\hat{=}\mathbf{u}$ fixes the distribution of $\mathbf{U}$, it is intuitive to suppose that a causal graph G of such a manipulated system will be such that $\boldsymbol{Pa}(\mathbf{U})_G = \emptyset$ (this intuition is, in fact, one justification for the causal faithfulness assumption: we are asserting that since $\mathbf{U}$ does not depend on $\mathbf{P}$ then that fact must be reflected in the structure of the causal graph).

Whatever other changes may occur to the causal graph under a manipulation are not necessarily specified by the intuitive notion of manipulation, and there are at least two differing opinions in the literature as to how the rest of the causal graph will be affected. By far the most common assumption is that the remainder of the model is unaffected, I refer to this view as the *arc-cutting* account of manipulation. The other viewpoint argues that the response of a system to manipulation is more complex and requires the COA for elucidation. I refer to this viewpoint as the *mechanism-altering* account of manipulation.

In the arc-cutting account, the fundamental knowledge of a system consists of causal parent-child relationships: these are expressed in terms of a causal model

$S = \langle\langle \mathbf{V}, \mathbf{E}\rangle, \phi\rangle$, where the function determining each variable $X \in \mathbf{V}$ is explicated by the mapping ϕ. Manipulating X is defined by replacing the equation $\phi(\mathbf{X})$ with a new equation, $X = x_0$, specifying the manipulated value of X. On the other hand, the mechanism-altering account, put forth by Simon [1953], defines manipulation on an *unmapped* SEM also as the altering of equations which constitute the fundamental building blocks of knowledge. Manipulation thus again involves striking equations from the model and replacing them with new equations. The difference between this approach and the arc-cutting approach is that the equation that is struck can be different in the two cases:

Definition 47 (*Alter* operator) *If $S = \langle \mathbf{V}, \mathbf{E}\rangle$ is an SEM, and $\mathbf{E_{del}}$ is a set of equations $\mathbf{E_{del}} \subseteq \mathbf{E}$, $\mathbf{E_{add}}$ is a set of equations such that $\mathbf{E}' \equiv \mathbf{E_{add}} \cup \{\mathbf{E} \setminus \mathbf{E_{del}}\}$ is self-contained with respect to $\mathbf{V}$, then Alter$(S, \mathbf{E_{del}}, \mathbf{E_{add}})$ is the SEM $S' = \langle \mathbf{V}, \mathbf{E}'\rangle$.*

In order to be consistent with the intuitive idea of manipulating a set $\mathbf{U} \hat{=} \mathbf{u}$, the list of equations to add, $\mathbf{E_{add}}$, obviously must take the form of $\mathbf{U} = \mathbf{u}$.

The *Do* operator is related to the *Alter* operator. In particular, given an unmapped SEM $S = \langle \mathbf{V}, \mathbf{E}\rangle$ and a causal model $S_m = \langle S, \phi\rangle$, performing $Do(G, V = v)$ on a variable $V \in \mathbf{V}$ in the causal graph G of S_m corresponds to *Alter*$(S, \mathbf{E_{del}}, \mathbf{E_{add}})$, where $\mathbf{E_{del}} \equiv \phi(\mathbf{V})$ and $\mathbf{E_{add}} \equiv \{V = v\}$. However, the *Alter* operator is more general in that it does not require the equation that was originally associated with V to be deleted from the set of equations. Because of this generality, the *Alter* operator is capable of modeling manipulations that result in "reversibility" in some systems (discussed in Section B).

It has been observed [Spirtes *et al.*, 1993; Druzdzel and van Leijen, 2001] that some systems appear to exhibit reversibility when manipulated. The standard example of a reversible system is the transmission of a bicycle. In normal operation, the following causal graph describes this system:

$$\textit{Pedal Rotation Rate} \rightarrow \textit{Wheel Rotation Rate;}$$

however, if the bike is propped up on a bike rack, the pedals left free to rotate, and the wheel directly rotated at some rate (in the backwards direction), the pedals will rotate in response. The causal ordering of the system under these circumstances yields:

$$\textit{Pedal Rotation Rate} \leftarrow \textit{Wheel Rotation Rate.}$$

This type of reversibility is present in any physical system where two or more components are connected by a set of rigid gears.

Another type of system that displays reversibility, in a probabilistic sense, occurs when a model is built on a mixed population, where in one subpopulation $\mathcal{A}$, a variable A causes a variable B; whereas in another subpopulation $\mathcal{B}$, B causes A. In this case, when A is manipulated (and not B), A will appear to cause B, and when A is released and B is manipulated, it will appear that B causes A.

Both types of reversibility require that some variables be "released" before reversibility will be present. For example, if the wheels of a bike are fixed to a certain rotation rate without releasing the pedals of the bike (e.g., while holding the pedals in place), then the chain will obviously break, and no reversibility will be apparent. Because of the need to release variables to expose reversibility in a system, the *Do* operator is incapable of modeling this behavior because it only specifies how to manipulate a variable, not how to release it. The *Alter* operator is capable of modeling these types of manipulations because of the added flexibility about which equations can be deleted. For example, if our SEM consists of two equations:

$$P = P_0 \tag{E_1}$$

$$W = \alpha P \tag{E_2}$$

where P denotes the rotation rate of the pedals and W denotes the rotation rate of the rear wheel, then manipulating the wheel and releasing the pedals can be achieved by striking out E_1 and replacing it with E_1':

$$W = W_0 \tag{E_1'}$$

This manipulation cannot be modeled by the *Do* operator because it requires the replacing of an equation for P with an equation for W.

APPENDIX C

CONSTRUCTION OF CAUSAL MODELS

In this section I review some basic approaches to using a database of non-experimental records to infer causal relationships. I discuss the two primary methods in use for this task. In Section C.1 I present the theory behind *constraint-based learning*, and in Section C.2 I discuss Bayesian methods for inferring causal structure.

C.1 Constraint-Based Learning

Constraint-Based (CB) learning methods [Verma and Pearl, 1991; Spirtes and Glymour, 1991; Pearl and Verma, 1991; Spirtes *et al.*, 2000; Cheng *et al.*, 2002] (also known as "conditional independence search" methods) build a CI model by systematically checking the data for conditional independence relations. The ability to infer causal structure from independence information requires the causal Markov and causal faithfulness conditions, defined in Sections A.2 and A.3, respectively. I assume the existence of a standard independence test $I(X, Y \mid \mathbf{Z}, D, \alpha)$, a set of variables $\mathbf{V}$ a complete database D, and a significance level α. I use $\boldsymbol{Adj}(V)$ to denote the set of adjacencies of V, i.e., if G is (in the most general case) a partially directed graph $G = \langle \mathbf{V}, \mathbf{U}, \mathbf{A} \rangle$:

$$\boldsymbol{Adj}(V) \equiv \{X \ : \ (X \to V) \in \mathbf{A} \vee (V \to X) \in \mathbf{A} \vee \{X, V\} \in \mathbf{U}\}.$$

The following straightforward algorithm could be proposed for the construction of a causal graph. I use the notation $\boldsymbol{Pairs}(\mathbf{V})$ to denote the set of all unordered disjoint pairs of $\mathbf{V}$: $\boldsymbol{Pairs}(\mathbf{V}) \equiv \{\{X, Y\} \ : \ X, Y \in \mathbf{V}, \ X \neq Y\}$:

Procedure 2 (simple constraint-based learning)

Input: *Set of variables* $\mathbf{V}$ *and database of values* $D(\mathbf{V})$.

Output: *a partially directed graph* G_p.

0. *Let the set* $\mathbf{I} = \emptyset$.

1. *For each pair* $\{X, Y\} \in \boldsymbol{Pairs}(\mathbf{V})$ *and all subsets* $\mathbf{Z} \subseteq \mathbf{V} \backslash \{X, Y\}$, *if* $I(X, Y \mid \mathbf{Z}, D, \alpha) =$ *true then add* $(X \perp Y \mid \mathbf{Z})$ *to the list* $\mathbf{I}$.

2. *Use Procedure 1 together with the CI model* $M = \langle \mathbf{V}, \mathbf{I} \rangle$ *to construct the causal graph* G_p.

In practice, Procedure 2 is prohibitive because if $|\mathbf{V}| = N$, Step 1 requires the checking of $O(N^{N/2})$ independence tests. A more practical algorithm will grow the independence graph as it checks for independencies so that it can reduce the number of independency checks required by using d-separation information provided by the graph-so-far. The PC algorithm [Spirtes *et al.*, 2000] does this by replacing Step 1 with one that constructs an independence graph by incrementally thinning a completely connected graph by checking various independence relations. The algorithm is sketched as follows:

Procedure 3 (PC algorithm)

Given: $\mathbf{V}$, D and α.

1. S_u = *PC-Find-Independence-Graph*$(\mathbf{V}, D, \alpha)$,

2. S = *Orient-Edges*(S_u, D),

3. *Return* S.

PC-Find-Independence-Graph$(\mathbf{V}, D, \alpha)$ takes a set of variables $\mathbf{V}$ and a database D as input and outputs an undirected graph S_u such that an edge X—Y exists in S_u iff there does not exist a subset $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ (including the empty set) such that

$I(X, Y \mid \mathbf{Z}, D) = true$. S_u is constructed by checking conditional independence relations and removing edges from an initially complete undirected graph whenever an independence is found. The PC algorithm makes this procedure efficient by successively checking higher-order dependencies while restricting the set of nodes that need to be conditioned on. Specifically, let $Adj(A)$ denote the set of variables that are adjacent to A, then *Find-Independence-Graph*$(\mathbf{V}, D, \alpha)$ can be sketched as follows:

Procedure 4 (*PC-Find-Independence-Graph*$(\mathbf{V}, D, \alpha)$)

```
1. Let $n = 0$.
2. Let $S_u$ be a complete undirected graph.
3. Repeat:
    (a) For all pairs of variables $(X, Y)$, check $I(X, Y \mid \mathbf{Z}, D, \alpha)$ for all subsets $\mathbf{Z}$
        such that $|\mathbf{Z}| = n$ and $\mathbf{Z} \subset \boldsymbol{Adj}(X)$ or $\mathbf{Z} \subset \boldsymbol{Adj}(Y)$. If there exists a $\mathbf{Z}$
        such that $I(X, Y \mid \mathbf{Z}, D, \alpha) = true$ then remove the edge $X$—$Y$ from $S_u$.
    (b) Set $n = n + 1$
   Until no variable has greater than $n$ adjacencies.
4. Return $S_u$.
```

The sub-procedure *Orient-Edges*(S_u, D) infers directionality of some arcs in S by searching for independence relations characteristic of v-structures and by avoiding cycles. For example, the four rules for orienting a pattern given in Chapter A.3 can be used.

The graphs produced by CI-based procedures are patterns as defined in Definition 43. As a reminder, patterns summarize the structure of a Bayesian network that can be inferred from a list of independencies alone. A loose-upper bound on the worst-case computational time complexity of PC is $O[n^2(n-1)^{k-1}/(k-1)!]$, where k is the maximum number of adjacencies of any node in the graph. In practice graphical models are typically assumed to be sparse. For example, the space complexity of a Bayesian network is exponential in the largest in-degree in the network. Therefore, this complexity result for the PC algorithm can be quite efficient for many problems considered in practice.

CB methods have the advantage of possessing clear stopping criteria and deterministic, systematic search procedures. On the other hand, they are subject to several instabilities. [Spirtes *et al.*, 2000; Dash and Druzdzel, 1999] Namely, if a mistake is made early on in the search, it can lead to incorrect sets $Adj(A)$ and $Adj(B)$ later in the search, which may in turn lead to bad decisions in the future, which can lead to even more incorrect sets $Adj(A)$, etc. This instability has the potential to cascade, creating many errors in the learned graph. Similarly, incorrect edges in S_u can lead to incorrectly oriented arcs in the final graph S. It is for these reasons that the quality and reliability of the independence test is critical for practical constraint-based algorithms. Unfortunately, when there is little data, when some configurations of variables are unlikely, or when there is missing data, hypothesis tests can be unreliable.

C.2 Bayesian Learning

Rather than finding a single model that maximizes the likelihood of the data D, Bayesian learning in general advocates averaging quantities of interest over all possible models M, weighting each model by its posterior probability $P(M \mid D)$. However, Bayesian-inspired methods exist which allow one to learn a single causal model in the form of a Bayesian network. In terms of Bayesian learning, one primary quantity of interest when using causal models is the joint probability distribution $P(\mathbf{X})$ over the variables $\mathbf{X}$. This chapter will review the techniques and assumptions that are used to allow Bayesian learning to be applied to this problem, and show how they can be simplified into a set of techniques for selection of a single causal model.

I begin by stating the assumptions that are required to make Bayesian learning tractable. First, I assume that all variables are discrete:

Assumption 7 (Multinomial variables) *Each variable $X_i \in \mathbf{V}$ is a discrete variable with r_i possible states $\{x_i^1, x_i^2, \ldots, x_i^{r_i}\}$.*

I let q_i denote the number of possible joint configurations of parents for node X_i, and I enumerate these configurations as $\{p_i^1, p_i^2, \ldots, p_i^{q_i}\}$. I use the shorthand that if Q_{ijk} is some quantity associated with coordinates (ijk), then $Q_{ij} \equiv \sum_k Q_{ijk}$.

I assume that all data is complete, although these techniques can be easily adapted to incomplete data via the EM algorithm [Dempster *et al.*, 1977]

Assumption 8 (Complete Labelled Data) *The training data set D contains no record $D_l \in D$ such that D_l is missing data.*

I let N_{ijk} denote the number of times in the database that the node X_i achieved state k when $\mathbf{P}_i$ was in the j-th configuration.

Assumption 9 (Dirichlet priors) *The prior beliefs over parameter values are given by a Dirichlet distribution.*

I let α_{ijk} denote the Dirichlet hyperparameters corresponding to the network parameter θ_{ijk}.

Assumption 10 (Parameter independence) *For any given network structure S, each probability distribution $\theta_{\mathbf{ij}}$ is independent of any other probability distribution $\theta_{\mathbf{i'j'}}$:*

$$P(\theta \mid S) = \prod_{i=0}^{N} \prod_{j=1}^{q_i} P(\theta_{\mathbf{ij}} \mid S) \tag{C.1}$$

C.2.1 Fixed Network Structure

For a fixed network structure S and a fixed set of network parameters $\mathbf{\Theta}$, the quantity $P(\mathbf{X} = \mathbf{x} \mid S, \mathbf{\Theta})$ can be calculated in $O(N)$ time:

$$P(\mathbf{X} = \mathbf{x} \mid S, \mathbf{\Theta}) = \prod_{i=0}^{N} \theta_{iJK}, \tag{C.2}$$

where all (j, k) coordinates are fixed by the configuration of $\mathbf{X}$ and the structure S to the value $(j, k) = (J, K)$ (J and K should technically be indexed as $J_{\mathbf{x}}^{S}$ and $K_{\mathbf{x}}^{S}$ because they are fixed by the structure and the configuration $\mathbf{x}$, but I refrain from doing this here to simplify the notation).

When, rather than a fixed set of parameters, a database D is given, from an ideal Bayesian perspective it is necessary to average over all possible configurations of the parameters $\mathbf{\Theta}$:

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \mid S, D) &= \int P(\mathbf{X} = \mathbf{x} \mid S, \mathbf{\Theta}) \cdot P(\mathbf{\Theta} \mid S, D) \cdot d\mathbf{\Theta} \\ &= \int \prod_{i=0}^{N} \theta_{iJK} \cdot P(\mathbf{\Theta} \mid S, D) \cdot d\mathbf{\Theta} \end{aligned}$$

where the second line follows from Equation C.2. Given the assumption of parameter independence and Dirichlet priors, this quantity can be written just in terms of sufficient statistics and Dirichlet hyperparameters [Cooper and Herskovits, 1992; Heckerman *et al.*, 1995]:

$$P(\mathbf{X} = \mathbf{x} \mid S, D) \;=\; \prod_{i=0}^{N} \frac{\alpha_{iJK} + N_{iJK}}{\alpha_{iJ} + N_{iJ}} \tag{C.3}$$

Comparing this result to Equation C.2 illustrates the well-known result that a single network with a fixed set of parameters $\hat{\mathbf{\Theta}}$ given by

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \tag{C.4}$$

will produce predictions equivalent to those obtained by averaging over all parameter configurations. [Heckerman, 1998] asserts that under a suitable coordinate transformation these parameters in fact coincide with the *maximum a posteriori* (MAP) configuration.

C.2.2 Unconstrained Averaging over Structures

Again, a holistic Bayesian approach would not consider a single structure, but would instead calculate predictions by averaging over all possible structures, in which case the quantity $P(\mathbf{X} = \mathbf{x} \mid D)$ is given by:

$$\begin{aligned} &P(\mathbf{X} = \mathbf{x} \mid D) \\ &\quad = \sum_S \int P(\mathbf{X} = \mathbf{x} \mid S, \mathbf{\Theta}) \cdot P(\mathbf{\Theta} \mid S, D) \cdot P(S \mid D) \cdot d\mathbf{\Theta} \\ &\quad = \sum_S \prod_{i=0}^{N} \hat{\theta}_{iJK} \cdot P(S \mid D) \end{aligned}$$

which according to Bayes' rule can be written as:

$$P(\mathbf{X} = \mathbf{x} \mid D) = \kappa \sum_S \prod_{i=0}^{N} \hat{\theta}_{iJK} \cdot P(D \mid S) \cdot P(S) \tag{C.5}$$

where κ is a constant depending only on the constant $P(D)$.

Given the assumptions of complete data, multinomial variables, Dirichlet priors and parameter independence, the marginal likelihood $P(D \mid S)$ can also be written just in terms of hyperparameters and sufficient statistics [Cooper and Herskovits, 1992; Heckerman *et al.*, 1995]:

$$P(D \mid S) = \prod_{i=0}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}. \tag{C.6}$$

The summation over all possible structures makes the full Bayesian model averaging approach intractable for any reasonable number of variables. Thus practical Bayesian algorithms resort to one of two solutions, either a MCMC model-averaging [Madigan and Raftery, 1994; Volinsky, 1997; Madigan and York, 1995] is performed by randomly searching the space of structures, or heuristic search is performed for a single model that maximizes Equation C.6 [Cooper and Herskovits, 1992; Heckerman *et al.*, 1995].

Both of these solutions require Equation C.6 to be calculated repeatedly during the course of a search procedure. Since the marginal likelihood of a given structure G is typically a very small number (on the order of $[\textit{total \# structures}]^{-1}$), it is usual to find the structure that maximizes the log-likelihood rather than the likelihood:

$$\begin{aligned}\log P(D \mid G) \quad = \quad & \sum_{i=1}^{n} \sum_{j=1}^{q_i} [\log \Gamma(\alpha_{ij}) - \log \Gamma(\alpha_{ij} + N_{ij})] \times \\ & \sum_{k=1}^{r_i} [\log \Gamma(\alpha_{ijk} + N_{ijk}) - \log \Gamma(\alpha_{ijk})],\end{aligned} \tag{C.7}$$

This criterion is important for several reasons. First, it allows the score of a given network to be calculated exactly in closed form. More importantly is that it is *separable*. That is, it takes the form:

$$\mathit{Score}(\mathbf{V}) = \sum_{i=1}^{N} c(V_i, \mathbf{P}_i), \tag{C.8}$$

where $c(V_i, \mathbf{P}_i)$ is a function that depends only on the sufficient statistics of V_i and $\mathbf{P}_i$. Thus, during a single-step search, after changing the parent set of a single node V_i, we can recalculate the entire score of the network by recalculating only the local statistics and the local function $c(V_i, \mathbf{P}_i)$. This allows the space of DAGs to be explored relatively quickly.

Bouckaert [1995] proves that, assuming a nonzero prior over structures, the existence of a P-map, and an infinite sample size, the posterior probability score will achieve a maximum for the P-map structure. This result justifies the use of this Bayesian search-based procedures for causal discovery.

APPENDIX D

THE CORRECTNESS OF THE CAUSAL ORDERING ALGORITHM

Comparison between the causal ordering given by COA and that given by an arbitrary expert is complicated by the fact that both procedures produce different types of directed graphs. I therefore must define precisely what I mean when I say that a DiG is consistent with a DPG.

Definition 48 (DiG/DPG consistency) *If* $\mathbf{V}$ *is a set of variables,* $G_p = \langle \mathcal{V}, \mathbf{A_p} \rangle$ *is a DPG over* $\mathbf{V}$*, and* $G = \langle \mathbf{V}, \mathbf{A} \rangle$ *is a DiG over* $\mathbf{V}$*, then* G *is consistent with* G_p *if and only if the following are true:*

1. *If an arc* $V_1 \rightarrow \mathbf{V^{(2)}}$ *exists in* $\mathbf{A_p}$ *then there exists a* $V_2 \in \mathbf{V^{(2)}}$ *such that the arc* $V_1 \rightarrow V_2$ *exists in* $\mathbf{A}$*.*

2. *If an edge* $V_1 \rightarrow V_2$ *exists in* $\mathbf{A}$*, then either* $V_1 \rightarrow \boldsymbol{Part}(V_2)$ *exists in* $\mathbf{A_p}$ *or* $\boldsymbol{Part}(V_1) = \boldsymbol{Part}(V_2)$*.*

Part 1 says that all arcs present in G_p must be represented in G, and Part 2 says that the only additional arcs that are allowed must be between variables that were "strongly coupled" in G_p. The DiG of Figure 1.2 is consistent with the DPG of Figure A.1.

Definition 49 (partial/total mapping consistency) *If* Φ_p *is a partial causal mapping over a self-contained set* $\mathbf{E}$ *with* $\mathbf{V} = \boldsymbol{Params}(\mathbf{E})$*, and* ϕ_t *is a total causal mapping over* $\mathbf{E}$*, then* ϕ_t *is consistent with* Φ_p *iff the following hold:*

1. *For each association* $\langle \mathbf{V^{(i)}}, \Phi_p(\mathbf{V^{(i)}}) \rangle \in \Phi_p$ *there exists for each* $V \in \mathbf{V^{(i)}}$ *an elementary association* $\langle V, E \rangle \in \phi_t$ *where* $E \in \Phi_p(\mathbf{V^{(i)}})$*.*

2. *An elementary association* $\langle V, E \rangle$ *exists in* ϕ_t *but not in* Φ_p *only if the non-elementary association* $\langle \boldsymbol{Part}(V), \boldsymbol{Part}(E) \rangle$ *exists in* Φ_p*.*

The following lemma shows that mapping consistency implies DiG/DPG consistency. I denote the DiG that corresponds to a total causal mapping ϕ_t as $DiG(\phi_t)$ and the DPG that corresponds to a partial causal mapping Φ_p as $DPG(\Phi_p)$:

Lemma 9 *Let* $\mathbf{E}$ *denote a self-contained set. If a total causal mapping* ϕ_t *over* $\mathbf{E}$ *is consistent with a partial causal mapping* Φ_p *over* $\mathbf{E}$*, then* $DiG(\phi_t)$ *is consistent with* $DPG(\Phi_p)$*.*

Proof: Let $G_p = \langle \mathcal{V}, \mathbf{A_p} \rangle$ denote $DPG(\Phi_p)$ and let $G_t = \langle \mathbf{V}, \mathbf{A} \rangle$ denote $DiG(\phi_t)$. Assume Conditions 49.1 and 49.2 are true.

Satisfaction of condition 48.1:

Assume an edge $V_1 \rightarrow \mathbf{V^{(2)}}$ exists in $\mathbf{A_p}$. Let $\langle \mathbf{V^{(2)}}, \Phi_p(\mathbf{V^{(2)}}) \rangle$ be the association corresponding to $\mathbf{V^{(2)}}$ in Φ_p. By condition 49.1 there exists in ϕ_t an elementary association of the form $\langle V_2, E_1 \rangle$ where $V_2 \in \mathbf{V^{(2)}}$ and $E_1 \in \Phi_p(\mathbf{V^{(2)}})$. Therefore in $DiG(\phi_t)$ there exists an edge from all $V_1^i \in \boldsymbol{Params}(E_1) \setminus \mathbf{V^{(2)}}$ to some $V_2 \in \mathbf{V^{(2)}}$. Finally, since $V_1 \rightarrow \mathbf{V^{(2)}}$ it must be the case that $V_1 \in \boldsymbol{Params}(E_1) \setminus \mathbf{V^{(2)}}$.

Satisfaction of condition 48.2:

Assume an edge $V_1 \rightarrow V_2$ exists in $\mathbf{A}$, then the elementary association $\langle V_2, E_1 \rangle$ must exist in ϕ_t such that $V_1 \in \boldsymbol{Params}(E_1)$. Then by condition 49.2, either $\langle \{V_2\}, \{E_1\} \rangle \in \Phi_p$ or $\langle \boldsymbol{Part}(V_2), \boldsymbol{Part}(E_1) \rangle \in \Phi_p$. Either way the association $\langle \boldsymbol{Part}(V_2), \boldsymbol{Part}(E_1) \rangle \in \Phi_p$. Therefore in $DPG(\Phi_p)$ an arc will be directed from all $V \in \boldsymbol{Params}(\boldsymbol{Part}(E_1)) \setminus \boldsymbol{Part}(V_2)$ to $\boldsymbol{Part}(V_2)$. Since $V_1 \in \boldsymbol{Params}(\boldsymbol{Part}(E_1))$, either there will exist an edge $V_1 \rightarrow \boldsymbol{Part}(V_2)$ or $V_1 \in \boldsymbol{Part}(V_2)$. □

Using this result, it can be shown that a DiG G_t generated by any total causal mapping ϕ_t over a set of equations $\mathbf{E}$ is consistent with the DPG G_p generated by applying COA to $\mathbf{E}$.

Theorem 12 *Given a self-contained set* $\mathbf{E}$*, any graph produced by constructing a total causal mapping on* $\mathbf{E}$ *is consistent with the graph specified by COA applied to* $\mathbf{E}$*.*

Proof: First I show that all total causal mappings must be consistent with the partial causal mapping generated by COA. Then the result follows from Lemma 9.

Satisfaction of condition 49.1:

I prove this result by induction. Let ϕ_t be an arbitrary total causal mapping on $\mathbf{E}$, and label the associations generated by COA as $\langle \boldsymbol{Params}(\mathbf{E_j^{(i)}}), \mathbf{\hat{E}_j^{(i)}}\rangle$ where $\mathbf{E_j^{(i)}}$ is the jth minimal self-contained subset found by COA in the ith level of recursion (e.g., the equations for the exogenous variables can be labeled as $E_1^{(0)}, E_2^{(0)}$, etc.). If V is an arbitrary variable such that $V \in \boldsymbol{Params}(\mathbf{E_j^{(i)}})$, I must show that V gets mapped to some equation $E \in \mathbf{\hat{E}_j^{(i)}}$. Let $\langle \boldsymbol{Params}(\mathbf{E_l^{(k)}}), \mathbf{\hat{E}_l^{(k)}}\rangle$ be an arbitrary association made by COA. Assume that condition 49.1 holds for all associations $\langle \boldsymbol{Params}(\mathbf{E_j^{(i)}}), \mathbf{\hat{E}_j^{(i)}}\rangle$ with all $i < k$. I show that it must also hold for the association $\langle \boldsymbol{Params}(\mathbf{E_l^{(k)}}), \mathbf{\hat{E}_l^{(k)}}\rangle$. Let $\langle V, E\rangle \in \phi_t$ be an arbitrary association such that $E \in \hat{E}_l^{(k)}$. By definition of a causal mapping, $V \in \boldsymbol{Params}(E)$, and therefore it must be the case that $V \in \boldsymbol{Params}(\hat{E}_l^{(k)})$. However, according to the induction hypothesis, all $V \in \boldsymbol{Params}(\hat{E}_l^{(k)}) \setminus \boldsymbol{Params}(\mathbf{E_l^{(k)}})$ have already been assigned to equations; therefore it must be the case that $V \in \boldsymbol{Params}(\mathbf{E_l^{(k)}})$. To complete the induction step, notice that for any association in the initial level of recursion $\langle \boldsymbol{Params}(\mathbf{E_l^{(0)}}), \mathbf{\hat{E}_l^{(0)}}\rangle$, it must be the case that $\boldsymbol{Params}(\mathbf{E_l^{(0)}}) \equiv \boldsymbol{Params}(\mathbf{\hat{E}_l^{(0)}})$ so for any $\langle V, E\rangle \in \phi_t$ it must be the case that $V \in \boldsymbol{Params}(\mathbf{E_l^{(0)}})$

Satisfaction of condition 49.2:

Let $\langle V, E\rangle$ be an elementary association in ϕ_t. Consider the association $\langle \boldsymbol{Part}(V), \Phi_p(\boldsymbol{Part}(V))\rangle \in \Phi_p$. By condition 49.1 there exists an elementary association

$\langle V, E' \rangle \in \Phi_p$ such that $E' \in \Phi_p(\boldsymbol{Part}(V))$. But since ϕ_t is one-to-one there can be only one equation associated with V. Therefore $E' = E$ (and $\boldsymbol{Part}(E') = \boldsymbol{Part}(E)$). □

BIBLIOGRAPHY

R. Bentzel and B. Hansen. On recursiveness and interdependency in economic models. *Review of Economic Studies*, 22:153–168, 1954.

R. Bouckaert. *Bayesian belief networks: From construction to inference.* PhD thesis, University Utrecht, 1995.

Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence*, 137(1):43–90, May 2002.

David Maxwell Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 87–98, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

Denver H. Dash and Marek J. Druzdzel. A hybrid anytime algorithm for the construction of causal models from sparse data. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 142–149, San Francisco, CA, 1999. Morgan Kaufmann Publishers, Inc.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B(39):1–38, 1977.

Marek J. Druzdzel and Hans van Leijen. Causal reversibility in Bayesian networks. *Journal of Experimental and Theoretical Artificial Intelligence*, 13(1):45–62, Jan 2001.

Franklin M. Fisher. A correspondence principle for simultaneous equation models. *Econometrica*, 38(1):73–92, January 1970.

D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1–2):9–43, 1997.

Moises Goldszmidt and Judea Pearl. Ranked-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, pages 661–672, San Mateo, CA, 1992. Morgan Kaufmann.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, January 1943.

Joseph Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

David Heckerman. A tutorial on learning with Bayesian networks. In Michael I. Jordan, editor, *Learning in Graphical Models*. The MIT Press, Cambridge, Massachusetts, 1998.

Yumi Iwasaki and Herbert A. Simon. Causality and model abstraction. *Artificial Intelligence*, 67(1):143–194, May 1994.

Benjamin Kuipers. Abstraction by time-scale in qualitative simulation. In *Proceedings of the National Conference on Artificial Intelligence, AAAI–87*, pages 621–625, Seattle, WA, July 1987. American Association for Artificial Intelligence, Morgan Kaufmann Publishers, Inc., San Mateo, CA.

David Madigan and Adrian E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

David Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 411–418, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

P. Pandurang Nayak. Causal approximations. *Artificial Intelligence*, 70(1–2):1–58, 1994.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452, Cambridge, MA, 1991. Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

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

Thomas Richardson. *Models of Feedback: Interpretation and Discovery.* PhD dissertation, Carnegie Mellon University, Department of Philosophy, 1996.

Paul A. Samuelson. *Foundations of Economic Analysis.* Harvard University Press, Cambridge, MA, 1947.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. Truth is among the best explanations: Finding causal explanations of conditional independence and dependence. In Clark Glymour and Gregory F. Cooper, editors, *Computation, Causation, and Discovery*, pages 167–209. AAAI Press, Menlo Park, 1999.

D. Serrano and D. Gossard. Constraint management in conceptual design. In D. Sriram and R.A. Adey, editors, *Knowledge Based Expert Systems in Engineering: Planning and Design*, pages 211–224. Computational Mechanics Publications, Southampton, 1987.

Herbert A. Simon and Nicholas Rescher. Cause and counterfactual. *Philosophy of Science*, 33(4):323–340, December 1966.

Herbert A. Simon. Causal ordering and identifiability. In William C. Hood and Tjalling C. Koopmans, editors, *Studies in Econometric Method. Cowles Commission for Research in Economics. Monograph No. 14*, chapter III, pages 49–74. John Wiley & Sons, Inc., New York, NY, 1953.

Herbert Simon. Causality and econometrics: Comment. *Econometrica*, 23(2):193–195, April 1955.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Springer Verlag, New York, 1993.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* The MIT Press, Cambridge, MA, second edition, 2000.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 491–498, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos with Applications to Physics, Biology, Chemistry, and Engineering.* Addison-Wesley, Publishers, Reading, MA, 1991.

Robert H. Strotz and H.O.A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis; Part I of a triptych on causal chain systems. *Econometrica*, 28(2):417–427, April 1960.

T.S. Verma and Judea Pearl. Equivalence and synthesis of causal models. In P.P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence 6*, pages 255 –269. Elsevier Science Publishing Company, Inc., New York, N. Y., 1991.

T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth Annual Conference on Uncertainty in Artificial Intelligence (UAI–92)*, pages 323–330, San Francisco, CA, 1992. Morgan Kaufmann Publishers.

C.T. Volinsky. *Bayesian Model Averaging for Censored Survival Models*. PhD dissertation, University of Washington, 1997.

Herman Wold. Causality and econometrics. *Econometrica*, 22(2):162–177, April 1954.

Herman Wold. Causality and econometrics: Reply. *Econometrica*, 23(2):196–197, April 1955.

Sewall Wright. The method of path coefficients. *Annals of Mathematical Statistics*, 5:161–215, 1934.

INDEX