

Lecture Notes in Statistics

81

Peter Spirtes, Clark Glymour,
Richard Scheines

Causation, Prediction,
and Search



Springer-Verlag

Lecture Notes in Statistics

81

Edited by J. Berger, S. Fienberg, J. Gani, K. Krickeberg,
I. Olkin, and B. Singer



Peter Spirtes
Clark Glymour
Richard Scheines

Causation, Prediction, and Search

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Peter Spirtes
Clark Glymour
Richard Scheines
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Mathematics Subject Classification: 62-07

Library of Congress Cataloging-in-Publication Data

Spirtes, Peter

Causation, prediction, and search / Peter Spirtes, Clark Glymour

Richard Scheines

p. cm. — (Lecture notes in statistics ; 81)

ISBN-13: 978-1-4612-7650-0

1. Mathematical statistics I. Glymour, Clark N. II. Scheines,
Richard. III. Title. IV. Series: Lecture notes in statistics

(Springer-Verlag) : v. 81.

QA276.S65 1993

519.5—dc20

92-40263

Printed on acid-free paper.

© 1993 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Camera ready copy provided by the authors.

9 8 7 6 5 4 3 2 1

ISBN-13: 978-1-4612-7650-0 e-ISBN-13: 978-1-4612-2748-9

DOI: 10.1007/978-1-4612-2748-9

To my parents, Morris and Cecile Spirtes - P.S.

In memory of Lucille Lynch Schwartz Watkins Speede Tindall Preston - C. G.

To Martha, for her support and love - R.S.

It is with data affected by numerous causes that Statistics is mainly concerned. Experiment seeks to disentangle a complex of causes by removing all but one of them, or rather by concentrating on the study of one and reducing the others, as far as circumstances permit, to comparatively small residuum. Statistics, denied this resource, must accept for analysis data subject to the influence of a host of causes, and must try to discover from the data themselves which causes are the important ones and how much of the observed effect is due to the operation of each.

--G. U. Yule and M. G. Kendall 1950

The Theory of Estimation discusses the principles upon which observational data may be used to estimate, or to throw light upon the values of theoretical quantities, not known numerically, which enter into our specification of the causal system operating.

-- Sir Ronald Fisher, 1956

George Box has [almost] said "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively." These words of caution about "natural experiments" are uncomfortably strong. Yet in today's world we see no alternative to accepting them as, if anything, too weak.

--G. Mosteller and J. Tukey, 1977

Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics.

-- P. Dawid, 1979

Preface

This book is intended for anyone, regardless of discipline, who is interested in the use of statistical methods to help obtain scientific explanations or to predict the outcomes of actions, experiments or policies.

Much of G. Udny Yule's work illustrates a vision of statistics whose goal is to investigate when and how causal influences may be reliably inferred, and their comparative strengths estimated, from statistical samples. Yule's enterprise has been largely replaced by Ronald Fisher's conception, in which there is a fundamental cleavage between experimental and non-experimental inquiry, and statistics is largely unable to aid in causal inference without randomized experimental trials. Every now and then members of the statistical community express misgivings about this turn of events, and, in our view, rightly so. Our work represents a return to something like Yule's conception of the enterprise of theoretical statistics and its potential practical benefits.

If intellectual history in the 20th century had gone otherwise, there might have been a discipline to which our work belongs. As it happens, there is not. We develop material that belongs to statistics, to computer science, and to philosophy; the combination may not be entirely satisfactory for specialists in any of these subjects. We hope it is nonetheless satisfactory for its purpose. We are not statisticians by training or by association, and perhaps for that reason we tend to look at issues differently, and, from the perspective common in the discipline, no doubt oddly. We are struck by the fact that in the social and behavioral sciences, epidemiology, economics, market research, engineering, and even applied physics, statistical methods are routinely used to justify causal inferences from data not obtained from randomized experiments, and sample statistics are used to predict the effects of policies, manipulations or experiments. Without these uses the profession of statistics would be a far smaller business. It may not strike many professional statisticians as particularly odd that the discipline thriving from such uses assures its audience that they are unwarranted, but it strikes us as very odd indeed. From our perspective outside the discipline, the most urgent questions about the application of statistics to such ends concern the conditions under which causal inferences and predictions of the effects of manipulations can and cannot reliably be made, and the most urgent need is a principled, rigorous theory with which to address these

problems. To judge from the testimony of their books, a good many statisticians think any such theory is impossible. We think the common arguments against the possibility of inferring causes from statistics outside of experimental trials are unsound, and radical separations of the principles of experimental and observational study designs are unwise. Experimental and observational design may not always permit the same inferences, but they are subject to uniform principles.

The theory we develop follows necessarily from assumptions laid down in the statistical community over the last fifteen years. The underlying structure of the theory is essentially axiomatic. We will give two independent axioms on the relation between causal structures and probability distributions and deduce from them features of causal relationships and predictions that can and that cannot be reliably inferred from statistical constraints under a variety of background assumptions. Versions of all of the axioms can be found in papers by Lauritzen, Wermuth, Speed, Pearl, Rubin, Pratt, Schlaifer, and others. In most cases we will develop the theory in terms of probability distributions that can be thought of loosely as propensities that determine long run frequencies, but many of the probability distributions can alternatively be understood as (normative) subjective degrees of belief, and we will occasionally note Bayesian applications. From the axioms there follow a variety of theorems concerning estimation, sampling, latent variable existence and structure, regression, indistinguishability relations, experimental design, prediction, Simpson's paradox, and other topics. Foremost among the "other topics" are the discovery that statistical methods commonly used for causal inference are radically suboptimal, and that there exist asymptotically reliable, computationally efficient search procedures that conjecture causal relationships from the outcomes of statistical decisions made on the basis of sample data. (The procedures we will describe require statistical decisions about the independence of random variables; when we say such a procedure is "asymptotically reliable" we mean it provides correct information if the outcome of each of the requisite statistical decisions is true in the population under study.)

This much of the book is mathematics: where the axioms are accepted, so must the theorems be, including the existence of search procedures. The procedures we describe are applicable to both linear and discrete data and can be feasibly applied to a hundred or more variables so long as the causal relations between the variables are sufficiently sparse and the sample

sufficiently large. These procedures have been implemented in a computer program, TETRAD II, which at the time of writing is publicly available.¹

The theorems concerning the existence and properties of reliable discovery procedures of themselves tell us nothing about the reliabilities of the search procedures in the short run. The methods we describe require an unpredictable sequence of statistical decisions, which we have implemented as hypothesis tests. As is usual in such cases, in small samples the conventional p values of the individual tests may not provide good estimates of type 1 error probabilities for the search methods. We provide the results of extensive tests of various procedures on simulated data using Monte Carlo methods, and these tests give considerable evidence about reliability under the conditions of the simulations. The simulations illustrate an easy method for estimating the probabilities of error for any of the search methods we describe. The book also contains studies of one large pseudo-empirical data set--a body of simulated data created by medical researchers to model emergency medicine diagnostic indicators and their causes--and a great many empirical data sets, most of which have been discussed by other authors in the context of specification searches.

A further aim of this work is to show that a proper understanding of the relationship between causality and probability can help to clarify diverse topics in the statistical literature, including the comparative power of experimentation versus observation, Simpson's paradox, errors in regression models, retrospective versus prospective sampling, the perils of variable selection, and other topics. There are a number of relevant topics we do not consider. They include problems of estimation with discrete latent variables, optimizing statistical decisions, many details of sampling designs, time series, and a full theory of "non-recursive" causal structures--i.e., finite graphical representations of systems with feedback.

Causation, Prediction and Search is not intended to be a textbook, and it is not fitted out with the associated paraphernalia. There are open problems but no exercises. In a textbook everything ought to be presented as if it were complete and tidy, even if it isn't. We make no such pretenses in this book, and the chapters are rich in unsolved problems and open questions. Textbooks don't usually pause much to argue points of view; we pause quite a lot.

The various theorems in this book often have a graph theoretic character; many of them are long, difficult case arguments of a kind quite unfamiliar in statistics. In order not to interrupt

¹To anyone with a workstation with a UNIX operating system, a PASCAL compiler and a network connection. A less flexible version of the program is available for IBM compatible 80-386 and 80-486 personal computers. Write Richard Scheines at RS2L@andrew.cmu.edu.

the flow of the discussion we have placed all proofs but one in a chapter at the end of the book. In the few cases where detailed proofs are available in the published literature, we have simply referred the reader to them. Where proofs of important results have not been published or are not readily available we have given the demonstrations in some detail.

The structure of the book is as follows. Chapter 1 concerns the motivation for the book in the context of current statistical practice and advertises some of the results. Chapter 2 introduces the mathematical ideas necessary to the investigation, and Chapter 3 gives the formal framework a causal interpretation, lays down the axioms, notes circumstances in which they are likely to fail, and provides a few fundamental theorems. The next two chapters work out the consequences of two of the axioms for some fundamental issues in contexts in which it is known, or assumed, that there are no unmeasured common causes affecting measured variables. In Chapter 4 we give graphical characterizations of necessary and sufficient conditions for causal hypotheses to be statistically indistinguishable from one another in each of several senses. In Chapter 5 we criticize features of model specification procedures commonly recommended in statistics, and we describe feasible algorithms that from properties of population distributions extract correct information about causal structure, assuming the axioms apply and that no unmeasured common causes are at work. The algorithms are illustrated for a variety of empirical and simulated samples. Chapter 6 extends the analysis of Chapter 5 to contexts in which it cannot be assumed that no unmeasured common causes act on measured variables. From both a theoretical and practical perspective, this chapter and the next form the center of the book, but they are especially difficult. Chapter 7 addresses the fundamental issue of predicting the effects of manipulations, policies, or experiments. As an easy corollary, the chapter unifies directed graphical models with Donald Rubin's "counterfactual" framework for analyzing prediction. Chapter 8 applies the results of the preceding chapters to the subject of regression. We argue that even when standard statistical assumptions are satisfied multiple regression is a defective and unreliable way to assess causal influence even in the large sample limit, and various automated regression model specification searches only make matters worse. We show that the algorithms of Chapter 6 are more reliable in principle, and we compare the performances of these algorithms against various multiple regression procedures on a variety of simulated and empirical data sets. Chapter 9 considers the design of empirical studies in the light of the results of earlier chapters, including issues of retrospective and prospective sampling, the comparative power of experimental and observational designs, selection of variables, and the design of ethical clinical trials. The chapter concludes with a look back at some aspects of the dispute over smoking and lung cancer. Chapters 10 and 11 further consider the linear case, and analyze algorithms for discovering or elaborating causal relations among measured and

unmeasured variables in linear systems. Chapter 12 is a brief consideration of a variety of open questions. Proofs are given in Chapter 13.

We have tried to make this work self-contained, but it is admittedly and unavoidably difficult. The reader will be aided by a previous reading of Pearl (1988), Whittaker (1990) or Neopolitan (1990).

Acknowledgments

One source of the ideas in this book is in work we began ten years ago at the University of Pittsburgh. We drew many ideas about causality, statistics and search from the psychometric, economic and sociological literature, beginning with Charles Spearman's project at the turn of the century and including the work of Herbert Simon, Hubert Blalock and Herbert Costner.

We obtained a new perspective on the enterprise from Judea Pearl's *Probabilistic Reasoning in Intelligent Systems*, which appeared the next year. Although not principally concerned with discovery, Pearl's book showed us how to connect conditional independence with causal structure quite generally, and that connection proved essential to establishing general, reliable discovery procedures. We have since profited from correspondence and conversation with Pearl and with Dan Geiger and Thomas Verma, and from several of their papers. Pearl's work drew on the papers of Wermuth (1980), Kiiveri and Speed (1982), Wermuth and Lauritzen (1983), and Kiiveri, Speed and Carlin (1984), which in the early 1980s had already provided the foundations for a rigorous study of causal inference. Paul Holland introduced one of us to the Rubin framework some years ago, but we only recently realized its logical connections with directed graphical models. We were further helped by J. Whittaker's (1990) excellent account of the properties of undirected graphical models.

We have learned a great deal from Gregory Cooper at the University of Pittsburgh who provided us with data, comments, Bayesian algorithms and the picture and description of the ALARM network which we consider in several places. Over the years we have learned useful things from Kenneth Bollen. Chris Meek provided essential help in obtaining an important theorem that derives various claims made by Rubin, Pratt and Schlaifer from axioms on directed graphical models.

Steve Fienberg and several students from Carnegie Mellon's department of statistics joined with us in a seminar on graphical models from which we learned a great deal. We are indebted to him for his openness, intelligence and helpfulness in our research, and to Elizabeth Slate for guiding us through several papers in the Rubin framework. We are obliged to Nancy Cartwright for her courteous but salient criticism of the approach taken in our previous book and continued here. Her comments prompted our work on parameters in

Chapter 4. We are indebted to Brian Skyrms for his interest and encouragement over many years, and to Marek Druzdzel for helpful comments and encouragement. We have also been helped by Linda Bouck, Ronald Christensen, Jan Callahan, David Papineau, John Earman, Dan Hausman, Joe Hill, Michael Meyer, Teddy Seidenfeld, Dana Scott, Jay Kadane, Steven Klepper, Herb Simon, Peter Slezak, Steve Sorensen, John Worrall and Andrea Woody. We are indebted to Ernest Seneca for putting us in contact with Dr. Rick Linthurst, and we are especially grateful to Dr. Linthurst for making his doctoral thesis available to us.

Our work has been supported by many institutions. They, and those who made decisions on their behalf, deserve our thanks. They include Carnegie Mellon University, the National Science Foundation programs in History and Philosophy of Science, in Economics, and in Knowledge and Database Systems, the Office of Naval Research, the Navy Personnel Research and Development Center, the John Simon Guggenheim Memorial Foundation, Susan Chipman, Stanley Collyer, Helen Gigley, Peter Machamer, Steve Sorensen, Teddy Seidenfeld and Ron Overmann. The Navy Personnel Research and Development Center provided us the benefit of access to a number of challenging data analysis problems from which we have learned a great deal.

Table of Contents

Preface	vii
Acknowledgments	xiii
Notational Conventions	xxi
1. Introduction and Advertisement	1
1.1 The Issue	1
1.2 Advertisements	10
1.2.1 Bayes Networks from the Data	11
1.2.2 Structural Equation Models from the Data	13
1.2.3 Selection of Regressors.....	14
1.2.4 Causal Inference without Experiment	17
1.2.5 The Structure of the Unobserved	19
1.3 Themes.....	21
2. Formal Preliminaries.....	25
2.1 Graphs.....	25
2.2 Probability.....	31
2.3 Graphs and Probability Distributions	32
2.3.1 Directed Acyclic Graphs.....	32
2.3.2 Directed Independence Graphs	34
2.3.3 Faithfulness	35
2.3.4 d-separation.....	36
2.3.5 Linear Structures.....	36
2.4 Undirected Independence Graphs	37
2.5 Deterministic and Pseudo-Indeterministic Systems	38
2.6 Background Notes	39
3. Causation and Prediction: Axioms and Explications	41
3.1 Conditionals	41
3.2 Causation	42
3.2.1 Direct vs. Indirect Causation	42
3.2.2 Events and Variables	43
3.2.3 Examples.....	45
3.2.4 Representing Causal Relations with Directed Graphs.....	47

3.3 Causality and Probability	49
3.3.1 Deterministic Causal Structures	49
3.3.2 Pseudo-Indeterministic and Indeterministic Causal Structures	51
3.4 The Axioms	53
3.4.1 The Causal Markov Condition.....	53
3.4.2 The Causal Minimality Condition	55
3.4.3 The Faithfulness Condition.....	56
3.5 Discussion of the Conditions	57
3.5.1 The Causal Markov and Minimality Conditions	57
3.5.2 Faithfulness and Simpson's Paradox	64
3.6 Bayesian Interpretations	70
3.7 Consequences of The Axioms	71
3.7.1 d-Separation	71
3.7.2 The Manipulation Theorem	75
3.8 Determinism	81
3.9 Background Notes	86
4. Statistical Indistinguishability	87
4.1 Strong Statistical Indistinguishability	88
4.2 Faithful Indistinguishability.....	89
4.3 Weak Statistical Indistinguishability	90
4.4 Rigid Indistinguishability	93
4.5 The Linear Case	94
4.6 Redefining Variables	99
4.7 Background Notes	101
5. Discovery Algorithms for Causally Sufficient Structures.....	103
5.1 Discovery Problems	103
5.2 Search Strategies in Statistics	104
5.2.1 The Wrong Hypothesis Space	105
5.2.2 Computational and Statistical Limitations.....	107
5.2.3 Generating a Single Hypothesis.....	108
5.2.4 Other Approaches	109
5.2.5 Bayesian Methods.....	109
5.3 The Wermuth-Lauritzen Algorithm.....	111
5.4 New Algorithms	112
5.4.1 The SGS Algorithm	114
5.4.2 The PC Algorithm.....	116

Contents	xvii
5.4.3 The IG (Independence Graph) Algorithm	124
5.4.4 Variable Selection.....	125
5.4.5 Incorporating Background Knowledge.....	127
5.5 Statistical Decisions.....	128
5.6 Reliability and Probabilities of Error.....	130
5.7 Estimation	132
5.8 Examples and Applications	132
5.8.1 The Causes of Publishing Productivity.....	133
5.8.2 Education and Fertility	139
5.8.3 The Female Orgasm.....	140
5.8.4 The American Occupational Structure	142
5.8.5 The ALARM Network.....	145
5.8.6 Virginity.....	147
5.8.7 The Leading Crowd	147
5.8.8 Influences on College Plans.....	149
5.8.9 Abortion Opinions	150
5.8.10 Simulation Tests with Random Graphs	152
5.9 Conclusion	161
5.10 Background Notes	162
6. Discovery Algorithms without Causal Sufficiency	163
6.1 Introduction	163
6.2 The PC Algorithm and Latent Variables	165
6.3 Mistakes	168
6.4 Inducing Paths	173
6.5 Inducing Path Graphs	174
6.6 Partially Oriented Inducing Path Graphs	177
6.7 Algorithms for Causal Inference with Latent Common Causes	181
6.8 Theorems on Detectable Causal Influence	190
6.9 Non-Independence Constraints.....	191
6.10 Generalized Statistical Indistinguishability and Linearity	193
6.11 The Tetrad Representation Theorem	196
6.12 An Example: Math Marks and Causal Interpretation	197
6.13 Background Notes	200
7. Prediction.....	201
7.1 Introduction.....	201
7.2 Prediction Problems	202

7.3 Rubin-Holland-Pratt-Schlaifer Theory	203
7.4 Prediction with Causal Sufficiency	213
7.5 Prediction without Causal Sufficiency	216
7.6 Examples.....	227
7.7 Conclusion	237
7.8 Background Notes	237
8. Regression, Causation and Prediction	238
8.1 When Regression Fails to Measure Influence	238
8.2 A Solution and Its Application	242
8.2.1 Components of the Armed Forces Qualification Test	243
8.2.2 The Causes of Spartina Biomass	244
8.2.3 The Effects of Foreign Investment on Political Repression	248
8.2.4 More Simulation Studies	250
8.3 Error Probabilities for Specification Searches.....	252
8.4 Conclusion	257
9. The Design of Empirical Studies	259
9.1 Observational or Experimental Study?.....	259
9.2 Selecting Variables	271
9.3 Sampling	272
9.4 Ethical Issues in Experimental Design	276
9.4.1 The Kadane/Sedransk/Seidenfeld Design.....	277
9.4.2 Causal Reasoning in the Experimental Design.....	280
9.4.3 Towards Ethical Trials.....	286
9.5 An Example: Smoking and Lung Cancer	291
9.6 Appendix.....	302
10. The Structure of the Unobserved	306
10.1 Introduction.....	306
10.2 An Outline of the Algorithm.....	307
10.3 Finding Almost Pure Measurement Models	310
10.3.1 Intra-Construct Foursomes	310
10.3.2 Cross-Construct Foursomes	311
10.4 Facts about the Unobserved Determined by the Observed.....	315
10.5 Unifying the Pieces.....	316
10.6 Simulation Tests	320
10.7 Conclusion	322

Contents	xix
11. Elaborating Linear Theories with Unmeasured Variables.....	323
11.1 Introduction.....	323
11.2 The Procedure	324
11.2.1 Scoring	324
11.2.2 Search	327
11.3 The LISREL and EQS Procedures	329
11.3.1 Input and Output	329
11.3.2 Scoring	330
11.3.3 The LISREL VI Search	331
11.3.4 The EQS Search.....	331
11.4 The Primary Study	332
11.4.1 The Design of Comparative Simulation Studies	332
11.4.2 Study Design.....	333
11.5 Results.....	343
11.6 Reliability and Informativeness	346
11.7 Using LISREL and EQS as Adjuncts to Search	349
11.8 Limitations of the TETRAD II Elaboration Search.....	351
11.9 Some Morals for Statistical Search.....	352
12. Open Problems	354
12.1 Feedback, Reciprocal Causation, and Cyclic Graphs.....	354
12.1.1 Mason's Theorem.....	355
12.1.2 Time Series and Cyclic Graphs	356
12.1.3 The Markov Condition, Factorizability and Faithfulness.....	359
12.1.4 Discovery Procedures	360
12.2 Indistinguishability Relations	361
12.3 Time series and Granger Causality	363
12.4 Model Specification and Parameter Estimation from the Same Data Base....	365
12.5 Conditional Independence Tests	366
13. Proofs of Theorems	367
13.1 Theorem 2.1	367
13.2 Theorem 3.1	367
13.3 Theorem 3.2	374
13.4 Theorem 3.3	376
13.5 Theorem 3.4	385
13.6 Theorem 3.5	386
13.7 Theorem 3.6 (Manipulation Theorem)	395

13.8 Theorem 3.7	398
13.9 Theorem 4.1	401
13.10 Theorem 4.2	403
13.11 Theorem 4.3	403
13.12 Theorem 4.4	404
13.13 Theorem 4.5	404
13.14 Theorem 4.6	405
13.15 Theorem 5.1	405
13.16 Theorem 6.1	408
13.17 Theorem 6.2	411
13.18 Theorem 6.3	414
13.19 Theorem 6.4	417
13.20 Theorem 6.5	418
13.21 Theorem 6.6	419
13.22 Theorem 6.7	424
13.23 Theorem 6.8	425
13.24 Theorem 6.9	425
13.25 Theorem 6.10 (Tetrad Representation Theorem)	426
13.26 Theorem 6.11	460
13.27 Theorem 7.1	460
13.28 Theorem 7.2	462
13.29 Theorem 7.3	463
13.30 Theorem 7.4	470
13.31 Theorem 7.5	471
13.32 Theorem 9.1	472
13.33 Theorem 9.2	472
13.34 Theorem 10.1	473
13.35 Theorem 10.2	476
13.36 Theorem 11.1	479
Glossary	481
Bibliography	495
Index.....	517

Notational Conventions

Text

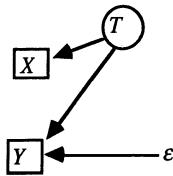
In the text, each technical term is written in boldface where it is defined.

Variables:	capitalized, and in italics, e.g., X
Values of variables:	lower case, and in italics, e.g., x
Sets:	capitalized, and in boldface, e.g., V
Values of sets of variables:	lower case, and in boldface, e.g., $V = v$
Members of X that are not members of Y :	$X \setminus Y$
Error variables:	ε, δ, e
Independence of X and Y :	$X \perp\!\!\!\perp Y$
Independence of X and Y conditional on Z :	$X \perp\!\!\!\perp Y Z$
$X \cup Y$:	XY
Covariance of X and Y :	$\text{COV}(X, Y)$ or γ_{XY}
Correlation of X and Y :	ρ_{XY}
Sample correlation of X and Y :	r_{XY}
Partial Correlation of X and Y , controlling for all members of set Z :	$\rho_{XY.Z}$

In all of the graphs that we consider, the vertices are random variables. Hence we use the terms "variables in a graph" and "vertices in a graph" interchangeably.

Figures

Figure numbers occur centered just below a figure, starting at 1 within each chapter. Where necessary, we distinguish between measured and unmeasured variables by boxing measured variables and circling unmeasured variables (except for error terms). Variables beginning with e , ε , or δ are understood to be "error," or "disturbance" variables. For example, in the figure below, X and Y are measured, T is not, and ε is an error term.

**Figure 1**

We will neither box nor circle variables in graphs in which no distinction need be made between measured and unmeasured variables, e.g., figure 2.

**Figure 2**

For simplicity, we state and prove our results for probability distributions over discrete random variables. However, under suitable integrability conditions, the results can be easily generalized to continuous distributions that have density functions by replacing the discrete variables by continuous variables, probability distributions by density functions, and summations by integrals.

If a description of a set of variables is a function of a graph G and variables in G , then we make G an optional argument to the function. For example, **Parents**(G, X) denotes the set of variables that are parents of X in graph G ; if the context makes clear which graph is being referred to we will simply write **Parents**(X).

If a distribution is defined over a set of random variables O then we refer to the distribution as $P(O)$. An equation between distributions over random variables is understood to be true for all values of the random variables for which all of the distributions in the equation are defined. For example if X and Y each take the values 0 or 1 and $P(X = 0) \neq 0$ and $P(X = 1) \neq 0$ then $P(Y|X) = P(Y)$ means $P(Y = 0|X = 0) = P(Y = 0)$, $P(Y = 0|X = 1) = P(Y = 0)$, $P(Y = 1|X = 0) = P(Y = 1)$, and $P(Y = 1|X = 1) = P(Y = 1)$.

We sometimes use a special summation symbol, $\sum_{\vec{X}}$, which has the following properties:

- (i) when sets of random variables are written beneath the special summation symbol, it is understood that the summation is to be taken over sets of values of the random variables, not the random variables themselves,
- (ii) if a conditional probability distribution appears in the scope of such a summation symbol, the summation is to be taken only over values of the random variables for which the conditional probability distributions are defined,
- (iii) if there are no values of the random variables under the special summation symbol for which the conditional probability distributions in the scope of the symbol are defined, then the summation is equal to zero.

For example, suppose that X , Y , and Z can each take on the values 0 or 1. Then if $P(Y=0;Z=0) \neq 0$

$$\sum_{\vec{X}} P(X|Y=0, Z=0) = P(X=0|Y=0, Z=0) + P(X=1|Y=0, Z=0)$$

However, if $P(Y=0;Z=0) = 0$, then $P(X=0|Y=0, Z=0)$ and $P(X=1|Y=0, Z=0)$ are not defined, so

$$\sum_{\vec{X}} P(X|Y=0, Z=0) = 0$$

We will adopt the following conventions for empty sets of variables. If $\mathbf{Y} = \emptyset$ then

- (i) $P(\mathbf{X}|\mathbf{Y})$ means $P(\mathbf{X})$.
- (ii) $\rho_{\mathbf{X}\mathbf{Z},\mathbf{Y}}$ means $\rho_{\mathbf{X}\mathbf{Z}}$.
- (iii) $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{Y}$ means $\mathbf{A} \perp\!\!\!\perp \mathbf{B}$.
- (iv) $\mathbf{A} \perp\!\!\!\perp \mathbf{Y}$ is always true.

Chapter 1

Introduction and Advertisement

1.1 The Issue

Statistics textbooks provide interesting examples of causal questions: Did halothane do more to cause surgical deaths than ether? Was the lower admission rate of women to graduate programs at the University of California caused by discrimination against women? Does smoking cause cancer? Issues about determining causes surround many of the introductory and even advanced topics in statistical pedagogy: experimental design, randomization, collinearity in multiple regression, observational versus experimental studies, and so forth. But except for the standard warnings that correlation is not causation, the textbooks include little if any systematic discussion of the connection between causation and probability. The mathematics of probability and statistical inference is explicit, but the connection between probability relations and causal dependencies is almost completely tacit. The same applies to prediction, at least outside of econometrics. The textbooks consider cases where policy interventions are at issue, but they tell us nothing systematic about the connections between statistical analysis of observations or experiments and predictions of the effects of policies, actions or manipulations.

Even more curious to an outsider, many statistics textbooks claim that the methods they describe *cannot* be reliably used to infer causal dependencies from random uncontrolled samples, or to predict the effects of manipulations, and they say or suggest that all possible statistical methods are equally impotent for these purposes. If widely believed, these bold claims would remove much of the market for the books that advance them, and at least some of the interest in statistics as a subject. Linear regression, for example, is taught as a means of fitting a line to sample data and also as a means of predicting new values of a variable. Yet many of the real applications of regression are to predict values of a variable when the regressors are *manipulated*, that is, when action or policy forces some novel distribution on the regressors. Every author of a regression textbook must know that were it not for such uses

much of the audience for the book would be lost. Modestly, the texts announce that regression cannot be used to infer causes from non-experimental samples, or to make predictions about manipulated systems; immodestly, often the same books announce that no *possible* statistical methods can be used to reliably infer causes or to make predictions about manipulated systems.

The arguments advanced against the possibility of reliable inference from statistical samples to causal structure are remarkably uniform from book to book. Mosteller and Tukey (1977), for example, provide an abundance of cases in which *regression* yields erroneous causal inferences. Rawlings (1988) makes the chief example of his textbook on applied regression a problem of causal inference pursued through several chapters, only to conclude that regression is unavailing for the purpose. But without some demonstration that regression is optimally reliable for such inference problems, these considerations only argue that rather than teaching regression we should look for better methods. An even more common textbook argument attempts to demonstrate that what regression cannot do for causal inference, nothing can. The argument asserts the underdetermination of causal structure by statistical dependencies. Readers are told (Younger, 1978; Rawlings, 1988) that a "relationship"--that is, some statistical dependency--between sample values of variables X and Y may be observed

- (i) when X causes Y ,
- (ii) when Y causes X ,
- (iii) when each causes the other,
- (iv) when some third variable causes both,
- (v) when the sample is not representative, or
- (vi) when the values of X and Y form time series.

Some such list is inevitably followed by the warning that "experimentation, and not the existence of the statistical relationship, is necessary to establish a cause" (Younger, p. 176). Consider the logical force of this argument:

It is true enough that we cannot distinguish among (i), (ii) (iii) and (iv) when we measure only two variables, X and Y , but what is the proof that we cannot distinguish among these alternative causal relations if additional variables are measured? We know that in other contexts identifiability and estimation properties of parameters relating a pair of variables can be changed if further variables are measured; the method of instrumental variables in econometrics (Bowden and Turkington, 1984) is a famous illustration. May the same not be the case with identifying causal structure? Where have the scores of textbooks so confident of the negative answer hidden their proofs? Or consider (v). The appeal to unrepresentative samples is as red a herring as can be. Virtually every statistical estimation procedure will be unreliable if the sampling

procedures are inappropriate, and bad luck in the draw can lead any estimation procedure astray. There is something disingenuous in applying to causal inference a demand for reliability that would be dismissed out of hand as a requirement on statistical estimation. Only (vi) makes a sound point. Yule (1926) noted that time series samples of variables that tend to increase or decrease will be correlated. One might conclude that such correlations are due to remote unmeasured common causes, but for unspecified reasons Yule rejected that hypothesis. Non-stationary time series are special in several ways; they represent, for instance, a mixture of distributions. One might reasonably wonder what conditions exclude such cases, and whether there are specifiable general conditions on distributions, or distributions and causal structures, that, when they obtain, make causal inference possible in principle; one might reasonably wonder whether regression or other standard methods are optimal for such inferences. Such questions are never raised in the textbooks. For comparison, what would one think of a textbook on estimation that, knowing of circumstances in which an estimator is biased or inconsistent, refuses to discuss circumstances in which it is neither, or that, noting that some latent variable models are not identifiable, concludes that none are?

Pedagogy reflects accepted statistical theory, and with some important exceptions, statistical theory tiptoes around issues of causal inference. Even so, many research issues in statistics are fundamentally motivated by problems about reliable causal inference, although the motivation is sometimes hidden by the details.

The statistical literature on *collapsibility* provides one illustration. Some years ago the National Halothane Study (Bunker, Forrest, Mosteller and Vandam, 1969) found itself with more relevant variables than could be analyzed by methods then available. At stake were a set of causal questions about the effects and side-effects of alternative anesthetics. The problem suggested an important theoretical question: when can the same conclusions about the existence and strength of an influence of one variable, A , on another, B , be obtained by analyzing a reduced set of variables, \mathbf{K} (containing A and B), rather than a larger set of variables that properly includes \mathbf{K} ? A little more exactly the question is this: when and how can the analysis of a set \mathbf{K} of variables, including A and B , reliably determine whether variable A causes B , even though there may be common causes of A and B that are not in \mathbf{K} ; and when and how, in such circumstances, can we reliably predict the effect on B of manipulating A ? The question isn't peculiar to the halothane study; the issue arises in nearly every non-experimental study that tries to establish or assess a causal relationship. The question haunts non-experimental epidemiology. A pessimistic answer was part of Fisher's (1959) criticism of epidemiological arguments that smoking causes lung cancer.

Motivated in part by the problems of the halothane study, the influential book of Bishop, Fienberg and Holland (1975) replaced the question of identifying and estimating causal influence in the presence of latent variables with another question altogether: when are the log-linear parameters for a model obtained by marginalizing out some variables the same as the corresponding parameters of the larger, unmarginalized model?¹ Other authors have produced variants of this collapsibility question, sometimes outside the log-linear formalism (Asmussen and Edwards, 1983; Whittaker, 1990). The various conceptions of collapsibility have led to some nice mathematical work, but their bearing on the original issue is not clear; there is no reason to think that when a log-linear statistical model for a causally related set of variables collapses over a subset of the variables, the collapsed log-linear parameters correctly characterize anything about the causal relations among the variables in the marginal set.

"Model selection" or "specification search" is an area of research in which questions of causal inference have been badly obscured. Statistical "models," whether log-linear models, structural equation models, regression models, or whatever, often have two distinct roles. One role is to restrict the class of possible probability distributions among a set of variables and to parametrize the family of distributions that satisfy the restriction. Thus a log-linear model, for example, is given by specifying that particular parameters vanish in a linear expansion of the logarithm of the probability of any cell. The importance of hypothesis selection in this respect is that the restrictions and the parametrization should aid one in understanding and efficiently estimating the distribution. The other role such models may have is to inform prediction; in econometrics, epidemiology, market research and elsewhere, the predictions are often about the effects of actions or events that, if they were to occur, would *alter* the probability distribution. These are causal claims; they do not follow from any estimate of an actual probability distribution, and they depend on a further interpretation of the representations through which the restrictions of a statistical model are expressed. Statistical hypotheses used with a causal interpretation would seem either to be correct or to be incorrect, and the difference is important: smoking does cause lung cancer; ether, not halothane, was the riskier anesthetic.

We naive outsiders might therefore expect that research on the selection of linear or logistic regressors, the selection of log-linear parameters, the modification of structural equation parameters, and so forth, would pursue model selection as a kind of estimation problem for causal structure, analogous to the estimation of features of probability distributions. One might expect theoretical work to investigate conditions under which procedures can be found that give correct information about structure in the large sample limit, to characterize their error

¹We are indebted to Steve Fienberg for this account of the genesis of collapsibility questions.

probabilities in various cases, and to describe other computational and statistical properties of such procedures. But there is little if any work of this sort in the statistical literature. Instead, model selection and specification search have been treated quite differently from estimation. While texts and papers sometimes note conditions under which particular search procedures--stepwise regression, for example--may fail, search methods come without any kind of relevant assurances about their asymptotic reliability. There are not even very many large simulation studies of the reliabilities of common methods for specifying models.

Any number of other topics in statistics that concern the relation between statistical dependency, on the one side, and causal dependency on the other, have been addressed without the guidance of any theory of that relation. They include the rare discussions of statistically indistinguishable structural equation models (Basmann, 1965; Stetzl, 1986; Joreskog, 1990), discussions of Simpson's "paradox," and many issues of experimental design, including retrospective versus prospective sampling, and randomization in experiments.

Why is so much of statistical application and so little of statistical theory concerned with causal inference? One reason sometimes given for avoiding any attempt at a mathematical analysis joining causality and probability is that the idea of causality involves a lot of metaphysical murk which a mathematical science does well to avoid. Some people try to give accounts of causation entirely in terms of probability relations, while others try to characterize causation in terms of counterfactual conditions. Who is to say which is right, or if the counterfactual characterizations are even partly right, what exactly is meant? (We, certainly, have no *definition* of causation to promote.) But surely, the thought continues, there cannot be any rigorous theory about what is undefined.

This explanation of the neglect of causation in statistical theory is unsatisfactory on two counts. First, the absence of a "definition" of causation doesn't keep statisticians from talking comfortably about causal inference in experimental contexts, and surely the very fact that one variable causes another cannot always depend essentially on whether we *discover* the fact by experimentation. Second, while the notion of causality is vague in many respects, and wrapped in metaphysical disputes, so is the idea of *probability*. Every interpretation of probability appeals to obscure counterfactual assumptions or to mysterious properties. The classical definition appealed to "equipossible" cases. Limiting frequency interpretations must associate real finite empirical sequences with imaginary infinite sequences. Subjective Bayesian interpretations rest on a particularly obscure psychological notion, *belief*, and require assignments of degrees of belief that, in complex cases, no cognitively limited human being can instantiate (see, for example, Fine, 1973). Few Bayesian models in the literature are

specifications of the actual beliefs of any person, a fact sometimes sidestepped by claims that probabilities are degrees of belief of *ideally rational* agents, yet another class of non-existent. The obscurity of the fundamental notion of probability has not prevented the idea from bearing enormous fruit, and one important step in that process was to give a clear mathematical form to the theory. That happened first through the analytic characterization of special probability distributions and their properties, and later through the theory of measure, the Kolmogorov axioms and their variants. Notions of causality have paralleled and motivated developments in the theory of probability since the 17th century, and are undeniably entangled with probability assessments in scientific practice. Why, therefore, should we not give notions of causal dependence a clear mathematical form and make the relation of the causal and stochastic formalisms explicit in a way that reflects scientific practice? There are two attempts in the statistical literature to do that very thing, each valuable and neither sufficient of itself.

A mathematical representation of causal dependencies among a set of variables has been in the statistical literature for most of this century. Sewell Wright (1934) used directed graphs to represent causal structures. The vertices of the graph represent variables and a directed edge from one variable to another represents the claim that the first variable has a direct influence on the second, an influence not blocked by holding constant all other variables considered. Ever since, directed graphs have occasionally been used in representations of regression models, factor models, simultaneous equation models, time series models, and elsewhere. From a mathematical point of view, directed graphs are implicit whenever a statistical model is specified through a set of algebraic equations in which a single variable occurs on one side and is treated as the effect of variables on the other side.

Of itself, the mathematical representation of causal influence by directed graphs is trivial. Things only get interesting when some condition is given connecting the graphical structure with restrictions on probabilities. Some such connection was already implicit in the use of linear models in the social and behavioral sciences, for the system of equations and independence assumptions about error variables always determined a directed graph, sometimes given explicitly, and entailed constraints on correlations and partial correlations (Simon, 1954; Blalock, 1961), which for normal distributions is the same as constraints on independence and conditional independence. But social scientists did not articulate any general principle connecting their graphs with their probability distributions, and the literature developed no further than analyses of various particular cases (Blalock, 1971). Without using graphical representations, a few philosophers of science, most notably Reichenbach (1956) and Suppes (1970) attempted to give analyses of the very notion of causality in terms of statistical dependencies (and, in Suppes' case, time order). An explicit, general mathematical connection

between directed graphs and probability distributions was, however, not introduced until about ten years ago.

Kiiveri and Speed (1982) related causal dependencies, represented as directed graphs without cycles, to conditional independence constraints. They formulated several equivalent versions of the idea that if Y is not a cause of X and X influences Y , if at all, only through an intermediary set Z of direct causes of Y , so that if the variables in Z are held constant no variation in X will produce a variation in Y , then X and Y are independent conditional on Z . Formal statements of the idea are called *Markov conditions*, and we use one such formulation in this book. Kiiveri and Speed showed that social scientists' claims about vanishing correlations and partial correlations required by various linear models follow from the Markov Condition. They further showed that any strictly positive probability density satisfying the Markov condition for a directed acyclic graph admits a "factorization" determined by the structure of the graph: any joint density satisfying the Markov condition for a directed graph must equal a product of terms, one term for each variable, with each term giving the conditional probability of that variable on its parent variables in the graph. Thus for discrete variables if the graph is

$$X \longrightarrow Y \longrightarrow Z$$

Figure 1

the joint distribution must satisfy

$$P(X,Y,Z) = P(X) P(Y|X) P(Z|Y).$$

The conditional independence constraints required by the Markov condition applied to the graph are a consequence of the factorization. A widespread practice in applied statistics was thus given an elegant formal foundation and connected with statistical work on factorization of distributions that had previously been carried out by Wermuth (1980) and others. Consequences of the Markov Condition and additional constraints on the connection between directed graphs and distributions were subsequently developed by a number of authors (Wermuth and Lauritzen, 1983; Pearl, 1988; Wermuth and Lauritzen, 1990).

There is another thread in statistics connecting probability and causality, a connection that emphasizes relations between the notion of causation and the predictability of the effects of manipulations. It arose from work on experimental design. As early as 1935, Neyman noted a counterfactual aspect to conclusions drawn from an experimental study in which some units are

treated one way and some another. Often the conclusion of such studies is about what *would happen if* all units were treated the same way, and is therefore in a sense about a condition the experiment did not examine, or about a distribution it did not actually sample. Donald Rubin (1974, 1977, 1978, 1986), and following him several others (Holland, 1986; Pratt and Schlaifer, 1988), have interpreted causal hypotheses as postulating a family of random variables, some of which never have their values observed. In an experiment, random variables whose values are never observed represent the value an outcome variable *would have* (but doesn't in fact have) for a unit subjected to one treatment *if* that unit had instead been subjected to another treatment condition of the experiment. Various assumptions may constrain the relations between the contrary-to-fact random variables and random variables whose values are observed; for example, it might be assumed that if a treatment were applied to one unit, the treatment of that unit would have no effect on the outcome of applying the same treatment to other units. Rubin has used the framework to argue for the importance of randomization in experimental design and to give methods for estimating the effect of one variable on another by means of trials in which treatment is determined by the value of a variable that covaries with the outcome variable. Pratt and Schlaifer have given rules for predicting the invariance of conditional probabilities under interventions or manipulations: when does $P(Y)$, in a population in which a variable X is *forced* to have a value x , equal $P(Y|X=x)$ in the population of unmanipulated units? While intuitively correct, their rules were not explicitly derived from any general principle.

From our perspective, the Rubin, Holland, Pratt and Schlaifer theory is essentially an account of special cases in which the effects of an intervention or manipulation can be predicted. The work makes no use of directed graphical methods, but it makes a great deal of use of conditional independence relations. And that ties the Rubin framework to the Markov Condition. For a causal structure the Markov condition provides a factorization of the probability distribution in terms of the conditional probability of each variable X on its parents: $P(X|V_1\dots V_k)$. We can think of a "direct manipulation" of X as an intervention that changes $P(X|V_1\dots V_k)$ to some other distribution $P^*(X|V_1\dots V_k)$ but leaves the other conditional probabilities in the original factorization unaltered. The result of a direct manipulation is thus a new distribution with a new factorization, obtained by replacing $P(X|V_1\dots V_k)$ by $P^*(X|V_1\dots V_k)$. So for example, suppose the causal structure is $X <- Z -> Y$ with factorization $P(X|Z)P(Y|Z)P(Z)$. Let some intervention change the distribution of X to $P^*(X|Z)$. Then if the intervention were applied to all units in the population the new joint distribution would be $P^*(X,Y,Z) = P^*(X|Z)P(Y|Z)P(Z)$. In this example the effect on Y of such an intervention is trivial to predict since the marginal distribution of Y is unchanged.

With an appropriate definition of "manipulation" the Markov Condition entails a generalization of the principle just illustrated, a result we call the Manipulation Theorem. Although it seems never to have been formulated, instances of the theorem are implicit in some treatments of shocks in econometrics and in many discussions of experimental design. Rubin's (1977) claims about unbiased estimates of differential effects from alternative manipulations and Pratt and Schlaifer's rules all follow immediately from the assumption that the interventions in question are special cases of "manipulations." As we will show in Chapter 7, their analysis applies in the special case in which the intervention makes X statistically independent of its causes in the original unmanipulated system.

The Markov Condition is not given by God; it can fail for various reasons we will discuss in the course of this book. The reliability of inferences based upon the Condition is only guaranteed if substantive assumptions obtain. But the Condition is weak enough that there is often reason to think it applies. In most of the investigations in this book we combine the Markov condition with a further condition that assumes that all conditional independence relations among variables occur because of the Markov Condition applied to the graph of causal relations among the variables. This assumption, which we call the Faithfulness Condition, can be thought of formally as the claim that when a causal graph is associated with a probability distribution, the Markov Condition applied to the graph characterizes *all* of the conditional independence relations that hold in the distribution. Informally, the Faithfulness Condition can be thought of as the assumption that conditional independence relations are due to causal structure rather than to accidents of parameter values. Sometimes we investigate consequences of the special case that assumes the variables are linearly related, and in that context it can be shown that a version of the Faithfulness Condition holds almost always for a "natural" probability distribution over the parameters.

The immediate value of the unification of the graphical and experimental design approaches to statistics and causality is aesthetic: two distinct theories are unified in a simple way, and principles that appear *ad hoc* are derived uniformly. But aesthetics is the least of it; the real value of the unified theory lies in recasting a number of research topics so that fundamental issues can be addressed more directly, and in providing new algorithmic techniques for practical inference problems.

The Markov and Faithfulness Conditions, and their consequences, provide a framework for analyzing many of the questions about causality that we think are fundamental to the application of statistics: When are regression methods reliable for causal inference? Under what conditions are they optimal? Are more reliable methods possible? Are various model selection procedures

reliable? Are there more reliable methods? When are two alternative causal structures statistically indistinguishable? When can the presence of latent variables be detected from the statistics? Can predictions reliably be made when the causal structure is not known? Can predictions reliably be made when it is not known beforehand whether there may be latent variables acting to produce measured statistical dependencies? How does the reliability of causal inference depend on sampling procedure? Is retrospective sampling as useful for causal inference as prospective sampling? What discriminations about causal structure can be made by experiment but not by observation, or--incredible thought--vice-versa? Under what additional assumptions can what features of latent structure be reliably inferred? The unification of the graphical and experimental design frameworks provides results pertinent to all of these questions. Sometimes the results are versions of statistical common sense, but rather often they reverse received opinions. The results we have obtained are in many respects incomplete, but they are complete enough to demonstrate that issues of causation, prediction and search form an area of research that is fundamental to many statistical problems, and to illustrate that within the unified theory there is no need either to ignore the essential questions about reliability or to have recourse to ersatz technical questions.

There are two sorts of practical consequences of the theory. The unified theory yields precise conditions under which a variety of causal inferences cannot be reliably made, and the unified theory yields algorithms for inferring causal structure from statistical decisions, algorithms that, under appropriate assumptions, can be shown in a special but well defined sense to be asymptotically reliable. With such proofs in hand, statistics can collaborate with computer science, and reliable algorithms that are computationally feasible and that require feasible sets of statistical decisions can be obtained. By implementing these algorithms in a computer, the reliabilities of the procedures on small and medium samples, and their robustness to violations of the assumptions, can be explored by simulation methods. Most of the algorithms we have derived from the theory have been implemented in experimental versions of the TETRAD II program, and the results of various simulation tests are described in the course of this book.

1.2 Advertisements

Each of the examples in this section is also treated in other chapters. Their function here is only to tempt the reader to muster the patience to work through a long book with a lot of mathematics. Although the examples are not themselves proofs of the reliability of anything,

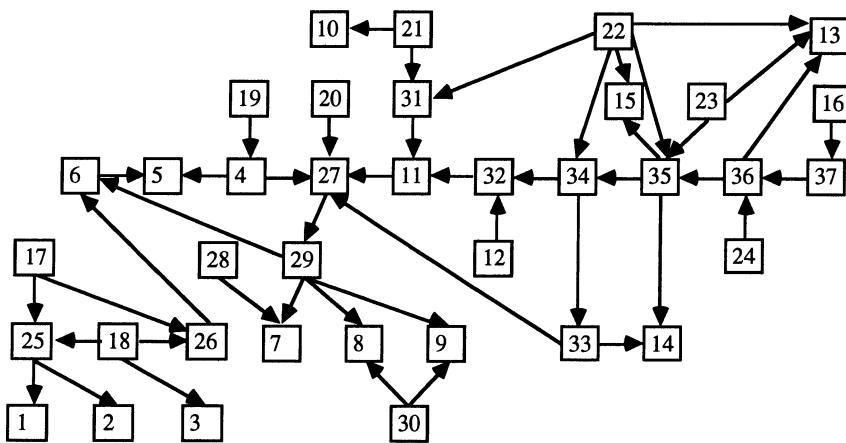
most of them illustrate procedures for which we will later give proofs of reliability assuming only the Markov and Faithfulness Conditions and the capacity to make correct statistical decisions.

1.2.1 Bayes Networks from the Data

The diagram shown below, called the ALARM network, was developed to model an emergency medical system (Beinlich, et al. 1989). The variables are all discrete, taking 2, 3 or 4 distinct values. In most instances a directed arrow indicates that one variable is regarded as a cause of another. The physicians who built the network also assigned it a probability distribution: each variable V is given a probability distribution conditional on each vector of values of the variables having edges directed into V . One use of such networks is to compute the probabilities of values of some variables for any new unit from measurements of values (for that unit) of other variables in the network. Of course such networks also constitute statistical models of discrete data, and as such they are alternatives to log-linear models, logistic regression models, or undirected graphical independence models. The directed graph has 37 variables and 46 edges. Herskovits and Cooper (1990) used the diagram to generate simulated emergency medicine statistics for 20,000 individuals.

For many statistical search procedures 37 might as well be infinity. Commonly recommended procedures for searching for models of discrete data in popular formalisms are stopped dead at ten or twelve variables. In contrast, Herskovits and Cooper describe a fast Bayesian procedure that, given a prior linear ordering of the variables consistent with the causal directions, makes only a single error in recovering the adjacencies in this graph from the discrete sample statistics. We will describe a procedure that recovers almost all of the ALARM network--including both the adjacencies and the directions of the edges--from the sample data. For example, there is an implementation on an ordinary workstation² that recovers most of the network in less than fifteen seconds from data generated by treating the dependencies in the graph as linear. For Herskovits and Cooper's discrete data generated from the same network, most of the structure can be recovered by computer with comparable reliability in a few minutes. Maximum likelihood estimates of the probabilities can be obtained directly. We emphasize that in these studies the procedure was not given any prior information about the graph, and that in most cases the computer determined the *directions* of the arrows.

²Decstation 3100.



The ALARM belief network

KEY:

- 1 - central venous pressure
- 2 - pulmonary capillary wedge pressure
- 3 - history of left ventricular failure
- 4 - total peripheral resistance
- 5 - blood pressure
- 6 - cardiac output
- 7 - heart rate obtained from blood pressure monitor
- 8 - heart rate obtained from electrocardiogram
- 9 - heart rate obtained from oximeter
- 10 - pulmonary artery pressure
- 11 - arterial-blood oxygen saturation
- 12 - fraction of oxygen in inspired gas
- 13 - ventilation pressure
- 14 - carbon-dioxide content of expired gas
- 15 - minute volume, measured
- 16 - minute volume, calculated
- 17 - hypovolemia
- 18 - left-ventricular failure
- 19 - anaphylaxis
- 20 - insufficient anesthesia or analgesia
- 21 - pulmonary embolus
- 22 - intubation status
- 23 - kinked ventilation tube
- 24 - disconnected ventilation tube
- 25 - left-ventricular end - diastolic volume
- 26 - stroke volume
- 27 - catecholamine level
- 28 - error in heart rate reading due to low cardiac output
- 29 - true heart rate
- 30 - error in heart rate reading due to electrocautery device
- 31 - shunt
- 32 - pulmonary-artery oxygen saturation
- 33 - arterial carbon-dioxide content
- 34 - alveolar ventilation
- 35 - pulmonary ventilation
- 36 - ventilation measured at endotracheal tube
- 37 - minute ventilation measured at the ventilator

Figure 2

1.2.2 Structural Equation Models from the Data

Rodgers and Maranto (1989) were interested in why some academic psychologists publish more than others. Is it abilities people already have before entering graduate school? Gender? The quality of their graduate training? The habits of publishing formed in graduate school? The quality of their first job? Rodgers and Maranto organized a survey and collected data on each of these variables, as well as others. Common sense determines the direction of most possible causal connections among their variables. For example, if gender and some of the other variables are causally connected, it cannot be because the other variables cause gender. Quality of first job, or publication rate after graduation cannot cause the quality of someone's graduate program. Rodgers and Maranto adapted two theories from economics to form two sets of linear equations, another from sociology, and still another from social psychology, estimating and testing each separate system of equations. Each of these theories failed a goodness of fit test. Combining the dependencies postulated in these rejected models, and adding more dependencies for good measure because they seemed plausible, Rodgers and Maranto then obtain a new model which they estimated, tested, and presented as a graph with eleven direct causal dependencies and associated standardized coefficients:

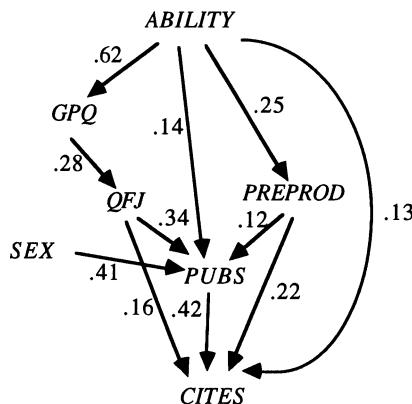


Figure 3

GPQ = graduate program quality

PUBS = publications since graduate school

QFJ = quality of first job

CITES = citation frequency

PREPROD = publications in graduate school

Rodgers and Maranto prefaced their paper with the remark that their elaborate exercise was necessary because causal dependencies cannot be inferred from the probabilities, but only from "theory." But given the common sense time order of the variables, a computer running a search algorithm will in a few seconds produce ten of the eleven hypothetical connections directly from their data, leaving out only one of the three dependencies associated with the smallest estimated linear coefficients.

Chapters 5 and 8 illustrate a variety of cases in which features of linear models that have been justified at length on theoretical grounds are produced immediately from empirical covariances by the procedures we describe. We also describe cases in which the algorithms produce plausible alternative models that show various conclusions in the social scientific literature to be unsupported by the data.

1.2.3 Selection of Regressors

Regression may be the method most commonly used in empirical studies to yield predictions about the effects of policies, often from variables measured roughly simultaneously and often without guarantees that no unmeasured factors affect the outcome variable and one or more regressors. Regression is something of a disaster when in such roles, and it is difficult to justify the prominence it is given in both pedagogy and arguments over policy.

Consider the following example, given by a set of linear equations with error terms and an unmeasured variable T , and also by a directed graph indicating the causal dependencies assumed for each unit in a population. The equations are

$$\begin{aligned} Y &= a_1X_1 + a_2X_5 + a_3T + \varepsilon_Y \\ X_1 &= a_4X_2 + a_5X_4 + \varepsilon_1 \\ X_3 &= a_6X_2 + a_7T + \varepsilon_3 \end{aligned}$$

The directed graph, with the error terms not represented, is:

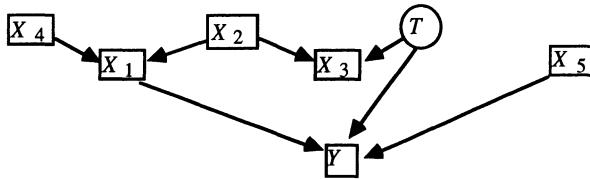


Figure 4

For definiteness, assume that all exogenous variables, including the error terms, are independent and distributed normally with zero mean and unit variance, and that none of the a_i vanish. Consider the multiple regression of Y on X_1-X_5 in such a population. The regression coefficient for X_3 provides an unreliable (biased, inconsistent) estimate of the influence of X_3 on Y , and this fact is often remarked in textbooks, usually in terms that do not explicitly mention causation. Perhaps exactly because they avoid explicit discussions of causality, however, the textbooks³ miss the more important fact that *when a regressor X and an outcome variable Y have an unmeasured common cause, the estimate of the influence on Y of every other regressor that directly influences X or has a common unmeasured cause with X will likewise be unreliable*. In the example in figure 4, even though only X_1 and X_5 are direct causes of Y , not mediated by other variables, only the regression coefficient for X_4 will be zero. For suitable ranges of values of the exogenous variances and linear coefficients, the regression coefficient of X_2 may even have a larger absolute value than that for X_1 . A similar phenomenon can occur if one of the regressors is actually an effect, rather than a cause, of the outcome variable, a circumstance that may not be uncommon in uncontrolled studies. As we will show with simulation studies later in this book, the usual computerized model selection techniques for regression typically fail in such cases; in fact, in simulation studies with large samples these methods do worse than simply regressing on all variables and choosing the significant regressors.

The unified theory yields an algorithm that under the statistical assumptions commonly made when such models are used (linearity, homoscedasticity, etc.) asymptotically produces correct information about the regressor structure. In the example just considered the procedure yields the object in figure 5, which we call a partially oriented inducing path graph.:

³Our nonrandom search through about 30 textbooks on regression found no mention of the fact in any book.

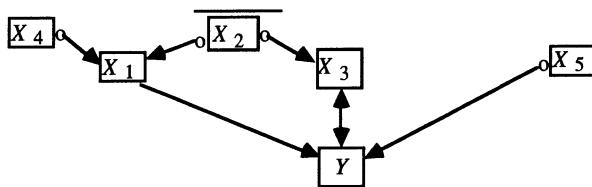


Figure 5

The double-headed arrow indicates the presence of unmeasured common causes; the circles at the ends of edges indicate that the algorithm cannot determine whether there should be arrowheads at those ends; that is, whether the $o \rightarrow$ edge should be \rightarrow or $\leftarrow\rightarrow$. The bar over X_2 indicates that not both of the small o marks on edges adjacent to X_2 may be arrowheads. If obtained in a context in which the Markov and Faithfulness Conditions are warranted, this graph would allow us to conclude that X_1 is a direct cause of Y , that X_5 may be a direct cause of Y , that X_3 does not cause Y and Y does not cause X_3 , that X_3 and Y are effects of a common cause, and that no other X variables are direct causes of Y .

Consider two empirical examples, which are described in more detail in Chapter 8. In the first, a sample of test results for more than 6,000 people includes values for each subject on seven test scores and on a combined score, $AFQT$, which is an average of three of the seven test scores as well as of other tests not included in the data. The dependency of $AFQT$ on the other recorded tests is therefore linear. The problem is to identify which of the seven tests are components of $AFQT$. Linear multiple regression of $AFQT$ on the other seven test scores gives significant regression coefficients to all of them. In contrast, an algorithm we will derive from the Markov and Faithfulness conditions correctly finds the three tests that are components of $AFQT$. In this case regression probably fails because the variables are related by an intricate structure of unmeasured common causes.

Rawlings (1988) describes a study (Linthurst, 1979) of 45 samples of Spartina grass from the Cape Fear Estuary. Besides the biomass of Spartina, 14 other variables were measured which might be thought to be relevant to growth. A linear multiple regression of biomass on the fourteen other variables found only two nutrient variables, copper and potassium concentration, to have significant coefficients, a result that is not plausible on biological grounds. With the same data, the alternative algorithm finds that pH controls growth in the sample, a result confirmed by experiment. In this case multiple regression may fail because significance tests of each regression coefficient must control for 13 other variables, which effectively reduces the

sample size from 45 to 32 and results in tests with little power against alternatives in which the partial correlations are not very large. In contrast the algorithms we use never require tests of vanishing partial correlations that control for more than one other variable in this case .

1.2.4 Causal Inference without Experiment

In his criticism of the epidemiological literature on smoking and lung cancer, Fisher (1959) emphasized that the correlation of two variables cannot distinguish a direct effect from an unmeasured common cause; the complaint was echoed in Brownlee's (1965) review in the *Journal of the American Statistical Association* of the first *Surgeon General's Report on Smoking and Health*. Fisher and Brownlee were, of course, both correct, but consider the cases illustrated in figure 6:

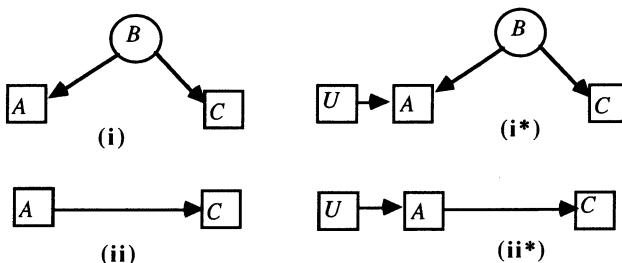


Figure 6

Suppose B is unmeasured. Fisher's point is that non-experimental studies of A and C cannot distinguish hypothesis (i) from hypothesis (ii), even if every unit in the population has the very same causal structure, but experiment can. But there are some points Fisher passed by.

If variable U is measured and causes A , (or has common causes with A), and U affects C only through A , as in the figures with an asterisk, then structures (i*), and (ii*) can be distinguished without experimental controls. It is not necessary that these restrictions on U be known independently of the data; all that is necessary is that they be true of the causal process generating the distribution.

Brownlee emphasized the case in which smoking is a cause of cancer, and smoking and cancer also have unmeasured common causes. As we will see in Chapter 9, that case presents

interesting complexities, but with modest prior knowledge it too can be distinguished from structures (i) and (ii) by appropriate observations.

The common presumption is that non-experimental evidence cannot determine whether measured statistical dependencies are due to unmeasured factors: one must decide on non-statistical grounds that there are unmeasured common causes at work producing statistical dependencies in a population. We can't help but think that in practice such decisions are often influenced more by convenience and the customs of disciplines than by any real knowledge. But under quite general conditions, whether for continuous linearly related variables or for discrete variables, the presumption is false. There are cases in which detailed inferences about causal structure can be made without any prior assumption as to whether or not unmeasured factors are at work, and in some of the same cases reliable predictions can be made about the effects of policies that directly manipulate certain variables while leaving the causal structure otherwise unaltered. Consider, for example, the following entirely hypothetical causal structure related to *Measured breathing dysfunction*, in which the variables in rectangles are measured, and those in ovals (*Environmental Pollution* and *Genotype*) are not.

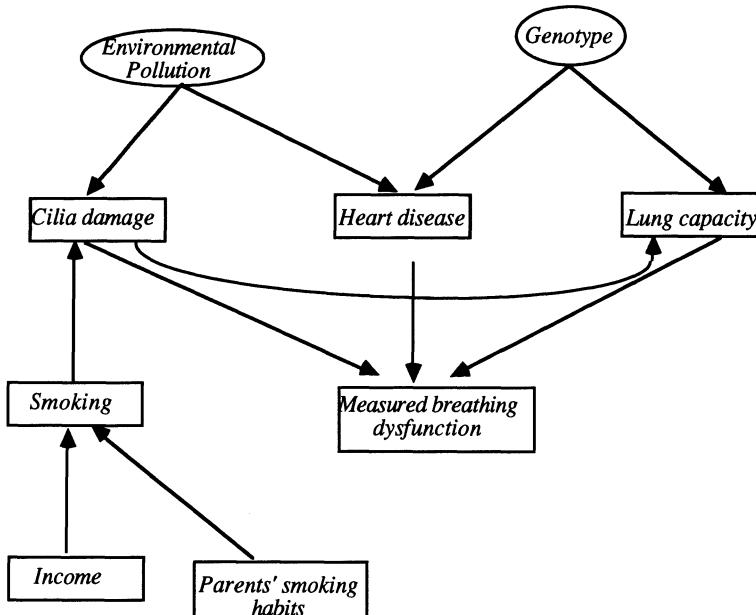


Figure 7

Suppose the structure in figure 7 represents the true causal dependencies among these variables in a population from which samples are drawn according to a multinomial distribution. The Markov and Faithfulness Conditions yield an algorithm that from the relations of conditional independence and dependence among the measured variables yields the graph in figure 8.

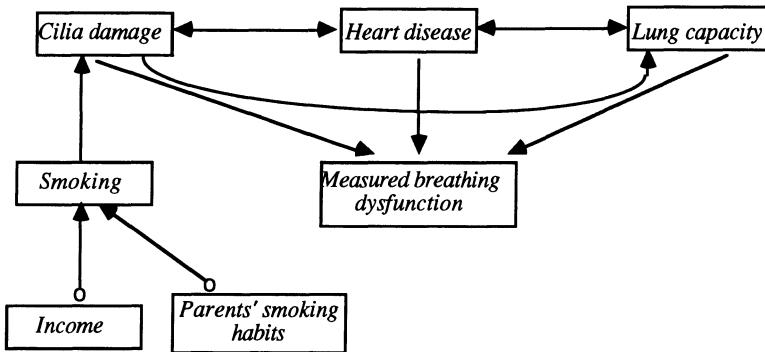


Figure 8

A double-headed arrow indicates the presence of unmeasured common causes; the circles at the ends of two of the edges indicate that the algorithm cannot determine whether there should be arrowheads at those ends. In this hypothetical example the statistics suffice to drive the conclusion that smoking indirectly affects lung capacity, and the statistics also answer most other questions about the causal dependencies among the variables, without any prior assumption of the existence or of the non-existence of unmeasured factors. Given the information about the population contained in the diagram in figure 8, and given adequate data, the effects of smoking cessation on cilia damage, heart disease, lung capacity and breathing dysfunction could be reliably predicted. Variants of the algorithm apply to continuous, linearly related variables.

1.2.5 The Structure of the Unobserved

Investigators who use psychometric or sociometric tests or questionnaires often have hundreds of item responses which in themselves are of no interest. What is of interest is the nature and causal relations of the features of persons or systems that the items indicate. On substantive or other grounds it may be believed that various items form clusters that have a common

unmeasured cause. The clusters may, of course, be heavily confounded: some of the items in the same cluster may have other common causes besides those common to all members of the cluster, and some members of a cluster may be affected by responses to other items, whether in that cluster or some other. Even if simplifying assumptions such as linearity are made, it is often thought to be utterly hopeless to extract from assumptions about the clusters and from the data any reliable conclusions about the causal relations of the unmeasured causes of the respective clusters. But provided the variables are to good approximation linearly related (or binary) and provided the clusters are not *too* confounded, reliable information about the causal relations among unmeasured variables can be obtained. If those causal relations are sufficiently sparse and the modeling assumptions apply, the causal structure may be identified almost uniquely. The procedures required have been fully automated. Consider the graph in figure 9, in which five (Y_1, Y_2, Y_4, Y_5 , and Y_{13}) of the 16 item responses are confounded.

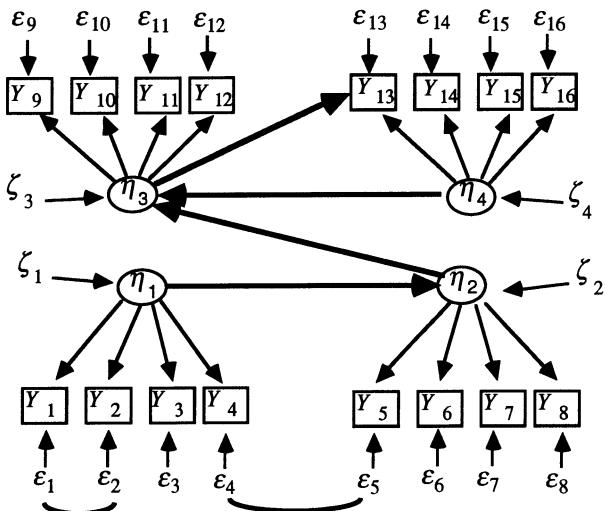


Figure 9

Given as input only the correlations among the Y variables and the four clusters shown, the procedure we will derive correctly reconstructs the graphical connections among the latent variables (the " η " variables) up to alternative orderings of the $\eta_1 - \eta_2$ connection, which are statistically indistinguishable.

1.3 Themes

A fundamental theme of this book is that there are systematic connections between causal dependency and stochastic dependency. By making use of these connections, the limits to reliable causal inference can be established theoretically. What can go wrong in causal inference and how wrong it can go only become clear once we have a theoretical understanding of the connection between causality and probability. One important class of limitations can be established by proving characterizations of equivalence relations between graph theoretic representations of alternative causal structures. In effect, one investigates the classification of directed graphs up to various relations of statistical indistinguishability. From the other side, one can consider characterizing the class of all graphs consistent with a given probability distribution. And the limits of causal inference can be investigated by characterizing circumstances in which the typical (in our case axiomatic) connections between causal and statistical structure are violated, a study begun by Yule and Pearson at the beginning of the century. Within these limits, we can investigate rigorously the reliabilities and computational properties of procedures that search for causal structure from statistical properties of samples. The asymptotic reliabilities of well-defined procedures can be determined mathematically, while short run reliabilities are estimated experimentally through simulation studies. The theory of model specification search consequently takes on some of the features of the theory of estimation.

Most current statistical modeling and search procedures can analyze only rather small models. Popular search and estimation procedures for log-linear models, for example, are stopped for realistic sample sizes at about a dozen variables or even fewer, depending on the number of values the variables may assume. By considering efficiencies of search, it is possible to find algorithms that (under explicit assumptions) will make reliable causal inferences for linear systems with a hundred or more variables when the causal connections among the variables are sparse. With discrete variables on sparse structures models can be reliably identified with fifty or more variables given samples of a few thousand units.

On the subject of latent variables, statistical practice is nearly schizophrenic. Statistical models are presented either with unmeasured variables or without them, but almost never with any credible argument from statistical considerations for their inclusion or their exclusion. Various model construction techniques proceed as if unmeasured common causes were the only possible kind, or alternatively as if they were absent and entirely irrelevant. Such assumptions need not

be left unargued: there exist asymptotically reliable methods to obtain information about the presence or absence of unmeasured common causes, and about their causal relations. Informative sufficient conditions exist for the presence of unmeasured common causes, assuming the Markov and Faithfulness Conditions. These conditions in turn yield theorems about causal conclusions and predictions that can be reliably drawn whether or not latent variables are present. If linearity is assumed then there are more powerful theorems, summarized in what we call the Tetrad Representation Theorem, that can be used to identify the presence of unmeasured common causes.

A further theme of this book concerns the conditions sufficient (or necessary) for correctly predicting the effects of a policy applied to a population that has been studied through a sample, and whose causal structure is not known prior to investigation. Statistical writers such as Rubin, Pratt, and Schlaifer have emphasized questions such as: When will the conditional distribution of Y on Z when X is forced to have a certain value equal the probability of Y conditional on X, Z in an observational or experimental population? Their answers assume that various counterfactual claims are known. In our terms, this amounts to knowing aspects of the causal structure of the systems under study. A natural question that follows is: When can the relevant causal knowledge necessary to answer the question above be obtained from sample data? We will describe results that answer this question and give information from sample data about when the conditional probability of Y on Z and X is invariant under a manipulation of X . In our view, however, the fundamental issue about prediction is this: *If the distribution of X_1, \dots, X_n is to be directly manipulated, when can the resulting distribution of a set Y of variables conditional on a set Z be calculated from the distribution of Y, Z, X_1, \dots, X_n and other variables in an observational or experimental population in which $X_1 \dots X_n$ were not so manipulated for each unit in the sample?* The formal connections between probability and causal structure determine an answer, and we will unravel part of it.

The rigorous investigation of the reliabilities of search algorithms leads directly to the conclusion that commonly used statistical search procedures are sub-optimal for causal inference. The criticism falls most heavily on regression. Better methods are available and easily applied to many regression problems. The best-founded objection to automated model search procedures, especially in linear and logistic regression, is that the procedures are asymptotically unreliable against alternative causal hypotheses that are often consistent with prior knowledge. If procedures give the right answers in the ideal case of perfect information about the population distribution, one can look around for better tests and more computationally efficient algorithms. But if, as in the case of regression and many other automated techniques, probability relations

and causal relations are incorrectly matched, all of the statistical subtlety in the world won't make for good inference.

Yet another theme of this book is the importance of causal reasoning in the design of empirical studies. The truth of the adage that correlation is not causation may be an obstacle to thinking through what *can be* determined about causal dependencies from statistical dependencies, and under what conditions. A recent textbook (Christensen, 1990, p. 279) on log-linear methods opens and closes the question with a remark that we think would be endorsed by the majority of statisticians:

Causation is not something that can be established by data analysis. Establishing causation requires logical arguments that go beyond the realm of numerical manipulation. For example, a well designed randomized experiment can be the basis for conclusions of causality but the analysis of an observational study yields information only on correlations. When observational studies are used as a basis for causal inference the jump from correlation to causation must be made on nonstatistical grounds.

The passage contrasts the information about causal structure in experimental studies with the information about causal structure in observational studies, and the view it expresses is perfectly standard. But is it correct? Once a formal understanding of the connection between causal structure and probability is in place, questions about the comparative power of experiment versus observation can be answered by a mathematical study of the causal information that can be extracted from experimental and from observational designs. The standard claims of the power of experiment and the impotence of observation turn out not to be so much false as misleading. There prove to be systematic parallels between the kinds of causal conclusions and predictions that can be drawn from experimental and observational data respectively. An intricate (and not yet complete) theory is required to understand the implications of the design of empirical studies for reliable causal inference and prediction.

The results we obtain about inference and prediction have interesting implications for a number of controversial subjects concerning the design of experiments, including the problem of designing "ethical" clinical trials in which subjects' preferences can influence the treatment they receive. We will describe a design that, depending on various discoverable empirical facts, permits patient preferences to have such a role without any loss in the power to obtain relevant information from experimental outcomes.

Although the methods we use are non-Bayesian, the broad issues of the subject cut across divisions between Bayesian and non-Bayesian statistics, and we will indicate some ways in which related inference results could be obtained by Bayesian procedures. The mathematical methods appropriate to the subject include more of graph theory and computation theory than measure theory. The empirical methods make heavy use of computer simulations to give evidence of reliability where analytic results are unavailable. Various pieces of methodological folklore are contradicted by the results already available in the subject, while others are given a new perspective and significance.

Whatever value the results in this book may have for practical scientific inference, together they illustrate a systematic and comparatively neglected area of inquiry that investigates causation, prediction and search. The subject is full of well-formed and almost well-formed open questions about systems with feedback, necessary and sufficient conditions for certain kinds of inferences, short-run reliabilities of procedures, the existence of optimal search procedures, optimal statistical decisions, indistinguishability properties of models, trade-offs between informativeness and computational feasibility, and more.

Chapter 2

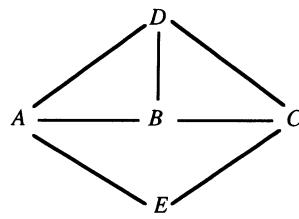
Formal Preliminaries

This chapter introduces some mathematical concepts used throughout the book. The chapter is meant to provide mathematically explicit definitions of the formal apparatus we use. It may be skipped in a first reading and referred to as needed, although the reader should be warned that for good reason we occasionally use nonstandard definitions of standard notions in graph theory. We assume the reader has some background in finite mathematics and statistics, including correlation analysis, but otherwise this chapter contains all of the mathematical concepts needed in this book. Some of the same mathematical objects defined here are given special interpretations in the next chapter, but here we treat everything entirely formally.

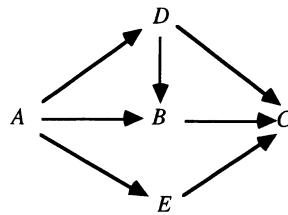
We consider a number of different kinds of graphs: directed graphs, undirected graphs, inducing path graphs, partially oriented inducing path graphs, and patterns. These different kinds of objects all contain a set of vertices and a set of edges. They differ in the kinds of edges they contain. Despite these differences, many graphical concepts such as undirected path, directed path, parent, etc., can be defined uniformly for all of these different kinds of objects. In order to provide this uniformity for the objects we need in our work, we modify the customary definitions in the theory of graphs.

2.1 Graphs

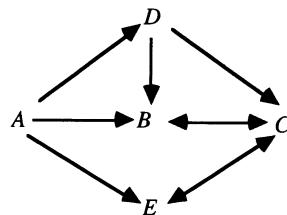
The undirected graph shown in figure 1 contains only undirected edges (e.g. $A - B$).

**Figure 1**

A directed graph, shown in figure 2, contains only directed edges, (e.g. $A \rightarrow B$).

**Figure 2**

An inducing path graph, shown in figure 3, contains both directed edges (e.g. $A \rightarrow B$) and bi-directed edges (e.g. $B \leftrightarrow C$). (Inducing path graphs and their uses are explained in detail in Chapter 6.)

**Figure 3**

A partially oriented inducing path graph, shown in figure 4, contains directed edges (e.g. $B \rightarrow F$), bi-directed edges (e.g. $B \leftrightarrow C$), nondirected edges (e.g. $E \circlearrowleft D$), and partially directed

edges (e.g. $A \rightarrow B$). (Partially oriented inducing path graphs and their uses are explained in detail in Chapter 6.)

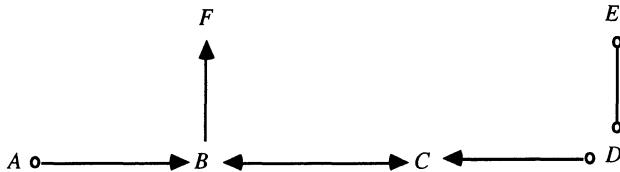


Figure 4

A pattern, shown in figure 5, contains undirected edges (e.g., $A - B$) and directed edges (e.g. $A \rightarrow E$). (Patterns and their uses are explained in detail in Chapter 5.)

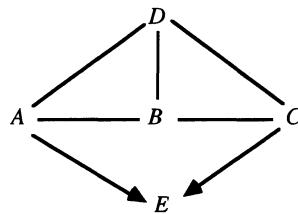


Figure 5

In the usual graph theoretic definition, a graph is an ordered pair $\langle V, E \rangle$ where V is a set of vertices, and E is a set of edges. The members of E are pairs of vertices (an ordered pair in a directed graph and an unordered pair in an undirected graph). For example, the edge $A \rightarrow B$ is represented by the ordered pair $\langle A, B \rangle$. In directed graphs the ordering of the pair of vertices representing an edge in effect marks an arrowhead at one end of the edge. For our purposes we need to represent a larger variety of marks attached to the ends of undirected edges. In general, we allow that the end of an edge can be unmarked, or can be marked with an arrowhead, or can be marked with an "o".

In order to specify completely the type of an edge, therefore, we need to specify the variables and **marks** at each end. For example, the left end of " $A \rightarrow B$ " can be represented as the

ordered pair $[A, o]$ ¹, and the right end can be represented as the ordered pair $[B, >]$. The first member of the ordered pair is called an endpoint of an edge, e.g. in $[A, o]$ the endpoint is A . The entire edge is a set of ordered pairs representing the endpoints, e.g. $\{[A, o], [B, >]\}$. The edge $\{[B, >], [A, o]\}$ is the same as $\{[A, o], [B, >]\}$ since it doesn't matter which end of the edge is listed first.

Note that a directed edge such as $A \rightarrow B$ has no mark at the A endpoint; we consider the mark at the A endpoint to be empty, but when we write out the ordered pair we will use the notation EM to stand for the empty mark, e.g. $[A, EM]$.

More formally, we say a **graph** is an ordered triple $\langle V, M, E \rangle$ where V is a non-empty set of vertices, M is a non-empty set of marks, and E is a set of sets of ordered pairs of the form $\{[V_1, M_1], [V_2, M_2]\}$, where V_1 and V_2 are in V , $V_1 \neq V_2$, and M_1 and M_2 are in M . Except in our discussion of systems with feedback we will always assume that in any graph, any pair of vertices V_1 and V_2 occur in at most one set in E , or, in other words, that there is at most one edge between any two vertices. If $G = \langle V, M, E \rangle$ we say that G is over V .

For example, the directed graph of figure 2 can be represented as $\langle \{A, B, C, D, E\}, \{EM, >\}, \{([A, EM], [B, >]), ([A, EM], [E, >]), ([A, EM], [D, >]), ([D, EM], [B, >]), ([D, EM], [C, >]), ([B, EM], [C, >]), ([E, EM], [C, >])\} \rangle$.

Each member $\{[V_1, M_1], [V_2, M_2]\}$ of E is called an **edge** (e.g. $\{[A, EM], [B, >]\}$ in figure 2.) Each ordered pair $[V_1, M_1]$ in an edge is called an **edge-end** (e.g. $[A, EM]$ is an edge-end of $\{[A, EM], [B, >]\}$.) Each vertex V_1 in an edge $\{[V_1, M_1], [V_2, M_2]\}$ is called an **endpoint** of the edge (e.g. A is an endpoint of $\{[A, EM], [B, >]\}$.) V_1 and V_2 are **adjacent** in G if and only if there is an edge in E with endpoints V_1 and V_2 (e.g. in figure 2, A and B are adjacent, but A and C are not.)

An **undirected graph** is a graph in which the set of marks $M = \{EM\}$. A **directed graph** is a graph in which the set of marks $M = \{EM, >\}$ and for each edge in E , one edge-end has mark EM and the other edge-end has mark " $>$ ".

An edge $\{<[A, EM], [B, >]\}$ is a **directed edge** from A to B . (Note that in an undirected graph there are no directed edges.) An edge $\{[A, M_1], [B, >]\}$ is **into** B . An edge $\{[A, EM], [B, M_2]\}$ is **out of** A . If there is a directed edge from A to B then A is a **parent** of B and B is a **child**

¹It is customary to represent the ordered pair A, B with angle brackets as $\langle A, B \rangle$, but for endpoints of an edge we use square brackets so that the angle brackets will not be misread as arrowheads.

(or **daughter**) of B . We denote the set of all parents of vertices in \mathbf{V} as **Parents**(\mathbf{V}) and the set of all children of vertices in \mathbf{V} as **Children**(\mathbf{V}). The **indegree** of a vertex V is equal to the number of its parents; the **outdegree** is equal to the number of its children; and the **degree** is equal to the number of vertices adjacent to V . (In a directed graph, the degree of a vertex is equal to the sum of its indegree and outdegree.) In figure 2, the parents of B are A and D , and the child of B is C . Hence, B is of indegree 2, outdegree 1, and degree 3.

We will treat an undirected path in a graph as a sequence of vertices that are adjacent in the graph. In other words for every pair X, Y adjacent on the path, there is an edge $\{[X,M_1],[Y,M_2]\}$ in the graph. For example, in figure 2, the sequence $\langle A,B,C,D \rangle$ is an undirected path because each pair of variables adjacent in the sequence (A and B , B and C , and C and D) have corresponding edges in the graph. The set of edges in a path consists of those edges whose endpoints are adjacent in the sequence. In figure 2 the edges in path $\langle A,B,C,D \rangle$ are $\{[A,EM],[B,>]\}$, $\{[B,EM],[C,>]\}$, and $\{[C,>],[D,EM]\}$.

More formally, an **undirected path** between A and B in a graph G is a sequence of vertices beginning with A and ending with B such that for every pair of vertices X and Y that are adjacent in the sequence there is an edge $\{[X,M_1],[Y,M_2]\}$ in G . An **edge** $\{[X,M_1],[Y,M_2]\}$ is **in path** U if and only if X and Y are adjacent to each other (in either order) in U . If an edge between X and Y is in path U we also say that X and Y are **adjacent** on U . If the edge containing X in an undirected path between X and Y is out of X then we say that the **path** is **out of** X ; similarly, if the edge containing X in a path between X and Y is into X then we say that the **path** is **into** X . In order to simplify proofs we call a sequence that consists of a single vertex an **empty path**. A path that contains no vertex more than once is **acyclic**; otherwise it is **cyclic**. Two paths **intersect** iff they have a vertex in common; any such common vertex is a **point of intersection**. If path U is $\langle U_1, \dots, U_n \rangle$ and path V is $\langle V_1, \dots, V_m \rangle$, then the **concatenation** of U and V is $\langle U_1, \dots, U_n, V_1, \dots, V_m \rangle$ denoted by $U \& V$. The concatenation of U with an empty path is U , and the concatenation of an empty path with U is U . Ordinarily when we use the term "path" we will mean acyclic path; in referring to cyclic path we will always use the adjective.

A **directed path** from A to B in a graph G is a sequence of vertices beginning with A and ending with B such that for every pair of vertices X, Y , adjacent in the sequence and occurring in the sequence in that order, there is an edge $\{[X,EM],[Y,>]\}$ in G . A is the **source** and B the **sink** of the path. For example, in figure 2 $\langle A,B,C \rangle$ is a directed path with source A and sink C . In contrast, in figure 2 $\langle A,B,D \rangle$ is an undirected path, but not a directed path because B and D occur in the sequence in that order, but the edge $\{[B,EM],[D,>]\}$ is not in G .

(although $\{[D,EM],[B,>]\}$ is in G .) Directed paths are therefore special cases of undirected paths. For a directed edge e from U to V ($U \rightarrow V$), $\text{head}(e) = V$ and $\text{tail}(e) = U$. A **directed acyclic graph** is a directed graph that contains no directed cyclic paths.

A **semi-directed path** between A and B in a partially oriented inducing path graph π is an undirected path U from A to B in which no edge contains an arrowhead pointing towards A (i.e. there is no arrowhead at A on U , and if X and Y are adjacent on the path, and X is between A and Y on the path, then there is no arrowhead at the X end of the edge between X and Y .) Of course every directed path is semi-directed, but in graphs with "o" end marks there may be semi-directed paths that are not directed.

A graph is **complete** if every pair of its vertices are adjacent. Figure 6 illustrates a complete undirected graph.

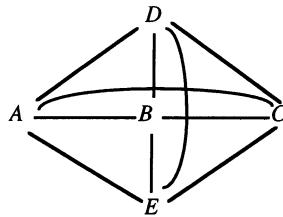


Figure 6

A graph is **connected** if there is an undirected path between any two vertices. Figures 1 - 6 are connected, but figure 7 is not.

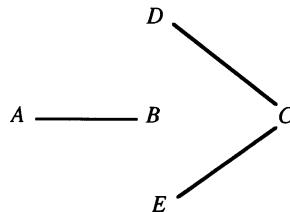


Figure 7

A **subgraph** of $\langle V, M, E \rangle$ is any graph $\langle V', M', E' \rangle$ such that V' is included in V , M' is included in M , and E' is included in E . Figure 7 is a subgraph of figure 1. The **subgraph of**

$\langle V, M, E \rangle$ over V' , where V' is included in V , is the subgraph $\langle V', M, E' \rangle$ in which an edge is in E' if and only if it is in E and has both endpoints in V' .

A **clique** in graph G is any subgraph of G that is complete. In figure 1, for example, the subgraph $G' =$

$$\langle \{A, B, D\}, \{EM\}, \{\{[A, EM], [B, EM]\}, \{[B, EM], [D, EM]\}, \{[A, EM], [D, EM]\}\} \rangle$$

is a clique with vertices A, B and D . A clique in G whose vertex set is not properly contained in any other clique in G is **maximal**. In figure 1, both G' and $G'' = \langle \{A, B\}, \{EM\}, \{\{[A, EM], [B, EM]\}\} \rangle$, are cliques, but G'' , unlike G' is not maximal because G'' is properly contained in G' .²

A **triangle** in a graph G is a complete subgraph of G with three vertices; in other words, vertices X, Y and Z form a triangle if and only if X and Y are adjacent, Y and Z are adjacent and X and Z are adjacent. In graph G a vertex V is a **collider on undirected path** U if and only if there are two distinct edges on U containing V as an endpoint and both are into V . Otherwise V is a **noncollider on U** . In graph G , vertex V is an **unshielded collider** on U if V is a collider on U , V is adjacent to distinct vertices V_1 and V_2 on U , and V_1 and V_2 are not adjacent in G . An **ancestor** of a vertex V is any vertex W such that there is a directed path from W to V . A **descendant** of a vertex V is any vertex W such that there is a directed path from V to W . In figure 2, A, B, C, D , and E are all ancestors of C , although neither A nor C is a parent of C . Similarly, C is a descendant of A, B, C, D , and E , although it is not a child of A or C . Since every vertex V is the source of a directed (empty) path from V to V , each vertex is its own descendant and its own ancestor, but not of course its own parent or its own child.

2.2 Probability

The vertices of the graphs we consider will always be random variables taking values in one of the following: a copy of the real line; a copy of the nonnegative reals; an interval of integers.

²Some writers, especially in statistics, understand "clique" as we have defined *maximal clique*.

By a joint distribution on the vertices of a graph we mean a countably additive probability measure on the Cartesian product of these objects. We say that two random variables, X, Y are **independent** when the joint density of (X, Y) is the product of the density of X and the density of Y for all values of X and Y . We write this as $X \perp\!\!\!\perp Y$. We generalize in the obvious way when asserting that one set of variables is independent of another set of variables. When we say a set of random variables is **jointly independent** we mean that any two disjoint subsets of the set are independent of one another. We say that random variables X, Y are **independent conditional on Z** (or given Z), when the density of X, Y given Z equals the product of the density of X given Z and the density of Y given Z , for all values of X, Y , and for all values z of Z for which the density of z is not equal to 0. We generalize in the obvious way for sets of random variables, X, Y, Z . If X is independent of Y given Z we write $X \perp\!\!\!\perp Y|Z$, and we say that the **order of the conditional independence** is equal to the number of variables in Z .

In the discrete case, we say that a distribution over V is positive if and only if for all values v of V , $P(v) \neq 0$. (In general, a distribution over V is positive if the density function is non-zero for all v .) If V is included in V' and

$$P(V) = \sum_{V' \setminus V}^{\rightarrow} P(V')$$

we will say that $P(V)$ is the **marginal** of $P(V')$ over V .

2.3 Graphs and Probability Distributions

We will examine several different graphical representations of conditional independence relations true in a distribution.

2.3.1 Directed Acyclic Graphs

A directed acyclic graph can be used to represent conditional independence relations in a probability distribution.

For a given graph G and vertex W let $\text{Parents}(W)$ be the set of parents of W , and $\text{Descendants}(W)$ be the set of descendants of W .

Markov Condition: A directed acyclic graph G over \mathbf{V} and a probability distribution $P(\mathbf{V})$ satisfy the Markov condition if and only if for every W in \mathbf{V} , W is independent of $\mathbf{V} \setminus (\text{Descendants}(W) \cup \text{Parents}(W))$ given $\text{Parents}(W)$.

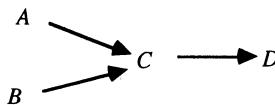


Figure 8

(Recall that W is its own descendant.) In the terminology of Pearl (1988) G is an **I-map** of P . In Figure 8, the Markov Condition entails the following conditional independence relations:³

$$\begin{aligned} A &\perp\!\!\!\perp B \\ D &\perp\!\!\!\perp \{A, B\} \mid C \end{aligned}$$

For all values of \mathbf{v} of \mathbf{V} for which $f(\mathbf{v}) \neq 0$, the joint density function $f(\mathbf{V})$ satisfying the Markov Condition is given by

$$f(\mathbf{V}) = \prod_{V \in \mathbf{V}} f(V \mid \text{Parents}(V))$$

where $f(V \mid \text{Parents}(V))$ denotes the density of V conditional on the (possibly empty) set of vertices that are parents of V . (See Kiiveri and Speed, 1982. Recall our notation convention that if $\text{Parents}(V) = \emptyset$, then $f(V \mid \text{Parents}(V)) = f(V)$.)

If a joint distribution over discrete variables satisfies the Markov Condition for figure 8 it can be factored in the following way:

$$P(A, B, C, D) = P(A) P(B) P(C \mid A, B) P(D \mid C)$$

³We do not include trivial independence relations, e.g., $C \perp\!\!\!\perp \emptyset \mid \emptyset$ which are true by definition.

for all values of A, B, C, D such that $P(A,B,C,D) \neq 0$. In a directed acyclic graph G , vertices of zero indegree are said to be **exogenous**. If G satisfies the Markov Condition for a distribution P , then for every pair of exogenous variables V_1 and V_2 , $V_1 \perp\!\!\!\perp V_2$ in P .

The Minimality Condition says, intuitively, that each edge in the graph prevents some conditional independence relation that would otherwise obtain.

Minimality Condition: If G is a directed acyclic graph over \mathbf{V} and P a probability distribution over \mathbf{V} , $\langle G, P \rangle$ satisfies the Minimality Condition if and only if for every proper subgraph H of G with vertex set \mathbf{V} , $\langle H, P \rangle$ does not satisfy the Markov Condition.

Returning to the example of figure 8, a distribution P' which satisfies the Markov Condition, but in which A is independent of $\{B, C, D\}$ does not satisfy the Minimality Condition, because P' also satisfies the Markov Condition for the subgraph in which the edge between A and C is removed. In the terminology of Pearl (1988) if a distribution $P(\mathbf{V})$ satisfies the Markov and Minimality conditions for a directed acyclic graph G , then G is a **minimal I-map** of P .

If a distribution P satisfies the Markov and Minimality Conditions for directed acyclic graph G , we will say that G **represents** P . For any directed acyclic graph G and for any probability distribution P satisfying the Markov and Minimality Conditions, if variables A and B are statistically dependent, then either:

- (i) there is a directed path in G from A to B ; or
- (ii) there is a directed path in G from B to A ; or
- (iii) there is a variable C and directed paths in G from C to B and from C to A .

A **trek** between distinct vertices A and B is an unordered pair of directed paths between A and B that have the same source, and intersect only at the source. The source of the pair of paths is also called the **source** of the trek. Note that one of the paths in a trek may be an empty path.

2.3.2 Directed Independence Graphs

Directed independence graphs are another (almost equivalent) way of representing conditional independence relations true of a probability distribution. Say that directed acyclic graph G is a **directed independence graph** of $P(\mathbf{V})$ (Whittaker 1990) for an ordering $>$ of the vertices of

G if and only if $A \rightarrow B$ occurs in G if and only if $\sim(A \perp\!\!\!\perp B \mid \mathbf{K}(B))$, where $\mathbf{K}(B)$ is the set of all vertices V such that $V \neq A$ and $V > B$.

Theorem 2.1: If $P(V)$ is a positive distribution, then for any ordering of the variables in V , P satisfies the Markov and Minimality conditions for the directed independence graph of $P(V)$ for that ordering.

If a distribution P is not positive, it is possible that the directed independence graph of P for a given ordering of variables is a subgraph of a directed acyclic graph for which P satisfies the Minimality and Markov conditions (Pearl, 1988).

2.3.3 Faithfulness

Given any graph, the Markov condition determines a set of independence relations. These independence relations in turn may entail others, in the sense that every probability distribution having the independence relations given by the Markov condition will also have these further independence relations. In general, a probability distribution P on a graph G satisfying the Markov condition may include other independence relations besides those entailed by the Markov condition applied to the graph. For example, A and D might be independent in a distribution satisfying the Markov Condition for the graph in figure 9, even though the graph does not entail their independence.

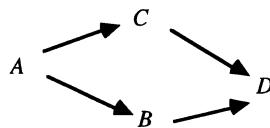


Figure 9

In linear models such an independence can arise if the product of the partial regression coefficients for D on C and C on A cancels the corresponding product of D on B and B on A .

If all and only the conditional independence relations true in P are entailed by the Markov condition applied to G , we will say that P and G are **faithful to one another**. We will, moreover, say that a distribution P is **faithful** provided there is some directed acyclic graph to which it is faithful. In the terminology of Pearl (1988) if P and G are faithful to one another

then G is a **perfect map** of P and P is a **DAG-Isomorph** of G . If distribution P is faithful to directed acyclic graph G , X and Y are dependent if and only if there is a trek between X and Y .

2.3.4 d-separation

Following Pearl (1988), we say that for a graph G , if X and Y are vertices in G , $X \neq Y$, and \mathbf{W} is a set of vertices in G not containing X or Y , then X and Y are **d-separated** given \mathbf{W} in G if and only if there exists no undirected path U between X and Y , such that (i) every collider on U has a descendent in \mathbf{W} and (ii) no other vertex on U is in \mathbf{W} . We say that if $X \neq Y$, and X and Y are not in \mathbf{W} , then X and Y are **d-connected** given set \mathbf{W} if and only if they are not d-separated given \mathbf{W} . If \mathbf{U} , \mathbf{V} , and \mathbf{W} are disjoint sets of vertices in G and \mathbf{U} and \mathbf{V} are not empty then we say that \mathbf{U} and \mathbf{V} are d-separated given \mathbf{W} if and only if every pair $\langle U, V \rangle$ in the cartesian product of \mathbf{U} and \mathbf{V} is d-separated given \mathbf{W} . If \mathbf{U} , \mathbf{V} , and \mathbf{W} are disjoint sets of vertices in G and \mathbf{U} and \mathbf{V} are not empty then we say that \mathbf{U} and \mathbf{V} are d-connected given \mathbf{W} if and only if \mathbf{U} and \mathbf{V} are not d-separated given \mathbf{W} . An illustration of d-connectedness is given in the following directed acyclic graph (but note that the definition also applies to other sorts of graphs such as inducing path graphs, as explained in Chapter 6):

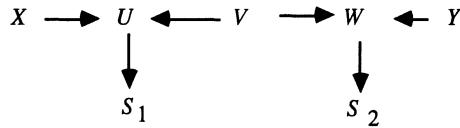


Figure 10

X and Y are d-separated given the empty set

X and Y are d-connected given set $\{S_1, S_2\}$

X and Y are d-separated given the set $\{S_1, S_2, V\}$

2.3.5 Linear Structures

A directed acyclic graph G over \mathbf{V} **linearly represents** a distribution $P(\mathbf{V})$ if and only if there exists a directed acyclic graph G' over \mathbf{V}' and a distribution $P'(\mathbf{V}')$ such that

- (i) V is included in V' ;
- (ii) for each endogenous (that is, with positive indegree) variable X in V , there is a unique variable ε_X in $V \setminus V$ with zero indegree, positive variance, outdegree equal to one, and a directed edge from ε_X to X ;
- (iii) G is the subgraph of G' over V ;
- (iv) each endogenous variable in G is a linear function of its parents in G' ;
- (v) in $P''(V')$ the correlation between any two exogenous variables in G' is zero;
- (vi) $P(V)$ is the marginal of $P''(V')$ over V .

The members of $V \setminus V$ are called **error variables** and we call G' the **expanded graph**. Directed acyclic graph G **linearly implies** $\rho_{AB,H} = 0$ if and only if $\rho_{AB,H} = 0$ in all distributions linearly represented by G . (We assume all partial correlations exist for the distribution.) If G linearly represents $P(V)$ we say that the pair $\langle G, P(V) \rangle$ is a **linear model** with directed acyclic graph G .

2.4 Undirected Independence Graphs

There is a well-known representation of statistical hypotheses about conditional independence by *undirected* graphs. The two representations, by directed and by undirected graphs, are closely related, but it is important not to confuse them.

An **undirected independence graph** G with a set of vertices V represents a probability distribution P if and only if there is no undirected edge between A and B just when A and B are conditionally independent given $V \setminus \{A, B\}$ in P . If an undirected independence graph G represents a distribution P , A and B are independent conditional on some set C if and only if every undirected path between A and B contains a member of C .

Suppose we consider a particular directed acyclic graph G and faithful probability distribution P . Let U be the undirected graph of adjacencies **underlying** G ; that is, U is the undirected graph with the same vertex set as G and the same adjacencies as G . Suppose that I is the undirected independence graph for the distribution P formed according to the definition just given. Then I and U are not in general the same, but U is always a subgraph of I . I and U will be the same if and only if G contains no unshielded colliders (Wermuth and Lauritzen 1983).

2.5 Deterministic and Pseudo-Indeterministic Systems

We will use the notion of a deterministic system in a technical sense: A joint probability distribution P on a set V of random variables represented by a directed acyclic graph G is **deterministic** if each of the vertices of G of non-zero indegree is a function of the vertices that are its immediate parents in G ; we will also say that G is a **deterministic graph** of P . By "function" we mean that for each assignment of a unique value to each of the parent vertices, there is a unique value of the dependent vertex.

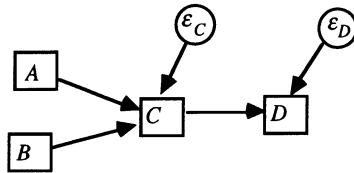


Figure 11

Suppose that the graph/distribution pair represented in figure 11 is deterministic, but that e_C and e_D are not measured. Were we to consider only the measured variables , i.e., A, B, C , and D , we would find that no variable has its value uniquely determined by the values of the others, although some of the variables are statistically dependent. The system looks indeterministic, although e_C and e_D are "hidden" variables which make it deterministic when added. Furthermore, it is not necessary to posit that two measured variables depend upon the same hidden variable, nor is it necessary to posit any dependence among the "hidden" variables in order to make the system deterministic. When a distribution represented by a directed acyclic graph among measured variables is not deterministic, but is embeddable in this way in a distribution represented by a directed acyclic graph that is, we say the distribution is pseudo-indeterministic.

In contrast, consider figure 12. Again suppose that only A, B, C , and D , were measured. In this case we could not make the system deterministic by adding hidden variables unless either

the hidden variables were associated or at least one hidden variable is adjacent to at least two of the measured variables.

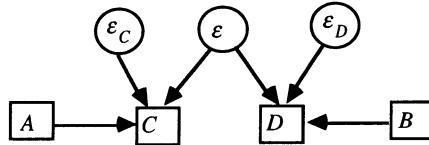


Figure 12

More formally, $\langle G, P \rangle$ is **pseudo indeterministic**, where P is a probability distribution over V and G is a directed acyclic graph over V , if and only if G is not a deterministic graph of P and there exists a distribution P' and a directed acyclic graph G' over a set of variables V' that properly includes V such that

- (i) G' is a deterministic graph of P' ;
- (ii) G is the subgraph of G' over V ;
- (iii) no vertex in V is an ancestor of a vertex in $V \setminus V'$;
- (iv) no vertex in $V \setminus V'$ is the source of a trek connecting two vertices in V ;
- (v) P is the marginal of P' ;
- (vi) G represents P .

If we say that $\langle P, G \rangle$ is **linear pseudo indeterministic** we mean that $\langle P, G \rangle$ is pseudo-indeterministic and in addition in G' , each vertex in V' is a linear function of its parents. A distribution linearly represented by a directed acyclic graph is pseudo-indeterministic. (Analogous definitions apply to Boolean pseudo indeterministic pairs of graphs and distributions, etc.)

2.6 Background Notes

Drawing from purely graph-theoretical work of Lauritzen, Speed and Vijayan (1978), and on statistical work in log-linear models (Bishop, Fienberg and Holland, 1975), in 1980 Darroch, Lauritzen and Speed introduced undirected graphical representations of log-linear hypotheses of conditional independence. Based on Kiiveri's thesis work, Kiiveri and Speed (1982) introduced versions of the Markov Condition, defined the notion of recursive causal model, obtained

maximum likelihood estimates for a multinomial distribution and provided a systematic survey of applications with both discrete and continuous variables. Shortly after, Kiiveri, Speed and Carlin (1984) further developed the formal foundations. Wermuth and Lauritzen (1983) introduced the notion of a recursive diagram, or what we have called a directed independence graph. The definitions of minimality and faithfulness are due to Pearl (1988).

Chapter 3

Causation and Prediction: Axioms and Explications

Views about the nature of causation divide very roughly into those that analyze causal influence as some sort of probabilistic relation, those that analyze causal influence as some sort of counterfactual relation (sometimes a counterfactual relation having to do with manipulations or interventions), and those that prefer not to talk of causation at all. We advocate no definition of causation, but in this chapter we try to make our usage systematic, and to make explicit our assumptions connecting causal structure with probability, counterfactuals and manipulations. With suitable metaphysical gyrations the assumptions could be endorsed from any of these points of view, perhaps including even the last.

3.1 Conditionals

Intelligent planning usually requires predicting the consequences of actions. Since actions change the states of affairs, assessing the consequences of actions not yet taken requires judging the truth or falsity of future conditional sentences--*If X were to be the case, then Y would be the case*. Judging the effects of past practice or policy requires judging the truth or falsity of counterfactual sentences--*If X had been the case , then Y would have been the case*.

Giving a detailed description of the conditions under which a future conditional or counterfactual conditional is true is a well-known and difficult philosophical problem. Lewis (1973) notes that *If kangaroos had no tails, they would topple over* is true even though we can imagine circumstances in which kangaroos use crutches. We mean that if things were pretty much as they are--given the scarcity of crutches for kangaroos and the disinclination of kangaroos to use crutches--if kangaroos had no tails they would topple over. But making this intuition precise is not easy.

It is widely recognized that causal regularities entail counterfactual conditionals; indeed this is often taken to be the feature that distinguishes a causal law from generalizations that are true, as it were, by accident. *All of the coins in your pocket are made of silver* does not entail the counterfactual *If this penny were in your pocket then it would be made of silver*. But the causal law *All collisions of electrons and positrons release energy* does entail the counterfactual *If this electron were to collide with this positron then energy would be released*.

The connection between causal regularities and the truth of future conditional and counterfactual sentences makes the discovery of causal structure essential for intelligent planning in many contexts. A linear equation relating the fatality rate in automobile accidents to car weight may be true of a given population, but unless it describes a robust feature of the world it is useless for predicting what would happen to the fatality rate if car weight was manipulated through legislation. Even quite accurate parametric representations of the distribution of values in a population may be useless for planning unless they also reflect the causal structure among the variables.

3.2 Causation

We understand causation to be a relation between particular events: something happens and causes something else to happen. Each cause is a particular event and each effect is a particular event. An event A can have more than one cause, none of which alone suffice to produce A . An event A can also be overdetermined: it can have more than one set of causes that suffice for A to occur. We assume that causation is transitive, irreflexive, and antisymmetric. That is, i) if A is a cause of B and B is a cause of C , then A is also a cause of C , ii) an event A cannot cause itself, and iii) if A is a cause of B then B is not a cause of A .

3.2.1 Direct vs. Indirect Causation

The distinction between direct and indirect causes is relative to a set of events. If C is the event of striking a match, and A is the event of the match catching on fire, and no other events are considered, then C is a direct cause of A . If, however, we added B : the sulfur on the match tip achieved sufficient heat to combine with the oxygen, then we would no longer say that C

directly caused A , but rather that C directly caused B and B directly caused A . Accordingly, we say that B is a **causal intermediary** between C and A if C causes B and B causes A .

Having fixed a context and a set of events, what is it for one event to be a direct cause of another? The intuition is this: once the events that are direct causes of A occur, then whether A occurs or not no longer has anything to do with whether the events that are indirect causes of A occur. The direct causes *screen off* the indirect causes from the effect. If a child is exposed to chicken pox at her daycare center, becomes infected with the virus, and later breaks out in a rash, the infection screens off the event of exposure from the occurrence of the rash. Once she is infected, whether she gets the rash has nothing to do with whether she was exposed to the virus from her daycare or from her Saturday morning playgroup.

Suppose V is a set of events including C and A . C is a **direct cause** of A relative to V just in case C is a member of some set C included in $V \setminus \{A\}$ such that (i) the events in C are causes of A , (ii) the events in C , were they to occur, would cause A no matter whether the events in $V \setminus \{A\} \cup C$ were or were not to occur, and (iii) no proper subset of C satisfies (i) and (ii).

3.2.2 Events and Variables

In order for causation to be connected with probabilities that can be estimated empirically, events must be sorted; some actual or possible events must be gathered together, declared to be of a type, and distinguished from other actual or possible events perhaps gathered into other types. The simplest classifications describe events as of a kind, e.g., solar eclipses, or declines in the Dow-Jones Industrial Average, and pair each event, A , of a kind with the event, $\neg A$, the non-occurrence of A . Such classifications permit us to speak intelligibly of *variables* as causes. We do so through the introduction of Boolean variables that take events of a kind, or their absences, as values. We say that **Boolean variable C causes Boolean variable A** if and only if at least one member of a pair (C , $\neg C$) causes at least one member of a pair (A , $\neg A$). Ordinarily no one would bother with collecting events into a type and examining causal relations among such variables unless the causal relations among events of the two types had some generality--that is, lots of events of type A have events of type C as causes and lots of events of type C have effects of type A , or none do.

Events can be aggregated into variables X and Y , such that some events of kind X cause some events of kind Y and some events of kind Y cause some events of kind X . In such cases there

will be no unambiguous direction to the causal relation between the variables. We consider this case in Chapter 12.

Some events are of a quantity taking a certain value, such as bringing a particular pot of water to a temperature of 100 degrees centigrade. Scales of many kinds are associated with an array of possible events in which a particular system takes on a scale value or takes on a value within a set of scale values. We can also speak of the variables of such scales as causes and effects, at least for particular systems over particular time intervals. For any particular system S we say that **scaled variable Q causes scaled variable R in S** provided that there is a value (or set of values) q for Q and a value (or set of values) r for R and a possible event in which Q taking value q in S would cause an event in which R takes value r in S . In practice we usually form scales only when we think the causal relations among values of different measures are not confined to particular values or particular systems but are more general. We sometimes say that the value r for R is caused by the value q for Q if the system taking on the value q for Q caused it to take on the value r for R . If \mathbf{K} is a collection of systems, we say that variable Q causes variable R in \mathbf{K} provided that for every system S in \mathbf{K} , Q causes R in S .

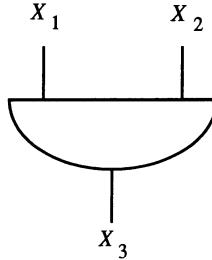
If our notion of causation between variables were strictly applied, almost every natural variable would count as a cause of almost every other natural variable, for no matter how remote two variables, A and B , may be, there is usually *some* physically possible--even if very unlikely--arrangement of systems such that variation in some values of A produces variation in some values of B . (A dictator could, we suppose, arrange circumstances so that the number of childbirths in Chicago is a function of the price of tea in China.) In practice, we always consider a restricted range of variation of other variables in judging whether A causes B . Strictly, therefore, our definitions of causal relations for variables should be relative to a set of possible values for other variables, but we will ignore this formality and trust to context. The notion of direct cause generalizes from events to variables in obvious parallel to the definition of causal dependence between variables: Variable C is a direct cause of variable A relative to V provided (i) C is a member of a set \mathbf{C} of variables included in V , (ii) there exists a set of values \mathbf{c} for variables in \mathbf{C} and a value a for A such that were the variables in \mathbf{C} to take on values \mathbf{c} , they would cause A to take on value a no matter what the values of other variables in V , and (iii) no proper subset of \mathbf{C} satisfies (i) and (ii). We say that a variable X is a **common cause** of variables Y and Z if and only if X is a direct cause of Y relative to $\{X,Y,Z\}$ and a direct cause of Z relative to $\{X,Y,Z\}$. If there is a sequence of variables in V beginning with A and ending with B such that for each pair of variables X and Y that are adjacent in the sequence in that order X is a direct cause of Y relative to V , then we say that there is a **causal chain** from A to B relative to V . A is an **indirect cause** of B relative to V if there is a causal chain from A to B

relative to \mathbf{V} of length greater than 2. We make the following two fundamental assumptions about causal relations: (i) if A is a cause of B then A is a direct cause or an indirect cause of B relative to \mathbf{V} ; (ii) if A , B , and C are in \mathbf{V} , and there exists a causal chain from A to B relative to \mathbf{V} that does not contain C , then for any set \mathbf{V}' that contains A and B there is a causal chain from A to B relative to \mathbf{V}' that does not contain C . When a cause is unmeasured it is sometimes called a **latent** variable. We say that two variables are **causally connected** in a system if one of them is the cause of the other or if they have a common cause. A **causal structure** for a population is an ordered pair $\langle \mathbf{V}, \mathbf{E} \rangle$ where \mathbf{V} is a set of variables, and \mathbf{E} is a set of ordered pairs of \mathbf{V} , where $\langle X, Y \rangle$ is in \mathbf{E} if and only X is a direct cause of Y relative to \mathbf{V} . We assume that in the population A is a direct cause of B either for all units in the population or no units in the population, unless explicitly noted otherwise. If it is obvious which population is intended we do not explicitly mention it. If $P(\mathbf{V})$ is a distribution over \mathbf{V} in a population with causal structure $C = \langle \mathbf{V}, \mathbf{E} \rangle$, we say that C **generated** $P(\mathbf{V})$. Two causal structures $\langle \mathbf{V}, \mathbf{E} \rangle$ and $\langle \mathbf{V}', \mathbf{E}' \rangle$ are **isomorphic** if and only if there is a one-to-one function f from \mathbf{V} onto \mathbf{V}' such that for any two members of A and B of \mathbf{V} , $\langle f(A), f(B) \rangle$ is in \mathbf{E}' if and only if $\langle A, B \rangle$ is in \mathbf{E} . A set \mathbf{V} of variables is **causally sufficient** for a population if and only if in the population every common cause of any two or more variables in \mathbf{V} is in \mathbf{V} , or has the same value for all units in the population.¹ We will often use the notion of causal sufficiency without explicitly mentioning the population.

3.2.3 Examples

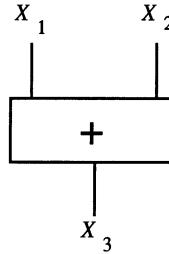
Simple digital logic circuit elements present concrete examples of causal structures. They are not of much intrinsic interest to most people, but they have the virtue that given a description of such a circuit element almost everyone can agree about which events pertaining to the circuit cause which other events. In the element illustrated below, the variables X_1 , X_2 and X_3 have two values, 1 and 0, accordingly as there is or is not a current through the corresponding line, and the semi-circle represents an "and" gate. Current flows from top to bottom. The value of the variable X_3 is thus a simple Boolean function of the values of X_1 and X_2 . If " \bullet " represents Boolean multiplication, $X_3 = X_1 \bullet X_2$.

¹Strictly, we require for causal sufficiency of \mathbf{V} for a population that if X is not in \mathbf{V} and is a common cause of two or more variables in \mathbf{V} , that the joint probability of all variables in \mathbf{V} be the same on each value of X that occurs in the population.

**Figure 1**

We understand the event of X_1 taking on value 1 and the event of X_2 taking on value 1 each to be a cause of the event in which X_3 takes on the value 1. We say that the Boolean variables X_1 and X_2 are each causes of the Boolean variable X_3 .

The form of the causal structure does not depend on the sort of variables involved or the particular class of functions among them. Isomorphic causal structures might be realized by a system of linear dependencies of continuous variables. Thus consider three different variables X_1 , X_2 , and X_3 , that represent the voltage in a given line and therefore range over the positive reals. Suppose we have a mechanism that outputs the sum of the voltage into it (figure 2).

**Figure 2**

In this case $X_3 = X_1 + X_2$, but the causal structure is isomorphic to the causal structure in figure 1: X_1 and X_2 each are causes of X_3 .

These examples suggest that the causal dependencies and the functional dependencies are related; X_3 is the effect of X_1 and X_2 , and X_3 is a function of X_1 and X_2 . In systems in which variables that are effects have their values uniquely determined by the values of all of the

variables that are their direct causes, functional dependence can be inferred from causal dependence by expressing each variable or event as a function of its direct causes. The converse does not hold: from the fact that an equation correctly describes a system one cannot infer that the direct causal dependencies in the system are reflected in the functional dependencies in the equation. For example, if the equation $X_3 = X_1 + X_2$ is true of a system then the equation $X_2 = X_3 - X_1$ is equally true of that system, but if X_1 and X_2 cause X_3 , then ordinarily X_3 and X_1 do not cause X_2 .²

3.2.4 Representing Causal Relations with Directed Graphs

Using the notion of a direct cause, it is trivial to represent causal structures with directed graphs:

Causal Representation Convention: A directed graph $G = \langle V, E \rangle$ represents a causally sufficient causal structure C for a population of units when the vertices of G denote the variables in C , and there is a directed edge from A to B in G if and only if A is a direct cause of B relative to V .³

We call a directed acyclic graph that represents a causal structure a **causal graph**. Figure 3 below is a causal graph for the circuit devices shown in figures 1 and 2.

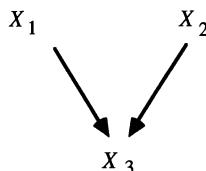


Figure 3

Consistently with our previous definition, if G is a causal graph and there is a vertex X in G and a directed path from X to Y that does not contain Z , and a directed path from X to Z that does not contain Y , we will say X is a **common cause** of Y and Z .

²Using the notion of identifiability, Simon (1953) proposed a general means to derive causal structure from a set of equations describing a system; later in the same paper Simon also proposed an account of causation using invariances under perturbations of linear coefficients.

³Since causation for variables is assumed to be transitive and irreflexive, the directed graph representing a causal structure must be acyclic. Introducing cyclic directed graphs requires a systematic reinterpretation.

There are important limitations to the Causal Representation Convention. Suppose drugs A and B both reduce symptoms C , but the effect of A without B is quite trivial, while the effect of B alone is not. The directed graph representations we have considered in this chapter offer no means to represent this interaction and to distinguish it from other circumstances in which A and B alone each have an effect on C . The interaction is only represented through the probability distribution associated with the graph. Consider another example, a simple switch. Suppose as in figure 4 battery A has two states: charged and uncharged. Charge in battery A will cause bulb C to light up provided the switch B is on, but not otherwise. If A and B are independent random variables, then A and C are dependent conditional on B and on the empty set, and B and C are dependent conditional on A and the empty set, and A and B are dependent conditional on C . The directed acyclic graph representing the distribution over A , B , and C therefore looks like the directed graph shown above. There is nothing wrong with this conclusion except that it is not fully informative. The dependence of A and C arises entirely through the condition $B = 1$. When $B = 0$, A and C are independent.

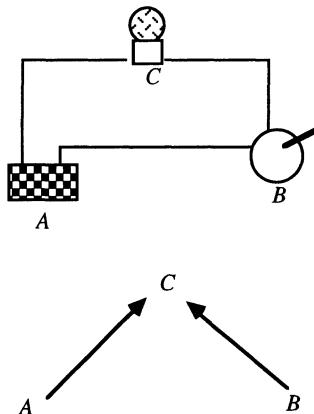


Figure 4

Since in discrete data the conditional independence facts, if known, identify the switch variables, a better representation would identify certain parents of a variable as switches. But a general representation of this sort would often not be very easy to grasp.⁴ Recent work on extending the directed acyclic graph representation to represent switches is described in Geiger and Heckerman (1991).

⁴A better practical arrangement might be a query system that, besides inferring the causal graph or graphs, responds to the user's questions about the effects of the manipulation of variables.

3.3 Causality and Probability

3.3.1 Deterministic Causal Structures

To good approximation the devices in figures 1 and 2 are **deterministic**, i.e., the effects are deterministic functions of their direct causes. If each effect is a linear function of its direct causes in the population, we say the system is a **linear deterministic causal structure** in the population.

Variables in a causal graph that have zero indegree, i.e., no causal input, are said to be **exogenous**. X_1 and X_2 are exogenous variables in the causal graph in figure 3. Variables that are not exogenous are **endogenous**. In a deterministic causal structure values for the exogenous variables determine unique values for the remaining variables.

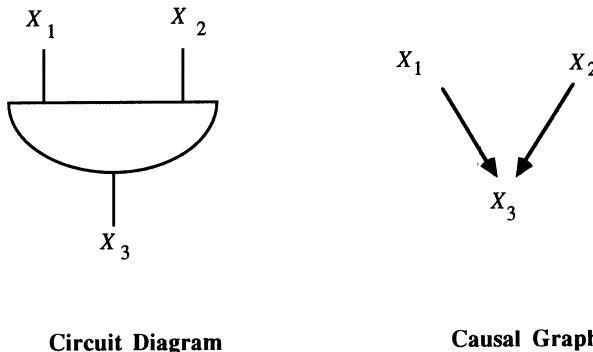


Figure 5

Consider the circuit element in figure 1 and its causal graph, both of which are shown in figure 5. Imagine an experiment to verify whether or not the device works as described. We would assign values to the **exogenous** variables, i.e., decide whether to put current into X_1 and X_2 , and then read whether or not X_3 has current. We can represent the experiment with the following table.

X_1	X_2	X_3
1	1	?
1	0	?
0	1	?
0	0	?

Suppose we were satisfied that the device usually worked as designed, but we wanted to know how often and in what way it fails. For each of a number of trials, we could randomly assign values to X_1 and X_2 , and then read whether or not X_3 has current. That is, we could assign a probability to each state the set of exogenous variables could occupy. For example,

$$\begin{aligned} P(X_1 = 1, X_2 = 1) &= 0.2 \\ P(X_1 = 1, X_2 = 0) &= 0.3 \\ P(X_1 = 0, X_2 = 1) &= 0.2 \\ P(X_1 = 0, X_2 = 0) &= 0.3 \end{aligned}$$

Because this causal structure is deterministic (even though the exogenous variables are random), a probability distribution over the exogenous variables determines a joint distribution for the entire set of variables in the system. For this example the joint distribution over (X_1, X_2, X_3) is:

$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 1) &= 0.2 \\ P(X_1 = 1, X_2 = 1, X_3 = 0) &= 0.0 \\ P(X_1 = 1, X_2 = 0, X_3 = 1) &= 0.0 \\ P(X_1 = 1, X_2 = 0, X_3 = 0) &= 0.3 \\ P(X_1 = 0, X_2 = 1, X_3 = 1) &= 0.0 \\ P(X_1 = 0, X_2 = 1, X_3 = 0) &= 0.2 \\ P(X_1 = 0, X_2 = 0, X_3 = 1) &= 0.0 \\ P(X_1 = 0, X_2 = 0, X_3 = 0) &= 0.3 \end{aligned}$$

We say that this distribution is generated by the causal structure of figure 5.

We use this example not to investigate sampling schemes for circuits but rather to illustrate how probability distributions are generated by deterministic causal devices. The only assumption we make about the connection between deterministic causal structures and the probability

distributions they may generate involves the distributions we will allow over the exogenous variables. *We assume that the exogenous variables are jointly independent in a probability distribution over the variables in a causally sufficient structure.* This is in part a substantive assumption--that statistical dependence is produced by causal connection--and in part a convention about representation. If exogenous variables in a structure are not independent, we expect that the causal graph is incomplete and there is some further causal mechanism, not represented in the graph, responsible for the statistical dependence. Either some of the input variables are causes of others (in which case we have equivocated, and the causal graph is not actually the graph of the causal structure of the structure) or else some nonconstant common causes of observed variables have not been included in the description of the structure.

3.3.2 Pseudo-Indeterministic and Indeterministic Causal Structures

In practice, the variables people measure are seldom deterministic functions of one another. We call a causal structure over a set V of variables for a population in which some variable is not a determinate function of its immediate causes in V an **indeterministic causal structure** for the population. An indeterministic causal structure might be pseudo-indeterministic. That is, a deterministic causal structure for which not all of the causes of variables in V are also members of V may appear to be indeterministic, even though there is no genuine indeterminism if the set of variables is enlarged by adding variables that are not common causes of variables in V . For example, suppose again that the device shown in figures 1 and 5 governs the current in line X_3 . Suppose also that X_2 is hidden from us so that only X_1 and X_3 occur in the causal structure we investigate. We might still hypothesize that X_1 is a cause of X_3 , thereby forming the causal graph on the right side of figure 6.

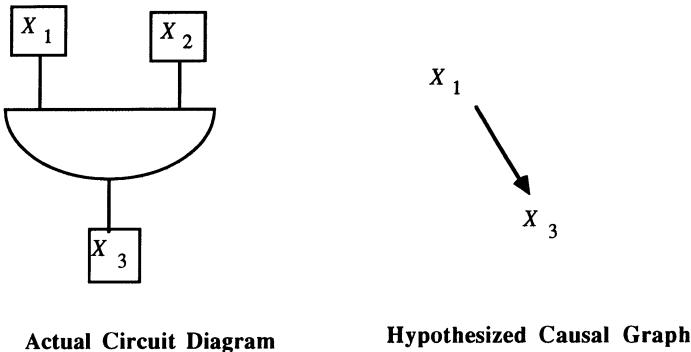


Figure 6

Assuming that the joint distribution $P(X_1, X_2, X_3)$ generated by the actual circuit device is the same as the one given for figure 5 in section 3.3.1, the observed distribution $P(X_1, X_3)$ is just the marginal of $P(X_1, X_2, X_3)$, namely:

$$\begin{aligned} P(X_1 = 1, X_3 = 1) &= 0.2 \\ P(X_1 = 1, X_3 = 0) &= 0.3 \\ P(X_1 = 0, X_3 = 1) &= 0.0 \\ P(X_1 = 0, X_3 = 0) &= 0.5 \end{aligned}$$

In the observed distribution X_3 is clearly not a function of its immediate parent X_1 and the causal structure appears to be indeterministic. We say the structure is pseudo-indeterministic. More formally, causal structure $C = \langle V, E \rangle$ is **pseudo-indeterministic** for a population, if and only if C is not a deterministic causal structure for the population and there exists a causal structure C' for the population over a set of variables V' that properly includes V such that

- (i) C' is a deterministic causal structure for the population;
- (ii) If A and B are in V , then $\langle A, B \rangle$ is in E if and only if $\langle A, B \rangle$ is in E' ;
- (iii) no variable in V is a cause of a variable in $V \setminus V'$;
- (iv) no variable in $V \setminus V'$ is a common cause of two variables in V ;

We say a structure is **linear pseudo-indeterministic** if all functional dependencies in C' are linear. Structural equation models in the social sciences are usually assumed to be pseudo-indeterministic causal structures. The error terms in such models are often interpreted as omitted causes.

A **genuinely indeterministic** causal structure for a population over a set of variables V is an indeterministic causal structure that is not pseudo-indeterministic. It is at least conceivable that there are genuinely indeterministic structures, even genuinely indeterministic macroscopic structures, whose variables have a causal structure. We will assume that the same relations between conditional independence and causal structure that obtain for pseudo-indeterministic structures hold as well for genuinely indeterministic causal relations, although as we will see later, there appear to be quantum mechanical systems for which that assumption must be carefully qualified. For a discussion of the case in which measured variables are exact functions of other measured variables see section 3.8.

3.4 The Axioms

We consider three conditions connecting probabilities with causal graphs: The Causal Markov Condition, the Causal Minimality Condition, and the Faithfulness Condition. These axioms are not independent. Consequences of various subsets of the conditions are investigated in the course of this book. We will consider justifications and objections to the conditions in the next section, but their importance—if not their truth—is evidenced by the fact that nearly every statistical model with a causal significance we have come upon in the social scientific literature satisfies all three: if the model were true, all three conditions would be met. While it is easy enough to construct models that violate the third of these conditions, Faithfulness, such models rarely occur in contemporary practice, and when they do, the fact that they have properties that are consequences of unfaithfulness is taken as an objection to them. In Chapters 5 and 8 we will consider published log-linear models, regression models, and structural equation models satisfying the three conditions.

3.4.1 The Causal Markov Condition

The intuitions connecting causal graphs with the probability distributions they generate are unified and generalized in one fundamental condition:

Causal Markov Condition: Let G be a causal graph with vertex set V and P be a probability distribution over the vertices in V generated by the causal structure represented by G . G and P satisfy the Causal Markov Condition if and only if for every W in V , W is independent of $V \setminus (\text{Descendants}(W) \cup \text{Parents}(W))$ given $\text{Parents}(W)$.

When G describes causal dependencies in a population with variables distributed as P satisfying the Causal Markov condition for G , we will sometimes say that P is **generated by** G . If V is not causally sufficient and V is a proper subset of the variables in a causal graph G generating a distribution P , we do *not* assume that the Causal Markov condition holds for the marginal over V of P .

The factorization results described in Chapter 2 apply to the joint probability distribution for a set V of variables in a population of systems with a causal structure satisfying the Causal Markov Condition. If $P(V | \text{Parents}(V))$ denotes the probability of V conditional on the (possibly empty) set of vertices that are direct causes of V , then

$$P(V) = \prod_{v \in V} P(v | \text{Parents}(v))$$

for all values of V for which each $P(v | \text{Parents}(v))$ is defined.

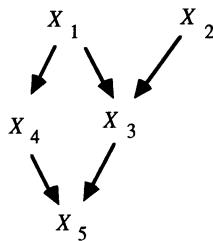


Figure 7

For the graph in figure 7 direct application of the Markov Condition yields a list of independence facts about the distribution generated by G .

$$\begin{aligned}
 & X_1 \perp\!\!\!\perp X_2 \\
 & X_2 \perp\!\!\!\perp \{X_1, X_4\} \\
 & X_3 \perp\!\!\!\perp X_4 \mid \{X_1, X_2\} \\
 & X_4 \perp\!\!\!\perp \{X_2, X_3\} \mid X_1 \\
 & X_5 \perp\!\!\!\perp \{X_1, X_2\} \mid \{X_3, X_4\}
 \end{aligned}$$

Other independence relations are entailed by these, for example

$$\{X_4, X_5\} \perp\!\!\!\perp X_2 \mid \{X_1, X_3\}$$

A discussion of axioms for conditional independence is found in Pearl (1988).

3.4.2 The Causal Minimality Condition

We will usually impose a further condition connecting probability with causality. The principle says that each direct causal connection prevents some independence or conditional independence relation that would otherwise obtain. For example, in the following causal graph G , C is a direct cause of A .

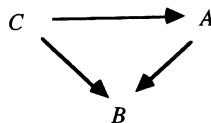


Figure 8

In a distribution P over $\{A, B, C\}$ for which $C \perp\!\!\!\perp A$, P satisfies the Markov condition even if the edge between C and A is removed from the graph.

Causal Minimality Condition: Let G be a causal graph with vertex set V and P a probability distribution on V generated by G . $\langle G, P \rangle$ satisfies the Causal Minimality condition if and only if for every proper subgraph H of G with vertex set V , the pair $\langle H, P \rangle$ does not satisfy the Causal Markov condition.

Since we will almost always give the graphs we consider a causal interpretation, we will in most cases hereafter simply describe these two conditions as the Markov and Minimality Conditions.

3.4.3 The Faithfulness Condition

Given a causal graph, the Markov condition determines a set of independence relations. These independence relations in turn may entail others, in the sense that every probability distribution having the independence relations given by the Markov condition will also have these further independence relations. In general a probability distribution P on a causal graph G satisfying the Markov condition may include other independence relations besides those entailed by the Markov condition applied to the graph. If, however, that does not occur, and all and only the independence relations of P are entailed by the Markov condition applied to G , we will say that P and G are **faithful to one another**. We will, moreover, say that a distribution P is faithful provided there is some directed acyclic graph to which it is faithful. So we consider a further axiom:

Faithfulness Condition: Let G be a causal graph and P a probability distribution generated by G . $\langle G, P \rangle$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in P is entailed by the Causal Markov Condition applied to G .

Note that a distribution P is faithful to G if and only if it satisfies *both* the Markov and Faithfulness Conditions. The Faithfulness and Markov Conditions entail Minimality, but Minimality and Markov do not entail Faithfulness. We will sometimes use the weaker axiom or axioms and more often the stronger one. Faithfulness turns out to be important to discovering causal structure, and it also turns out to be the "normal" relation between probability distributions and causal structures.

3.5 Discussion of the Conditions

When and why should it be thought that probability and causality together satisfy these conditions, and when can we expect the conditions to be violated? When should the values of variables in a population be thought to be distributed in accordance with the conditions?

3.5.1 The Causal Markov and Minimality Conditions

If we consider probability distributions for the vertices of causal graphs of deterministic or pseudo-indeterministic systems in which the exogenous variables are independently distributed, then the Markov Condition must be satisfied. A proof is given in the last chapter. We conjecture the Minimality Condition is true of all pseudo-indeterministic systems. The warrant for the conditions lies in this fact, and in the history of human experience with systems that we can largely control or manipulate. Electrical devices, mechanical devices, chemical devices all satisfy the condition. Large areas of science and engineering--from auto mechanics to chemical kinetics to digital circuit design--would be impossible without using the principles to diagnose failures and infer mechanisms.

In an important class of cases the application of the Minimality and Markov Conditions may be unclear. In 1903 G. Udny Yule concluded his fundamental paper on the theory of association of attributes in statistics with a section "On the fallacies that may be caused by the mixing of distinct records". (Yule uses $|AB|C$ to denote "the association between A and B in the universe of C 's" (p. 131)):

It follows from the preceding work that we cannot infer independence of a pair of attributes within a sub-universe from the fact of independence within the universe at large...The theorem is of considerable practical importance from its inverse application; i.e. even if $|AB|$ have a sensible positive or negative value we cannot be sure that nevertheless $|AB|C$ and $|AB|\gamma$ are not both zero. Some given attribute might, for instance, be inherited neither in the male line nor the female line; yet a mixed record might exhibit a considerable apparent inheritance. Suppose for instance that 50% of the fathers and of the sons exhibit the attribute, but only 10% of the mothers and daughters. Then if there be no inheritance in either line of descent the record must give (approximately)

fathers with attribute and sons with attribute:	25%
fathers with attribute and sons without attribute:	25%
fathers without attribute and sons with attribute:	25%
fathers without attribute and sons without attribute:	25%
mothers with attribute and daughters with attribute:	1%
mothers with attribute and daughters without attribute:	9%
mothers without attribute and daughters with attribute:	9%
mothers without attribute and daughters without attribute	81%

If these two records be mixed in equal proportions we get

parents with attribute and offspring with attribute	13%
parents with attribute and offspring without attribute	17%
parents without attribute and offspring with attribute	17%
parents without attribute and offspring without attribute	53%

Here $13/40 = 43$ [and] 1/3% of the offspring of parents with the attribute possess the attribute themselves, but only 30% of offspring in general, i.e. there is quite a large but illusory inheritance created simply by the mixture of the two distinct records. A similar illusory association, that is to say an association to which the most obvious physical meaning must not be assigned, may very probably occur in any other case in which different records are pooled together or in which only one record is made of a lot of heterogeneous material.

The fictitious association caused by mixing records finds its counterpart in the spurious correlation to which the same process may give rise in the case of continuous variables, a case to which attention was drawn and which was fully discussed by Professor Pearson in a recent memoir. If two separate records, for each of which the correlation is zero, be pooled together, a spurious correlation will necessarily be created unless the mean of one of the variables, at least, be the same in the two cases.

Yule's example seems to present a problem for the Causal Markov condition. Let a **mixture** over V be any population that consists of a combination of some finite number of subpopulations P_i each having different joint distributions over the variables in V , with each distribution satisfying the Causal Markov Condition for some graph. Consider a population that

is a mixture of structures $\langle G, P_1 \rangle$ and $\langle G, P_2 \rangle$ where P_1 and P_2 are distinct and satisfy the Markov Condition for G . Let the proportions in the mixture be $n:m$.

Let $P(X,Y,Z) = nP_1(X,Y,Z) + mP_2(X,Y,Z)$, with $n + m = 1$. A little algebra shows that $P(XY|Z) = P(X|Z)P(Y|Z)$ if and only if

$$(1) \quad n^2P_1(X,Y,Z)P_1(Z) + nmP_2(X,Y,Z)P_1(Z) + mnP_1(X,Y,Z)P_2(Z) + m^2P_2(X,Y,Z)P_2(Z) = \\ n^2P_1(X,Z)P_1(Y,Z) + nmP_1(X,Z)P_2(Y,Z) + mnP_2(X,Z)P_1(Y,Z) + m^2P_2(X,Z)P_2(Y,Z).$$

If $n, m > 0$ and in both distributions, X, Y are independent conditional on Z , that is $P_1(X,Y|Z) = P_1(X|Z)P_1(Y|Z)$ and $P_2(X,Y|Z) = P_2(X|Z)P_2(Y|Z)$, then equation (1) reduces to

$$(2) \quad P_2(X|Z)P_2(Y|Z) + P_1(X|Z)P_1(Y|Z) = P_1(X|Z)P_2(Y|Z) + P_2(X|Z)P_1(Y|Z)$$

The old but still rather surprising conclusion is that when we mix probability distributions we may find all possible conditional *dependence* relations. Thus, it seems, in many mixed populations conditional independence and dependence will not be a reliable guide to causal structure.

In the case of linear pseudo-indeterministic systems, when populations with two different distributions each associated with a linear structure are mixed, vanishing correlations in each separate distribution will not produce vanishing correlations in the mixed distribution, and vanishing partial correlations in each separate distribution will not produce vanishing partial correlations in the mixed distribution. It is easy to verify that for any mixture of two distributions--based on linear structures or not--the covariance of two variables vanishes in the mixture if and only if

$$k_1\text{COV}_1(XY) + k_2\text{COV}_2(XY) = k_1k_2[\mu_1X \mu_2Y + \mu_1Y \mu_2X] + \\ k_1(k_1 - 1)\mu_1X\mu_1Y + k_2(k_2 - 1)\mu_2X\mu_2Y$$

where the proportion of population 1 to population 2 is $n:m$ and $k_1 = n/(n+m)$, $k_2 = m/(n+m)$, and " μ_i " denotes the mean in population i .

So the situation is that we can have population 1 with causal graph G_1 and population 2 with causal graph G_2 , and the joint population will have a distribution that does not satisfy the Markov Condition for either graph. The question is whether such a mixed population violates

the Causal Markov Condition. *When a cause of membership in a subpopulation is rightly regarded as a common cause of the variables in V, the Causal Markov Condition is not violated in a mixed population;* instead, we have a population of systems satisfying the Causal Markov Condition but with a common cause (or causes) that may not have been measured. In some cases the cause of membership in a subpopulation may act like a latent switch variable of the kind considered in section 3.2.4; the distributions conditional on different values of the latent variable determine probability relations that are faithful to distinct causal graphs. In Yule's example, the missing common cause is gender. If, to take another example, we form a mixed sample of lead and copper pennies, within each subpopulation density and electrical conductivity will be independent, but in the mixed population they will be statistically dependent. We should say that is because chemical composition is a common cause of density and conductivity. In other cases the cause or causes of membership in relevant subpopulations may seem like *unnatural* kinds, or may at least not be the sort of causes a scientist seeks. Thus an important controversy (Caramazza, 1986) in contemporary cognitive neuropsychology concerns the use of statistical results for samples of people selected by syndrome, for example subjects with Broca's aphasia. One aim of studying such groups may be to discover if two or more normal capacities have a common cause damaged in Broca's aphasics. Suppose in a sample of Broca's aphasics a correlation is observed in scores on tests of two cognitive skills. Should the psychologist conclude that the test performances have a common latent cause? Perhaps, but the common cause need not be any functional capacity--damaged or otherwise--that causes both skills. Instead, the sample of Broca's aphasics might be a mixture of people with different sorts of brain damage, and within each subgroup the skills in question might be independently distributed. The common cause is only a variable representing membership in a subpopulation.

There are contexts in which the statistics of mixtures do not reflect any variable for population membership. In linear models the correlations and partial correlations are determined by the linear coefficients and the variances of exogenous variables. These parameters themselves may be treated as random variables and the resulting population distribution is a (generally uncountable) mixture of distributions. Statistical, but not causal, inference has been extensively studied in such settings (Swamy, 1971). If X is a random variable, we denote the expected value of X by $E(X)$.

Theorem 3.1 Let M be a linear model with directed acyclic graph G and linear coefficients a_{ij} . Let M' be a linear model with directed acyclic graph G , such that the linear coefficients in M' are random variables a'_{ij} that are jointly independent of all other random variables in M' , and $E(a'_{ij}) = a_{ij}$. Suppose the variances of the exogenous non-

coefficient random variables are the same in M and M' . Then $\rho_{AB,C} = 0$ in M' if and only if $\rho_{AB,C} = 0$ in M .

Thus a population that is a mixture of linear pseudo indeterministic causally sufficient systems with the same causal graph and with parameters independently distributed will satisfy the Causal Markov Condition for that graph without any unmeasured common cause.

Professional philosophers have offered a spate of criticisms of consequences of the Causal Markov Condition. Most of them appear to depend on omitting relevant latent variables. Wesley Salmon (1984) claims that "[t]here is another, basically different, sort of common cause situation" that cannot appropriately be characterized in terms of the Causal Markov Condition. Salmon calls this other causal relation an "interactive fork."

One putative example of an "interactive fork" is from Davis (1988):

Imagine a television set with a balky switch: it usually turns the set on, but not always. When the set is on, it produces both sound and picture. Then the probability of a picture given that the switch is on and given sound is greater than the probability of a picture given just that the switch is on. (Davis 88, p.156)

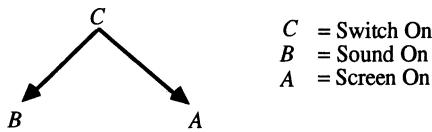
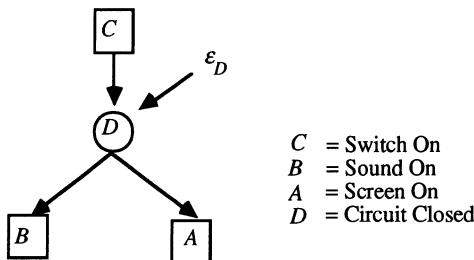


Figure 9

So, $P(B|C) < P(B|A \ \& \ C)$.

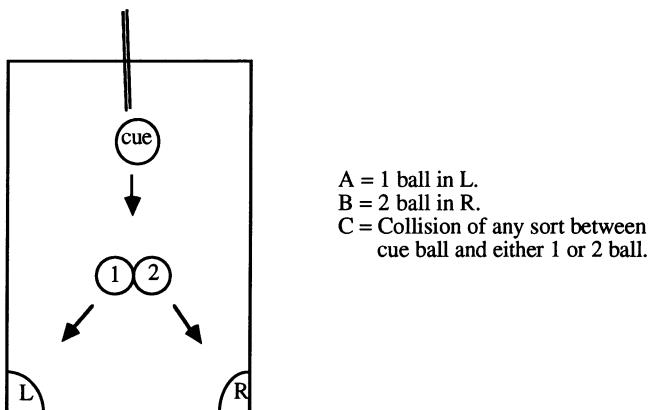
Davis' example gives an inaccurate picture of the causal situation, which is better depicted as follows:

**Figure 10**

The state of the circuit, or some variable downstream from the switch event, makes A and B independent.

Salmon's own illustration uses a slight variant of the following example from the game of pool (where we replace his events by Boolean variables).

C is the description of causal conditions relevant to both A and B , but A and B are not independent conditional on C .

**Figure 11**

Knowing C (that there was a collision) and A (that ball 1 dropped into its pocket) tells us more about whether B occurred (the 2 ball dropped into its pocket) than just knowing C . A and B are not directly causally connected, and they are not independent conditional on C .

In Salmon's example, event C does not completely describe *all* of the common causes of A and B . C tells us that there was a collision of some sort between the cue ball and the 1 or 2 balls, but it does not tell us the nature of the collision. A informs us about the nature of the collision and therefore tells us more about B . Were the prior event more informative--for example, were it to specify the exact momentum of the cue ball on striking the two target balls--conditional independence would be regained. The example simply reflects a familiar problem in real data analysis that arises whenever some proxy variable is used in causal analysis or distinct values of a variable are collapsed. In our view these examples give no reason to doubt the Causal Markov Condition.

Elliott Sober (1987) argues that we routinely find correlations for which there are no common causes, or for which residual correlations remain after conditioning on known common causes. The correlation of bread prices in England and the sea level in Venice may have some common causes (perhaps the industrial revolution), but not enough to account for all of the dependency. His point seems to be Yule's: if we consider a series in which variable A increases with time and a series in which variable B increases with time, then A and B will be correlated in the population formed from all the units-at-times, even though A and B have no causal connection. Any such combined population is obviously a mixture of populations given by the time values.

There is a more fundamental objection to the Causal Markov Condition, namely that there exist non-deterministic causal systems for which, to the best of current knowledge, the condition is false. Consider pair production: a quantum mechanical event produces two particles which move off in different directions. Because of conservation laws, dynamical variables in the two particles must be correlated; if one has a component of spin up, for example, the other must have that spin component down. We can do experiments in which for pairs we measure either of two different components of spin at two spatially separated sensors and compute the correlations. Suppose there is some state S of the system at the moment the pair of particles is produced such that, conditional on S , the dynamical variables of the two particles are uncorrelated. J. S. Bell (1964) argued that on such an assumption there follows an inequality constraining the correlations of the measured dynamical variables. While the assumptions needed for the derivation are controversial, the empirical facts seem beyond doubt: Bell's inequality is violated in certain quantum mechanical experiments. In such experiments the correlated variables are associated with spatially remote subsystems, so unless principles constraining causal processes to act "locally" that is, not instantaneously over a distance, are abandoned, any statistical dependency is presumably *not* due to the effect of one sub-system on

the other or to a common cause. Thus unless the locality principles are abandoned, the Causal Markov Condition appears to be false (Elby, 1992).

In our view the apparent failure of the Causal Markov Condition in some quantum mechanical experiments is insufficient reason to abandon it in other contexts. We do not, for comparison, abandon the use of classical physics when computing orbits simply because classical dynamics is literally false. The Causal Markov Condition is used all the time in laboratory, medical and engineering settings, where an unwanted or unexpected statistical dependency is *prima facie* something to be accounted for. If we give up the Condition everywhere, then a statistical dependency between treatment assignment and the value of an outcome variable will never require a causal explanation and the central idea of experimental design will vanish. No weaker principle seems generally plausible; if, for example, we were to say only that the causal parents of Y make Y independent of more remote causes, then we would introduce a very odd discontinuity: So long as X has the least influence on Y , X and Y are independent conditional on the parents of X . But as soon as X has no influence on Y whatsoever, X and Y may be statistically dependent conditional on the parents of Y .

The basis for the Causal Markov Condition is, first, that it is necessarily true of populations of structurally alike pseudo-indeterministic systems whose exogenous variables are distributed independently, and second, it is supported by almost all of our experience with systems that can be put through repetitive processes and whose fundamental propensities can be tested. Any persuasive case against the Condition would have to exhibit macroscopic systems for which it fails and give some powerful reason why we should think the macroscopic natural and social systems for which we wish causal explanations also fail to satisfy the condition. It seems to us that no such case has been made.

3.5.2 Faithfulness and Simpson's Paradox

Faithfulness can be violated in cases that realize variants of Simpson's "paradox" as Simpson originally presented it. We have already seen that both Yule and Pearson observed that two variables may be independent in subpopulations but dependent in a combined population. In 1948, M. G. Kendall used an example in his *Advanced Theory of Statistics* illustrating the reverse situation: two binary variables are independent but are dependent conditional on a third variable. Kendall's case was given a twist in a paper by Simpson (1951) a few years later, who thought his example introduced difficulties about the relation between causal dependencies and contingency tables. Subsequently the phenomenon the example exhibits has been referred to as

"Simpson's paradox." Like examples have since become standard puzzlers in discussions of the connection between causality and probability.

Kendall's example⁵ was as follows:

Consider the case in which a number of patients are treated for a disease and there is noted the number of recoveries. Denoting A by recovery, $\sim A$ by non-recovery, B by treatment, $\sim B$ by not-treatment⁶, suppose the frequencies are

	B	$\sim B$	Totals
A	100	200	300
$\sim A$	50	100	150
Totals	150	300	450

Here $(AB) = 100 = (A)(B)/N$, so that the attributes are independent. So far as can be seen, treatment exerts no effect on recovery. Denoting male sex by S_M and female sex by S_F , suppose the frequencies among males and females are

	Males		
	BS_M	$\sim B S_M$	Totals
AS_M	80	100	180
$\sim AS_M$	40	80	120
Totals	120	180	300

	Females		
	BS_F	$\sim B S_F$	Totals
AS_F	20	100	120
$\sim AS_F$	10	20	30
Totals	30	120	150

In the male group we now have

⁵p. 319. Q is Yule's $Q = (ad - bc)/(ad + bc)$ when the first row is a,b and the second c,d in a 2 X 2 table.

⁶(sic) Kendall means, of course, that the symbols denote the respective treatment and recovery states, not vice-versa.

$$Q_{AB,SM} = 0.231$$

and in the female group

$$Q_{AB,SF} = -0.429$$

Thus among the males treatment is positively associated with recovery, and among the females negatively associated. The apparent independence in the two together is due to canceling of these associations in the sub-populations.

Kendall's example is thus of a mixture of two distributions, one for males and one for females, such that the positive association between two variables in one population is exactly canceled by the negative association in the other. There is nothing paradoxical in that, and one may find empirical examples for which the same structure is claimed. The mixed distribution will violate the Faithfulness Condition, because it will exhibit a statistical independence relation that does not follow from the Markov condition applied to the causal graph common to all units.

Kendall's explanation of his contingency table depends on the fact that in one population the association of two variables is positive, and in the other negative. But what can be going on if in *both sub-populations the association is positive*, and yet in the mixed population it vanishes? That is exactly the question Simpson posed in 1951⁷. Simpson gave the following table and commentary:

	Male		Female	
	Untreated	Treated	Untreated	Treated
Alive	4/52	8/52	2/52	12/52
Dead	3/52	5/52	3/52	15/52

This time...there is a positive association between treatment and survival both among males and among females; but if we combine the tables we...find that there is no association between treatment and survival in the combined population. What is the

⁷Fienberg (1977), citing Darroch, attributes the issue to Yule "since Yule discussed it in the final section of his 1903 paper on the theory of association of attributes."(p. 51.) But save for the first sentence of that section, Yule actually discusses the reverse issue of mixtures, namely circumstances in which variables are statistically dependent in a population but independent in sub-populations.

"sensible" interpretation here? The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.

The question is what causal dependencies can produce such a table, and that question is properly known as "Simpson's paradox."⁸

In Simpson's example the variables G (male or female), T (treated or untreated) and S (survives or does not) are given an interpretation that imposes tacit restrictions on causal structure. When we read the example we naturally assume that gender G cannot be caused by treatment T or survival S , but may cause them. As with Kendall's example, the distribution in Simpson's table satisfies the Causal Markov Condition for a graph in which G causes T and S and T causes S . Simpson's distribution is not, however, faithful to such a graph, because T and S are independent in the distribution even though T is a parent of S in the graph.

Suppose for a moment that we ignore the interpretation that Simpson gave to the variables in his example, which was, after all, entirely imaginary, and let ourselves consider causal structures that would be excluded by that interpretation. To avoid substantive associations, we substitute A for T , B for G and C for S and obtain graph (i) in figure 12. Distributions such as Simpson's and Kendall's can also be realized by a graph in which A and C are not adjacent but each causes B , as in graph (ii) in figure 12.⁹

With the substitution of variables just noted, Simpson's distribution is faithful to graph (ii) but not to graph (i); moreover (ii) is the only graph faithful to the distribution.

⁸The subsequent literature has confused it with a number of other questions about how independence and dependence relations in a population may be related to independence and dependence relations in sub-populations, and the causal significance of such facts. The unfortunate aspect of collapsing these questions is that they have distinct answers. A circumstance attributed to Simpson and now often called "Simpson's paradox," but nonetheless distinct from the question Simpson actually posed, was described by Colin Blyth (1972):

It is possible to have simultaneously

(1) $P(A|B) < P(A|B')$

and

(2) $P(A|BC) \geq P(A|B'C)$

(3) $P(A|BC') \geq P(A|B'C')$

In fact, Simpson has equality in (1) and $>$ in (2) and (3).

⁹The point is implicit in Blalock (1961) and no doubt other sources as well.

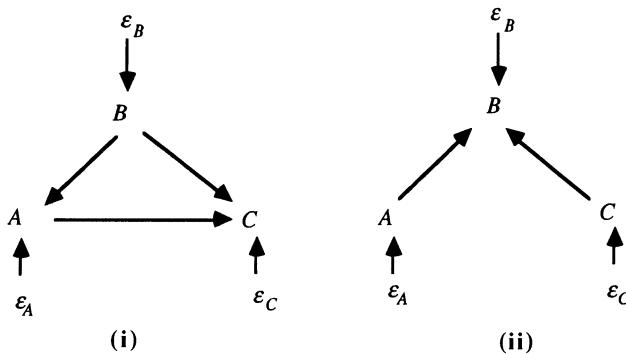


Figure 12

Judea Pearl (1988) offers a Bayesian example that illustrates why, when a causal structure like that in graph (ii) obtains, one should expect that A and C , though independent, are dependent conditional on B : Whether or not your car starts depends on whether or not the battery is charged and also on whether or not there is fuel in the tank, but these conditions are independent of one another. Suppose you find that your car won't start, and you hold in that case that there is some probability that the fuel tank is empty and some probability that the battery is dead. Suppose next you find that the battery is not dead. Doesn't the probability that the fuel tank is empty change when that information is added?

Were we to find that A and C are independent but dependent conditional on B , the Faithfulness Condition requires that if any causal structure obtains, it is structure (ii). Still, structure (i) is logically possible, and if the variables had the significance Simpson gives them we would of course prefer it. But if prior knowledge does not require structure (i), what do we lose by applying the Faithfulness Condition; what, in other words, do we lose by excluding causal structures that are not faithful to the distribution?

In the linear case, the parameter values--values of the linear coefficients and exogenous variances of a structure--form a real space, and the set of points in this space that create vanishing partial correlations not implied by the Markov Condition have Lebesgue measure zero.

Theorem 3.2: Let M be a linear model with directed acyclic graph G and n linear coefficients a_1, \dots, a_n and k positive variances of exogenous variables v_1, \dots, v_k . Let

$M(<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>)$ be the distributions consistent with specifying values $<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>$ for a_1, \dots, a_n and v_1, \dots, v_k . Let Π be the set of probability measures P on the space \Re^{n+k} of values of the parameters of M such that for every subset V of \Re^{n+k} having Lebesgue measure zero, $P(V) = 0$. Let Q be the set of vectors of coefficient and variance values such that for all q in Q every probability distribution in $M(q)$ has a vanishing partial correlation that is not linearly implied by G . Then for all P in Π , $P(Q) = 0$.

The theorem can be strengthened a little; it is not really necessary that the set of exogenous and error variables be jointly independent--pairwise independence is sufficient. In the pseudo-indeterministic case, faithfulness can be violated, if at all, only by very special choices of the functional dependencies between variables. Consider a population of linear, pseudo-indeterministic systems in which the exogenous variables are independently and normally distributed. The conditional independence relations required by the Markov Condition will be automatically fulfilled for every possible value of the linear coefficients--they are guaranteed just by the way the device acts to compose linear functions. But conditional independence relations that are *not* required by the Markov Condition--the sorts of conditional independence relations that characterize distributions that are unfaithful to the causal structure of the devices--either cannot be produced at all or can only be produced if the linear coefficients satisfy very strong constraints.

The same moral applies to other classes of functions. While for discrete variables we have not attempted a formal proof of a theorem analogous to 3.2, such a result should be expected on intuitive grounds. The factorization formula for distributions satisfying the Markov Condition for a graph provides a natural parametrization of the distributions. If an exogenous variable has n values, it determines $n-1$ parametric dimensions consisting of a copy of the open interval $(0,1)$. If an endogenous variable X has n values, a conditional probability $P(X|Parents(X))$ in the factorization determines another $n-1$ parametric dimensions consisting of a copy of $(0,1)$ for each vector of values of the parents of X . One expects that the set of probability values that generate conditional independence relations not entailed by the factorization itself will be measure zero in this parameter space.

3.6 Bayesian Interpretations

We have interpreted the conditions as about frequencies in populations in which all units have the same causal structure. We wish to consider how the conditions can be given a Bayesian interpretation in which the probabilities are subjective. Current subjectivist interpretations hold that probability is an idealization of rational, subjective degree of belief. On a strict subjectivist view there can be finite frequencies, but there is no such thing as objective probability. One assumes the systems under study in the sciences are deterministic, and any appearance of indeterminacy is due simply to ignorance. The likelihood structures of Bayesian statistical models often look like ordinary un-Bayesian statistical models; Bayesians add a prior probability distribution over the free parameters. For example, Bayesian linear models specify a distribution over a parameter Θ representing linear coefficients, variances, means, and so on. The Bayesian model is thus a mixture of ordinary linear models, and the joint distribution over the measured variables does not satisfy the conditions we have considered in this section.

Consider a study of systems with causal graph G . Suppose a Bayesian agent's degrees of belief are represented by a density, f , satisfying the condition $f(X|\text{Parents}(G,X)) = h(\text{Parents}(G,X); \Theta)$, where Θ is a parameter whose values determine a density for X conditional on its parents. Let the Bayesian agent also have a distribution over Θ . In such a case we understand the Causal Markov and Causal Minimality conditions to constrain the agent's degrees of belief conditional on Θ . The subjective joint distribution over the variables conditional on Θ will satisfy the conditions, but typically the unconditional joint distribution will not.

Suppose now that the agent entertains a set G of alternative possible causal structures, and holds that in each structure G $f(X|\text{Parents}(G,X)) = h(\text{Parents}(G,X); \Theta_G)$, as before. Then we understand the Causal Markov and Causal Minimality conditions to constrain the agent's degrees of belief conditional on Θ_G, G .

So understood, the conditions are normative principles about "reasonable" degrees of belief. In a later chapter we will consider in some detail a Bayesian proposal for clinical trials and argue that the assumptions the proposal makes about the degrees of belief of scientific experts accords with the Markov Condition.

3.7 Consequences of The Axioms

Consequences of the Causal Markov, Minimality and Faithfulness Conditions are developed throughout this book, but some important connections between causal dependency and statistical dependency should be noted here.

3.7.1 d-Separation

Given a causal graph G , the Markov Condition axiomatizes the set of independence and conditional independence relations true of any distribution P faithful to G . But which conditional independence relations follow from the Markov Condition for a given graph may not be obvious. Suppose one wanted to know, for each pair of vertices X and Y and each set of vertices Q not containing X and Y , whether or not X and Y are independent conditional on Q , i.e., all the atomic independence facts among sets of variables. Applying the Markov Condition directly to G , that is, applying the definition for each vertex, does not in general suffice.

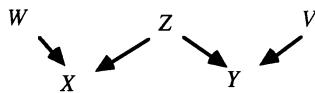


Figure 13

For example, in a distribution faithful to the graph in figure 13, suppose we wanted to know whether X and Y are independent conditional on the set $Q = \{Z\}$. Applying the Markov Condition directly to figure 13, we obtain:

$$\begin{aligned}
 &W \perp\!\!\!\perp \{Z, Y, V\} \\
 &X \perp\!\!\!\perp \{Y, V\} \mid \{W, Z\} \\
 &Z \perp\!\!\!\perp \{W, V\} \\
 &Y \perp\!\!\!\perp \{W, X\} \mid \{V, Z\} \\
 &V \perp\!\!\!\perp \{W, X, Z\}
 \end{aligned}$$

It is not obvious that these facts entail $X \perp\!\!\!\perp Y \mid \{Z\}$. Pearl proposed a purely graphical characterization--which he called **d-separation**--of conditional independence, and Geiger,

Pearl, and Verma (Geiger and Pearl 1989a; Verma 1987) proved that d-separation in fact characterizes all and only the conditional independence relations that follow from satisfying the Markov condition for a directed acyclic graph.

The definition of d-separation is sufficiently unintuitive that an analogy may be helpful. Consider first the situation with unconditional independence. Think of an undirected path in a graph as a pipe carrying causal flow, and each of the vertices are valves that are either active (open), or inactive (closed). If a vertex is a collider, then causal flow cannot get through it, and it is thus inactive. For example, in the causal graph at the top of figure 14, X and V are d-separated by the empty set because Y is a collider on the only path between them.

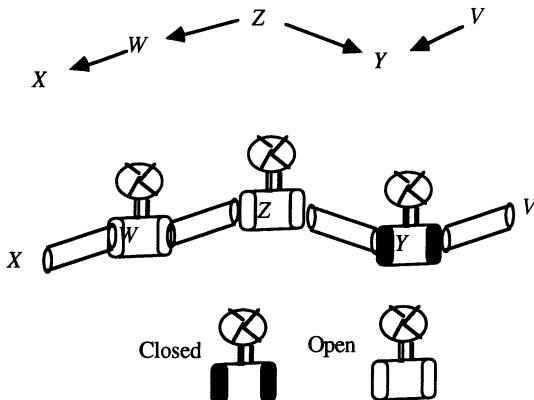


Figure 14

Conditioning on a node flip-flops its status. Whereas X and Y are not d-separated given the empty set in the graph in figure 14, they are d-separated given $\{W\}$, $\{Z\}$ or $\{W, Z\}$. X and V are d-separated given the empty set, but are not d-separated given $\{Y\}$. That conditioning on a non-collider makes it inactive is similar to the intuition behind the Markov Condition. A non-collider is either a common cause, e.g. Z , or part of a causal chain, e.g., W . Effects are made independent when we condition on their common causes, and effects are made independent of their remote causes when we condition on their more proximate ones. That conditioning on a collider makes it active was noted in section 3.5.2 above.

Given a graph G , checking whether any two vertices X and Y are d-separated by a set Q and thus independent conditional on Q in a distribution faithful to G appears straightforward. A path is active if all of its vertices are active, i.e. if all colliders are in Q and all of its non-colliders are

not in \mathbf{Q} . X and Y are d-separated given \mathbf{Z} if no undirected path between X and Y is active given \mathbf{Z} . In figure 15, for example, X and Y are not d-separated given U , but they are d-separated given $\{V, Z\}$.

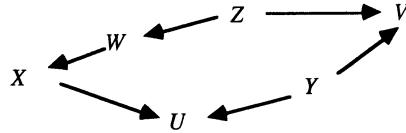


Figure 15

In the first case, conditioning on U activates the $X \rightarrow U \leftarrow Y$ path, so X and Y are not d-separated given U . The $X \leftarrow W \leftarrow Z \rightarrow V \leftarrow Y$ path is inactive given U , but X and Y are not d-separated given U as long as there is one undirected path between X and Y that is active given U . In the second case, conditioning on $\{V, Z\}$ activates V on the $X \leftarrow W \leftarrow Z \rightarrow V \leftarrow Y$ path, but conditioning inactivates Z on this path and thus makes it inactive; the $X \rightarrow U \leftarrow Y$ path is also inactive given $\{V, Z\}$ because U is a collider on the path that is not in $\{V, Z\}$. Because all of the undirected paths between X and Y are inactive given $\{V, Z\}$, X and Y are d-separated given $\{V, Z\}$.

Unfortunately, the full story is not quite so simple. Conditioning on a collider activates it, and so does conditioning on any of its descendants. In the graph in figure 16, for example, X and Y are not d-separated given W , because W is a descendant of U .

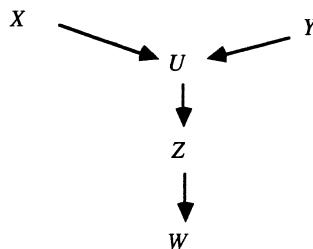


Figure 16

For a directed acyclic graph G , if X and Y are vertices in G , $X \neq Y$, and \mathbf{W} is a set of vertices in G not containing X or Y , then X and Y are **d-separated** given \mathbf{W} in G if and only if there exists no undirected path U between X and Y , such that (i) every collider on U has a descendent

in \mathbf{W} and (ii) no other vertex on U is in \mathbf{W} . We say that if $X \neq Y$, and X and Y are not in \mathbf{W} , then X and Y are **d-connected** given set \mathbf{W} if and only if they are not d-separated with respect to \mathbf{W} . If U , V , and \mathbf{W} are disjoint sets of vertices in G and U and V are not empty then we say that U and V are **d-separated** given \mathbf{W} if and only if every pair $\langle U, V \rangle$ in the cartesian product of U and V is d-separated given \mathbf{W} . If U , V , and \mathbf{W} are disjoint sets of vertices in G and U and V are not empty then we say that U and V are **d-connected** given \mathbf{W} if and only if U and V are not d-separated given \mathbf{W} .

The essential results are the following:

Theorem 3.3: $P(\mathbf{V})$ is faithful to directed acyclic graph G with vertex set \mathbf{V} if and only if for all disjoint sets of vertices \mathbf{X} , \mathbf{Y} , and \mathbf{Z} , \mathbf{X} and \mathbf{Y} are independent conditional on \mathbf{Z} if and only if \mathbf{X} and \mathbf{Y} are d-separated given \mathbf{Z} .

Theorem 3.4 provides a slightly more intuitive characterization of faithfulness, which motivates algorithms developed in Chapter 5.

Theorem 3.4: If $P(\mathbf{V})$ is faithful to some directed acyclic graph, then $P(\mathbf{V})$ is faithful to directed acyclic graph G with vertex set \mathbf{V} if and only if

- (i) for all vertices X , Y of G , X and Y are adjacent if and only if X and Y are dependent conditional on every set of vertices of G that does not include X or Y ; and
- (ii) for all vertices X , Y , Z such that X is adjacent to Y and Y is adjacent to Z and X and Z are not adjacent, $X \rightarrow Y \leftarrow Z$ is a subgraph of G if and only if X , Z are dependent conditional on every set containing Y but not X or Z .

The study of correlation is historically tied to the normal distribution, and for that distribution vanishing partial correlations and conditional independence are equivalent. But the Markov and Faithfulness Conditions tie vanishing correlation and partial correlation to graphical and causal structure for linear systems, without any normality assumption. Thus for linear systems, correlational structure is a guide to causal structure. We will say that a distribution P is **linearly faithful** to a graph G if and only if for vertices A and B of G and all subset C of the vertices of G , A and B are d-separated given C if and only $\rho_{AB|C} = 0$.

Theorem 3.5: If G is a directed acyclic graph with vertex set \mathbf{V} , A and B are in \mathbf{V} , and \mathbf{H} is included in \mathbf{V} , then G linearly implies $\rho_{AB|\mathbf{H}} = 0$ if and only A and B are d-separated given \mathbf{H} .

It follows that a distribution P is linearly faithful to a graph G if and only if for vertices A and B of G and all subsets C of the vertices of G , A and B are d-separated given C if and only $\rho_{AB|C} = 0$. Theorem 3.5 is the general principle behind all of the path analysis examples (Wright, 1934; Simon, 1954; Blalock, 1961; Heise, 1975) connecting causal structure in "recursive" (i.e., acyclic) linear models with vanishing partial correlations.

In the chapters that follow we will frequently remark that some conditional independence or conditional dependence relation, or vanishing or non-vanishing partial correlation, follows from a causal structure, assuming the distribution is faithful. Conversely, we will often observe that given certain conditional independence and dependence relations, or partial correlation facts, the causal structure must have certain properties if the distribution is faithful. Whenever we make such claims, we are using tacit corollaries of Theorems 3.3, 3.4 and 3.5.

3.7.2 The Manipulation Theorem

The fundamental aim of many empirical studies is to predict the effects of changes, whether the changes come about naturally or are imposed by deliberate policy. How can an observed distribution P be used to obtain reliable predictions of the effects of alternative policies that would impose a new marginal distribution on some set of variables? The very idea of imposing a policy that would directly change the distribution of some variable (e.g., drug use) necessitates that the resulting distribution P_{Man} will be different from P . P alone cannot be used to predict P_{Man} , but P and the causal structure can be.

Suppose that the Surgeon General is considering discouraging smoking, and he asks "What would the distribution of *Cancer* be if no one in the U.S. were allowed to smoke?" Let $V = \{Drinking, Smoking, and Cancer\}$. For the purpose of illustration assume that in the actual population in the U.S. the causal structure shown in figure 17 is correct.

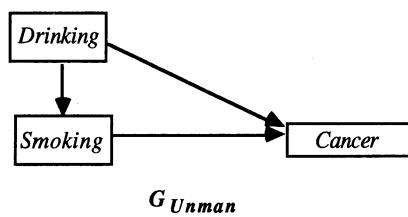


Figure 17

Let us call the population actually sampled (or produced by sampling and some experimental procedure) the **unmanipulated** population, and the hypothetical population for which smoking is banned the **manipulated** population. Suppose that if the policy of banning smoking were put into effect it would be completely effective, stopping everyone from smoking, but would not affect the value of *Drinking* in the population. Then the causal graph for the hypothetical manipulated population will be different than for the unmanipulated population, and the distribution of *Smoking* is different in the two populations. The manipulated causal graph is shown in figure 18.

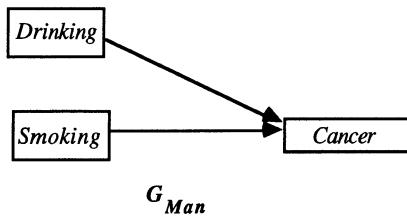


Figure 18

The difference between the unmanipulated graph and the manipulated graph is that some vertices that are parents of the manipulated variables in G_{Unman} may not (depending upon the precise form of the manipulation) be parents of manipulated variables in G_{Man} and vice-versa.

How can we describe the change in the distribution of *Smoking* that will result from banning smoking? One way is to note that the value of a variable that represents the policy of the federal government is different in the two populations. So we could introduce another variable into the causal graph, the *Ban Smoking* variable, which is a cause of *Smoking*. The full causal graph, including the new variable representing smoking policy, is then shown in figure 19. In the actual unmanipulated population the *Ban Smoking* variable is *off*, and in the hypothetical population the *Ban Smoking* variable is *on*. In the actual population we measure $P(Smoking|Ban\ Smoking = off)$; in the hypothetical population that would be produced if smoking were banned $P(Smoking = 0 | Ban\ Smoking = on) = 1$. For any subset X of $V = \{Smoking, Drinking, Cancer\}$ in the causal graph, let $P_{Unman}(Ban\ Smoking)(X)$ be $P(X|Ban\ Smoking = off)$ and let $P_{Man}(Ban\ Smoking)(V)$ be $P(V|Ban\ Smoking = on)$.

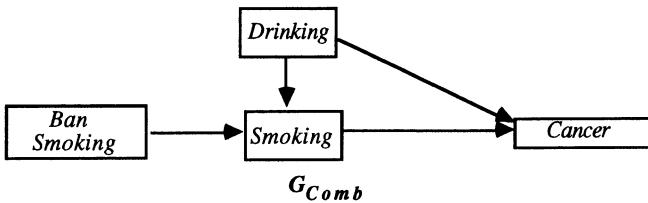


Figure 19

We can now ask if $P_{Unman}(Ban\ Smoking)(Cancer|Smoking) = P_{Man}(Ban\ Smoking)(Cancer|Smoking)$ (for those values of *Smoking* for which $P_{Man}(Ban\ Smoking)(Cancer|Smoking)$ is defined, namely $Smoking = 0$)? Clearly the answer is affirmative exactly when *Cancer* and *Ban Smoking* are independent given *Smoking*; but if the distribution is faithful this just reduces to the question of whether *Cancer* and *Ban Smoking* are d-separated given *Smoking*, which they are not in this causal graph. Further, $P_{Unman}(Ban\ Smoking)(Cancer) \neq P_{Man}(Ban\ Smoking)(Cancer)$ because *Cancer* is not d-separated from *Ban Smoking* given the empty set. But in contrast $P_{Unman}(Ban\ Smoking)(Cancer|Smoking, Drinking) = P_{Man}(Ban\ Smoking)(Cancer|Smoking, Drinking)$ (for those values of *Smoking* for which $P_{Man}(Ban\ Smoking)(Cancer|Smoking, Drinking)$ is defined, namely $Smoking = 0$), because *Ban Smoking* and *Cancer* are d-separated by $\{Smoking, Drinking\}$. The importance of this invariance is that we can predict the distribution of cancer if smoking is banned by considering the conditional distribution of cancer given drinking in the observed subpopulation of non-smokers, and by considering the distribution of drinking in the unmanipulated population.

Note that one of the *inputs* to our conclusion about $P_{Man}(Ban\ Smoking)(Cancer)$ is that the ban on smoking is completely successful and that it does not affect *Drinking*; this knowledge does not come from the measurements that we have made on *Smoking*, *Drinking* and *Cancer*, but is assumed to come from some other source. Of course, if the assumption is incorrect, there is no guarantee that our calculation of $P_{Man}(Ban\ Smoking)(Cancer)$ will yield the correct result. If we had instead considered a policy that does not effectively ban smoking, but intervenes to make smoking less likely without affecting drinking, then the graph of the entire system including the manipulation variable *Ban Smoking*, would be the same as in figure 19, and the graph G_{Unman} would be as in figure 17, but the manipulated graph G_{Man} would look like figure 17 rather than 18. Intervention would not remove the influence of drinking on smoking.

The analysis of prediction for a system involves three distinct graphs: a causal graph G_{Comb} which includes variables \mathbf{W} representing manipulations, and a causal graph G_{Unman} which is

the subgraph of G_{Comb} over a set of variables V not including the variables representing manipulations, and a graph G_{Man} over V which represents the causal relations among variables in V that result from a manipulation. G_{Man} may be a subgraph of G_{Unman} if the manipulation "breaks" causal dependencies in G_{Unman} ; otherwise G_{Unman} and G_{Man} will be the same graph.

Here are the formal definitions: If G is a directed acyclic graph over a set of variables $V \cup W$, and $V \cap W = \emptyset$, then W is **exogenous with respect to V** in G if and only if there is no directed edge from any member of V to any member of W . If G_{Comb} is a directed acyclic graph over a set of variables $V \cup W$, and $P(V \cup W)$ satisfies the Markov condition for G_{Comb} , then changing the value of W from w_1 to w_2 is a **manipulation** of G_{Comb} with respect to V if and only if W is exogenous with respect to V , and $P(V|W = w_1) \neq P(V|W = w_2)$.

We define $P_{Unman}(W)(V) = P(V|W = w_1)$, and $P_{Man}(W)(V) = P(V|W = w_2)$, and similarly for various marginal and conditional distributions formed from $P(V)$.

We refer to G_{Comb} as the **combined graph**, and the subgraph of G_{Comb} over V as the **unmanipulated graph** G_{Unman} . (Note that while $P_{Unman}(W)(V)$ satisfies the Markov Condition for G_{Unman} , it may also satisfy the Markov Condition for a subgraph of G_{Unman} . This is because G_{Comb} , and hence its subgraph G_{Unman} , may contain edges that are needed to represent the distribution of the manipulated subpopulation but not needed to represent the distribution of the unmanipulated subpopulation.)

V is in **Manipulated(W)** (that is, V is a variable directly influenced by one of the manipulation variables) if and only if V is in **Children(W) $\cap V$** ; we will also say that the variables in **Manipulated(W)** have been **directly manipulated**. We will refer to the variables in W as **policy variables**.

The **manipulated graph**, G_{Man} is a subgraph of G_{Unman} for which $P_{Man}(W)(V)$ satisfies the Markov Condition and which differs from G_{Unman} in at most the parents of members of **Manipulated(W)**. Exactly which subgraph G_{Man} is depends upon the details of the manipulation and what the causal graph of the subpopulation where $W = w_2$ is. For example, if smoking is banned, then G_{Man} contains no edge between income and smoking. On the other hand, if taxes are raised on cigarettes, G_{Man} does contain an edge between income and smoking. We will prove (in Chapter 13) that given a manipulation as defined, there always exists a subgraph of G_{Unman} for which $P_{Man}(W)(V)$ satisfies the Markov Condition. All of our theorems about manipulations hold for any G_{Man} that is a subgraph of G_{Unman} for which

$P_{Man(W)}(V)$ satisfies the Markov Condition, and which differs from G_{Unman} in at most the parents of members of $\text{Manipulated}(W)$.

These definitions entail the **Manipulation Theorem**:

Theorem 3.6: (Manipulation Theorem): Given directed acyclic graph G_{Comb} over vertex set $V \cup W$ and distribution $P(V \cup W)$ that satisfies the Markov condition for G_{Comb} , if changing the value of W from w_1 to w_2 is a manipulation of G_{Comb} with respect to V , G_{Unman} is the unmanipulated graph, G_{Man} is the manipulated graph, and

$$P_{Unman(W)}(V) = \prod_{X \in V} P_{Unman(W)}(X | \text{Parents}(G_{Unman}, X))$$

for all values of V for which the conditional distributions are defined, then

$$\begin{aligned} P_{Man(W)}(V) = \\ \prod_{X \in \text{Manipulated}(W)} P_{Man(W)}(X | \text{Parents}(G_{Man}, X)) \times \\ \prod_{X \in V \setminus \text{Manipulated}(W)} P_{Unman(W)}(X | \text{Parents}(G_{Unman}, X)) \end{aligned}$$

for all values of V for which each of the conditional distributions is defined.

The importance of this theorem is that if the causal structure and the direct effects of the manipulation (i.e. $P_{Man(W)}(X | \text{Parents}(X))$ for each X in $\text{Manipulated}(W)$) are known, then the joint distribution can be estimated from the unmanipulated population.

The Manipulation Theorem is not applicable when a causal mechanism between a pair of variables is reversible, in which case there can be two subpopulations in which the direction of the causal relationship between a pair of variables is reversed. For example, the movement of a motor of a car may cause the wheels to turn (as when the gas pedal is pressed), but also the turning of the wheels can cause the motor to move (as when the car rolls downhill).¹⁰ An intervention in a causal system which reverses the direction of some causal relationship is not a manipulation in our technical sense because there is no one combined graph representing the causal relations in the combined population. We are not suggesting any non-experimental

¹⁰We thank Marek Drzadzel for suggesting this example, and pointing out the problem of reversible mechanisms to us.

methods for determining whether a given mechanism is reversible. In some cases, such as smoking and yellow fingers, it is obvious from background knowledge that the mechanism is not reversible, because yellow fingers cannot cause smoking. In other cases, the relevant background knowledge may not be available, in which case it is not known whether the Manipulation Theorem is applicable.

Rubin (1977; 1978), and following him Pratt and Schlaifer (1988), have offered rules for when conditional probabilities in an observed population of systems will equal conditional probabilities for the same variables if the population is altered by a direct manipulation of some variables for all population units. We will show in Chapter 7 that their various rules are direct consequences of the special case of the Manipulation Theorem, illustrated in the discussion of figures 17, 18 and 19, in which one variable is manipulated and the intervention makes that variable independent of its causes in the unmanipulated graph.

Because the Manipulation Theorem is a consequence of the Markov condition, it requires no separate justification. Although the Manipulation Theorem is abstract, it is just the general formulation of inferences that are routine, if not always correct. When, for example, a regression model is used to predict the effects of a policy that would force values on some of the regressors, we have an application of the Manipulation Theorem. Of course the prediction may be incorrect if the causal or statistical assumptions of the regression model are false, or if the changes actually carried out do not satisfy the conditions for a manipulation. There are striking examples of both sorts of failure. Application of the Manipulation Theorem may give misleading predictions if the values of variables for each unit depend on the values of other units and if that dependency is not represented in the causal graph. Some public policy debates illustrate absurd violations of this requirement. Recently a research institute funded by automobile insurers carried out a non-linear regression of the rate of fatalities of occupants of various kinds of cars against car length, weight and other variables, finding unsurprisingly that the smaller the car the higher the fatality rate. This statistical analysis was then used by others to argue that proposed federal policies to downsize the American automobile fleet would increase highway fatalities. But of course the fatality rate in cars of a given size depends on the distribution of sizes of other cars in the fleet.

One can mistake which variables will be directly affected by a policy or intervention. Tacit applications of the Manipulation Theorem in such cases can lead to disappointment. As we will see in a later chapter, the literature on smoking, lung cancer and mortality provides vivid examples of predictions that went wrong, arguably because of misjudgments as to which variables would be directly manipulated by an intervention.

There is no reason why every intervention to deliberately alter the distribution of values of a set of variables V among units in a population (or sample) *must* satisfy the conditions for a direct manipulation of V and no others. But one of the chief aims in the design of experiments is to see to it that experimental manipulations are in fact direct manipulations of the intended variables and no others. The point of blind and double blind designs, for example, is exactly to obtain in experiment a direct manipulation of only the treatment variables. The concern with chronic wounding in drug trials with animals is essentially a worry that with respect to the outcome variables of interest, outcome variables as well as pharmacological variables have been directly manipulated. Typically, when we mistake the variables an intervention will directly manipulate, predictions of the outcomes of intervention will fail.

Our discussion in this section has assumed that the causal structure of the system is fully known. In Chapters 6 and 7 we will consider when and how the effects of interventions can be predicted from an unmanipulated distribution, assuming the distribution is the marginal over the measured variables of a distribution faithful to an unknown causal graph, and assuming the intervention constitutes a direct manipulation in the sense we have defined here.

3.8 Determinism

Another way that the Faithfulness Condition can be violated is when there are deterministic relationships between variables. In this section, we will give some rules for determining what extra conditional independence relations are entailed by deterministic relationships among variables.

We will say that a set of variables Z **determines** the set of variables A , when every variable in A is a deterministic function of the variables in Z , and not every variable in A is a deterministic function of any proper subset of Z . When there are deterministic relationships among variables in a graph, there are conditional independencies that are entailed by the deterministic relationships and the Markov condition that are not entailed by the Markov condition alone. For example, if G is a directed acyclic graph over V , V contains Z and A , and Z determines A , then A is independent of $V \setminus (Z \cup \{A\})$ given Z . If Z is a proper subset of the parents of A then this entails that A is independent of its other parents given Z , and also independent of its descendants as well as its non-descendants given Z . But it could also be the case that the

members of \mathbf{Z} are children of A , in which case given its children, A is independent of all other variables including its parents. It is also possible that \mathbf{Z} could contain non-parental ancestors of A . Each of these cases entails conditional independence relations not entailed by the Markov condition alone. For example, consider the graph in figure 20.

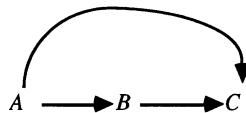


Figure 20

No conditional independence relations among A , B , or C are entailed by the Markov Condition alone. However, if the grandparent A determines the grandchild C , then $C \perp\!\!\!\perp B|A$. If the parent B determines the child C then $C \perp\!\!\!\perp A|B$. If the child C determines the parent B then $B \perp\!\!\!\perp A|C$.

Hence d-separability relations do not capture all of the conditional independencies entailed by the Markov condition and a set of deterministic relations. We will look for a graphical condition which entails the conditional independence of variables given the Markov condition and a set of deterministic relations among variables.

Geiger has proposed a simple, provably complete rule for graphically determining the conditional independencies entailed by the Markov and Minimality conditions and one kind of deterministic relationship among variables. Following Geiger (1990), in a directed acyclic graph G over \mathbf{V} that includes A and \mathbf{Z} , say that vertex A is a **deterministic variable** if it is a deterministic function of its parents in G . (Note that if a variable A has no parents in G , but has a constant value, then A is a deterministic variable.) A is **functionally determined** by \mathbf{Z} if and only if A is in \mathbf{Z} , or A is a deterministic variable and all of its parents are functionally determined by \mathbf{Z} . If \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are three disjoint subsets of variables in \mathbf{V} , \mathbf{X} and \mathbf{Y} are **D-separated** given \mathbf{Z} if and only if there is no undirected path U between any member of \mathbf{X} and any member of \mathbf{Y} such that each collider has a descendant in \mathbf{Z} and no other variable on U is functionally determined by \mathbf{Z} . Geiger has shown that \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} if and only if for every distribution that satisfies the Markov and Minimality Conditions for G , and the deterministic relations, \mathbf{X} and \mathbf{Y} are independent given \mathbf{Z} . We will prove that Geiger's rule is correct for a much wider class of deterministic relations; we do not know if it is complete for this wider class of deterministic relationships.

Suppose G is a directed acyclic graph over \mathbf{V} , and **Deterministic(V)** is a set of ordered tuples of variables in \mathbf{V} , where for each tuple D in **Deterministic(V)**, if D is $\langle V_1, \dots, V_n \rangle$ then V_n is a deterministic function of V_1, \dots, V_{n-1} and is not a deterministic function of any subset of V_1, \dots, V_{n-1} ; we also say $\{V_1, \dots, V_{n-1}\}$ determines V_n . Note that V_n could be an ancestor in G of members of V_1, \dots, V_{n-1} . Also, if A determines B and B determines A then **Deterministic(V)** contains both $\langle A, B \rangle$ and $\langle B, A \rangle$. We assume that **Deterministic(V)** is complete in the sense that if it entails some deterministic relationships among variables, those deterministic relations are in **Deterministic(V)**. (For example, if A determines B and B determines C , then A determines C .) **Det(Z)** is the set of variables determined by some subset of \mathbf{Z} . If a variable A has a constant value, then we say that it is determined by the empty set, and is in **Det(Z)** for all \mathbf{Z} .

Note that **Deterministic(V)** can entail dependencies between variables as well as independencies. If Z determines A , and Z is a member of \mathbf{Z} , then A is dependent on $\mathbf{Z} \setminus \{Z\}$ given Z . (Other dependencies may be entailed by **Deterministic(V)** as well.) These dependencies may conflict with independencies entailed by satisfying the Markov Condition for a directed acyclic graph G , so not every **Deterministic(V)** is compatible with every directed acyclic graph with vertex set \mathbf{V} . If **Deterministic(V)** and directed acyclic graph G are incompatible, Theorem 3.7 stated below is vacuously true, but obviously it would be desirable to have a test for determining whether **Deterministic(V)** and G are compatible.

We will expand Geiger's concept of D-separability so that it is not limited to the kind of deterministic relations that he considers. If G is a directed acyclic graph with vertex set \mathbf{V} , \mathbf{Z} is a set of vertices not containing X or Y , $X \neq Y$, then X and Y are **D-separated** given \mathbf{Z} and **Deterministic(V)** if and only if there is no undirected path U in G between X and Y such that each collider on U has a descendant in \mathbf{Z} , and no other vertex on U is in **Det(Z)**; otherwise if $X \neq Y$ and X and Y are not in \mathbf{Z} , then X and Y are **D-connected** given \mathbf{Z} and **Deterministic(V)**. Similarly, if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint sets of variables, and \mathbf{X} and \mathbf{Y} are non-empty, then \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} and **Deterministic(V)** if and only if each pair $\langle X, Y \rangle$ in the Cartesian product of \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} and **Deterministic(V)**; otherwise if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint, and \mathbf{X} and \mathbf{Y} are non-empty, then \mathbf{X} and \mathbf{Y} are D-connected given \mathbf{Z} and **Deterministic(V)**.

Theorem 3.7: If G is a directed acyclic graph over \mathbf{V} , \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint subsets of \mathbf{V} , and $P(\mathbf{V})$ satisfies the Markov condition for G and the deterministic relations in **Deterministic(V)** then if \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} and **Deterministic(V)**, \mathbf{X} and \mathbf{Y} are independent given \mathbf{Z} in P .

For example, suppose G is the graph in figure 21, and $\text{Deterministic}(V) = \{\langle A,B \rangle, \langle B,C \rangle, \langle A,C \rangle\}$.

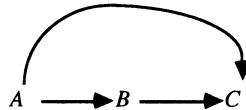


Figure 21

B and C are D-separated given A and $\text{Deterministic}(V)$, and A and C are D-separated given B and $\text{Deterministic}(V)$.

Suppose that G is still the graph in figure 21, but now $\text{Deterministic}(V) = \{\langle A,B \rangle, \langle B,A \rangle, \langle B,C \rangle, \langle C,B \rangle, \langle A,C \rangle, \langle C,A \rangle\}$. In addition to the previous D-separability relations, now A and B are D-separated given C and $\text{Deterministic}(V)$ because C determines A .

In some cases, conditional independencies are entailed because a parent is determined by its child. Consider the graph in figure 22, where $\text{Deterministic}(V) = \{\langle Y,W,Z \rangle, \langle Z,Y \rangle, \langle Z,W \rangle\}$. X and T are D-separated given Z and $\text{Deterministic}(V)$ because Z determines Y and W , and Y and W are non-colliders on the only undirected path between X and T .

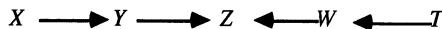


Figure 22

Finally, we note that it is possible that some non-parental ancestor X of Z determines Z , even though X does not determine any of the parents of Z . Let G be the graph in figure 23 and $\text{Deterministic}(V) = \{\langle X,Z \rangle\}$. Suppose X , R , and Z each have two values, and Y has four values. Consider the following distribution (where we give the probability of each variable conditional on its parents):

$$\begin{aligned} P(X = 0) &= .2 \\ P(R = 0) &= .3 \\ P(Y = 0|X = 0, R = 0) &= 1 \\ P(Y = 1|X = 0, R = 1) &= 1 \end{aligned}$$

$$P(Y = 2|X = 1, R = 0) = 1$$

$$P(Y = 3|X = 1, R = 1) = 1$$

$$P(Z = 0|Y = 0) = 1$$

$$P(Z = 0|Y = 1) = 1$$

$$P(Z = 1|Y = 2) = 1$$

$$P(Z = 1|Y = 3) = 1$$

In effect Y encodes the values of both R and X , and Z decodes Y to match the value of X .

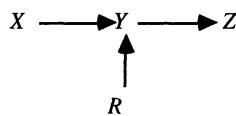


Figure 23

It follows that Y and Z are D-separated given X and **Deterministic(V)**, and X and Z are D-separated given Y and **Deterministic(V)**.

The following example points up an interesting difference between the set of distributions that satisfy the Markov condition for a given directed acyclic graph G , and the set of distributions that satisfy the Markov condition and a set of deterministic relationships among the variables in G . Suppose G is the graph shown in figure 24. For any directed acyclic graph, the set of probability distributions that satisfy the Markov condition for the graph includes some distributions that also satisfy the Minimality Condition for the graph. Suppose however, that **Deterministic(V) = {<X,Y>}**. In this case, among the distributions that satisfy the Markov Condition *and* the specified deterministic relations, there is no distribution that also satisfies the Minimality Condition. All distributions that satisfy the Markov Condition and the specified deterministic relation are faithful to the subgraph of figure 24 that does not contain the $Z \rightarrow Y$ edge. This suggests that to find all of the conditional independence relations entailed by satisfying the Markov Condition for a directed acyclic graph G and a set of deterministic relations, one would need to test for D-separability in various subgraphs G' of G with vertex set V in which for each Y in V no subset of **Parents**(G', Y) determines Y .



Figure 24

We will not consider algorithms for constructing causal graphs when such deterministic relations obtain, nor will we consider tests for deciding whether a set of variables \mathbf{X} determines a variable Y .

3.9 Background Notes

The ambiguous use of hypotheses to represent both causal and statistical constraints is nearly as old as statistics. In modern form the use of the idea by Spearman (1904) early in the century might be taken to mark the origins of statistical psychometrics. Directed graphs at once representing both statistical hypotheses and causal claims were introduced by Sewell Wright (1934) and have been used ever since, especially in connection with linear models. For a number of particular graphs the connections between linear models and partial correlation constraints were described by Simon (1954) and by Blalock (1961) for theories without unmeasured common causes, and by Costner (1971) and Lazarsfeld and Henry (1968) for theories with latent variables, but no general characterization emerged. A distribution-free connection of graphical structure for linear models with partial correlation was developed in Glymour, Scheines, Spirtes and Kelly (1987), for first order partials only, but included cyclic graphs. Geiger and Pearl (1989a) showed that for any directed acyclic graph there exists a faithful distribution. The general characterization given here as Theorem 3.5 is due to Spirtes (1989), but the connection between the Markov Condition, linearity and partial correlation seems to have been understood already by Simon and Blalock and is explicit in Kiiveri and Speed (1982). The Manipulability Theorem has been used tacitly innumerable times in experimental design and in the analysis of shocks in econometrics but seems never to have previously been explicitly formulated. A special case of it was first given in Spirtes, Glymour, Scheines, Meek, Fienberg and Slate (1991). The Minimality Condition and the idea of d-separability are due to Pearl (1988), and the proof that d-separability determines the consequences of the Markov condition is due to Verma (1987), Pearl, and Geiger (1989a). A result entailing theorem 3.4 was stated by Pearl, Geiger, and Verma (1990). Theorem 3.4 was used as the basis for a causal inference algorithm in Spirtes, Glymour, and Scheines (1990c). D-separability is described in Geiger (1990).

Chapter 4

Statistical Indistinguishability

Without experimental manipulations, the resolving power of any possible method for inferring causal structure from statistical relationships is limited by statistical indistinguishability. If two causal structures can equally account for the same statistics, then no statistics can distinguish them. The notions of statistical indistinguishability for causal hypotheses vary with the restrictions one imposes on the connections between directed graphs representing causal structure and probabilities representing the associated joint distribution of the variables. If one requires only that the Markov and Minimality Conditions be satisfied, then two causal graphs will be indistinguishable if the same class of distributions satisfy those conditions for one of the graphs as for the other. A different statistical indistinguishability relation is obtained if one requires that distributions be faithful to graph structure; and still another is obtained if the distributions must be consistent with a linear structure, and so on. For each case of interest, the problem is to characterize the indistinguishability classes graph-theoretically, for only then will one have a general understanding of the causal structures that cannot be distinguished under the general assumptions connecting causal graphs and distributions.

There are a number of related considerations about the resolving power of any possible method of causal inference from statistical properties. Given axioms about the connections between graphs and distributions, what graph theoretic structure must two graphs share in order also to share *at least one* probability distribution satisfying the axioms? When, for example, do two distinct graphs admit one and the same distribution satisfying the Minimality and Markov Conditions? When do two distinct graphs admit one and the same distribution satisfying the Minimality and Markov Conditions for one and the Faithfulness and Markov Conditions for the other? Reversing the question, for any given probability distribution that satisfies the Markov and Minimality Conditions (or in addition the Faithfulness Condition) for *some* directed acyclic graph, what is the *set of all such graphs* consistent with the distribution and these conditions? Finally, there are relevant measure-theoretic questions. If procedures exist that will identify causal structure under a more restrictive assumption such as Faithfulness, but not always under

weaker assumptions such as the Markov and Minimality Conditions, how likely are the cases in which the procedures fail? Under various natural measures on sets of distributions, for example, what is the measure of the set of distributions that satisfy the Minimality and Markov Conditions for a graph but are not faithful to the graph?

These are fundamental questions about the limits of any possible inference procedure--whether human or computerized--from non-experimental data to structure. We will provide answers for many of these questions when the system of measured variables is causally sufficient. Statistical indistinguishability is less well understood when graphs can contain variables representing unmeasured common causes.

4.1 Strong Statistical Indistinguishability

Two directed acyclic graphs G, G' are **strongly statistically indistinguishable (s.s.i)** if and only if they have the same vertex set V and every distribution P on V satisfying the Minimality and Markov Conditions for G satisfies those conditions for G' , and vice-versa.

That two structures are s.s.i. of course does not mean that the causal structures are one and the same, or that the difference between them is undetectable by any means whatsoever. From the correlation of two variables, X and Y , one cannot distinguish whether X causes Y , Y causes X or there is a third common cause, Z . But these alternatives may be distinguished by experiment or, as we will see, by other means.

Strong statistical indistinguishability is characterized by a simple relationship, namely that two graphs have the same underlying undirected graph and the same collisions:

Theorem 4.1: Two directed acyclic graphs G_1, G_2 , are strongly statistically indistinguishable if and only if (i) they have the same vertex set V , (ii) vertices V_1 and V_2 are adjacent in G_1 if and only if they are adjacent in G_2 , and (iii) for every triple V_1, V_2, V_3 in V , the graph $V_1 \rightarrow V_2 \leftarrow V_3$ is a subgraph of G_1 if and only if it is a subgraph of G_2 .

Given an arbitrary directed acyclic graph G , the graphs s.s.i. from G are exactly those that can be obtained by any set of reversals of the directions of edges in G that preserves all collisions in

G . A decision as to whether or not two graphs are s.s.i. requires $O(n^3)$ computations, where n is the number of vertices.

In figure 1 graphs G_1 and G_2 are s.s.i., but G_1 and G_3 , and G_2 and G_3 are not s.s.i.

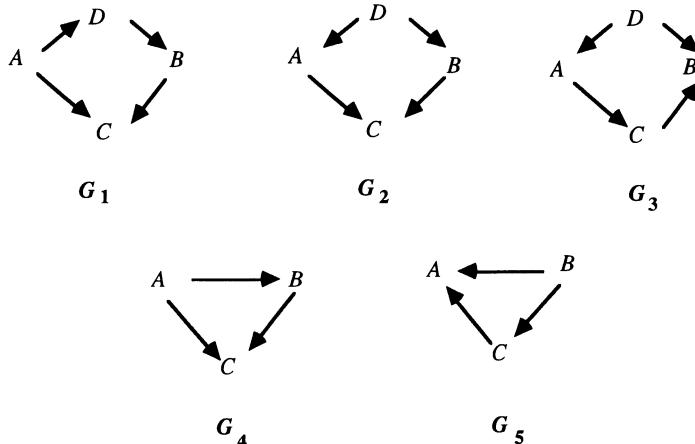


Figure 1

Note, however, if a set of variables \mathbf{V} is totally ordered, as for example by a known time order, and $P(\mathbf{V})$ is positive, then there is a *unique* graph for which $P(\mathbf{V})$ satisfies the Minimality and Markov conditions. (See Corollary 3 in Pearl 1988.)

4.2 Faithful Indistinguishability

Suppose we assume that all pairs $\langle G, P \rangle$ are faithful: all and only the conditional independence relations true in P are a consequence of the Markov condition for G . We will say that two directed acyclic graphs, G, G' are **faithfully indistinguishable** (f.i.) if and only if every distribution faithful to G is faithful to G' and vice-versa. The problem is to characterize faithful indistinguishability graphically.

Theorem 4.2: Two directed acyclic graphs G and H are faithfully indistinguishable if and only if (i) they have the same vertex set, (ii) any two vertices are adjacent in G if and only if they are adjacent in H , and (iii) any three vertices, X, Y, Z , such that X is adjacent to Y and Y is adjacent to Z but X is not adjacent to Z in G or H , are oriented as $X \rightarrow Y \leftarrow Z$ in G if and only if they are so oriented in H .

The question of faithful indistinguishability for two graphs can be decided in $O(n^3)$ where n is the number of vertices.

It is immediate from Theorems 4.1 and 4.2 that if two graphs are strongly statistically indistinguishable they are faithfully indistinguishable, but not necessarily conversely. The graphs G_4 and G_5 in figure 1 are not s.s.i. but they are f.i.

A class of f.i. graphs may be represented by a **pattern**. A pattern Π is a mixed graph with directed and undirected edges. A graph G is in the set of graphs **represented by** Π if and only if:

- (i) G has the same adjacency relations as Π ;
- (ii) if the edge between A and B is oriented $A \rightarrow B$ in Π , then it is oriented $A \rightarrow B$ in G ;
- (iii) if Y is an unshielded collider on the path $\langle X, Y, Z \rangle$ in G then Y is an unshielded collider on $\langle X, Y, Z \rangle$ in Π .

For example, the set of all complete, acyclic directed graphs on three vertices forms a faithful indistinguishability class that can be represented by a pattern consisting of the complete undirected graph on the same vertex set. When the pattern of the faithful indistinguishability class of a directed acyclic graph has no directed edges, and so is purely undirected, the statistical hypothesis represented by the directed graph is equivalent to the statistical hypothesis of the undirected independence graph corresponding to the pattern.

4.3 Weak Statistical Indistinguishability

The indistinguishability relations characterized in the two previous sections ask for the graphs that can accommodate the same class of probability distributions as a given graph. We can turn the tables, at least partly, by starting with a particular probability distribution on a set of variables and asking for the set of all directed acyclic graphs on those vertices that are consistent

with the given distributions. The answers characterize how much the probabilities and our assumptions about the connection between probabilities and causes underdetermine the causal structure. Assuming Markov and Minimality only, the equivalence of these two conditions (under positivity) with the defining conditions for a directed independence graph provides an (impractical) algorithm for generating the set of all graphs that satisfy the two conditions for a given distribution P . For every ordering of the variables in P there is a directed acyclic graph G compatible with that ordering (i.e. A precedes B in the ordering only if A is not a descendant of B in G) satisfying the Minimality and Markov Conditions for P . It can be generated by assuming the ordering and the conditional independence relations in P and applying the definition of directed independence graph. An algorithm that does not assume positivity is given by Pearl (1988). According to that algorithm let Ord be a total ordering of the variables, and $\text{Predecessors}(Ord, X)$ be the predecessors of X in the ordering Ord . For each variable X , let the parents of X in G be a smallest subset R of $\text{Predecessors}(Ord, X)$ such that X is independent of $\text{Predecessors}(Ord, X) \setminus R$ given R in P . It follows that P satisfies the Minimality and Markov Conditions for P .

The alternatives are more limited if we start with P and assume that any graph must be faithful to P . In that case all of the graphs faithful to P form a faithful indistinguishability class, i.e., the set of all graphs f.i. from any one graph faithful to P . The next chapter presents a number of algorithms that generate the faithful indistinguishability classes from properties of distributions.

Given axioms connecting causal graphs with probability distributions it makes sense to ask for which pairs G, G' of graphs there exists some probability distribution satisfying the axioms for both G and G' . Let us say that two graphs are **weakly faithfully indistinguishable** (w.f.i.) if and only if there exists a probability distribution faithful to both of them. We say that two graphs are **weakly statistically indistinguishable** (w.s.i.) if and only if there exists a probability distribution meeting the Minimality and Markov Conditions for both of them. Weak faithful indistinguishability proves to be equivalent to faithful indistinguishability:

Theorem 4.3: Two directed acyclic graphs are faithfully indistinguishable if and only if some distribution faithful to one is faithful to the other and conversely; i.e. they are f.i. if and only if they are w.f.i.

This theorem tells us that faithfulness divides the set of probability distributions over a vertex set into equivalence classes that exactly correspond to the equivalence classes of graphs induced by faithful indistinguishability. It follows that if a distribution is faithful to some graph G then it is faithful to all and only the graphs faithfully indistinguishable from G .

There is no reason to expect so nice a match in general. Suppose we assume only the Minimality and Markov Conditions. Under what conditions will there exist a distribution P satisfying those axioms for two distinct graphs, G , and G' ? The answer is *not*: exactly when G and G' are strongly statistically indistinguishable. The two graphs shown in figure 2 are not s.s.i., but there exist distributions that satisfy the Minimality and Markov Conditions for both:

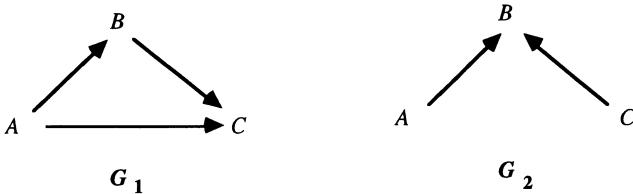


Figure 2

The distributions in Simpson's "paradox" provide an example, as we have already seen in Chapter 3. We conjecture that if a distribution satisfies the Minimality and Markov Conditions for two graphs G and G' , then G and G' have the same edges and the same colliders, save that triangles such as G_1 in one graph may be replaced by collisions such as G_2 in the other, provided appropriate conditions are met by other edges. We don't know how to characterize the "appropriate" conditions. There is, however, a related property of interest we can characterize.

No distribution that is faithful to graph G_1 in figure 2 can be faithful to graph G_2 , but a distribution that satisfies the Minimality and Markov Conditions for G_1 can be faithful to graph G_2 . Just when can this sort of thing happen? When, in other words, can the generalization of Simpson's "paradox" arise? If probability distribution P satisfies the Minimality and Markov Conditions for G , and P is faithful to graph H , what is the relation between G and H ?

Theorem 4.4: If probability distribution P satisfies the Markov Condition for directed acyclic graphs G and H , and P is faithful to H , then for all vertices X, Y , if X, Y are adjacent in H they are adjacent in G .

Theorem 4.5: If probability distribution P satisfies the Markov and Minimality Conditions for directed acyclic graphs G and H , and P is faithful to graph H , then (i) for all X, Y, Z such that $X \rightarrow Y \leftarrow Z$ is in H and X is not adjacent to Z in H , either $X \rightarrow Y \leftarrow Z$ in G or X, Z are adjacent in G and (ii) for every triple X, Y, Z of vertices such that

$X \rightarrow Y \leftarrow Z$ is in G and X is not adjacent to Z in G , if X is adjacent to Y in H and Y is adjacent to Z in H then $X \rightarrow Y \leftarrow Z$.

Corollary 4.1: If probability distribution P satisfies the Markov Condition for directed acyclic graph G , P is faithful to directed acyclic graph H , and G and H agree on an ordering of the variables (as, for example, by time) such that $X \rightarrow Y$ only if $X < Y$ in the order, then H is a subgraph of G .

4.4 Rigid Indistinguishability

In addition to the notions of strong, faithful and weak statistical indistinguishability, there is still another. Suppose two directed acyclic graphs, G and G' , are statistically indistinguishable in some sense over a common set \mathbf{O} of vertices. Then without experiment, no measurement of the variables in \mathbf{O} will reliably determine which of the graphs correctly describes the causal structure that generated the data. It might be, however, that G and G' can be distinguished if other variables besides those in G or G' are measured and stand in appropriate causal relations to the variables in \mathbf{O} . For example, the following simple graphs are both s.s.i. and f.i. (where A and B are assumed to be measured and in \mathbf{O} .)



Figure 3

But if we also measure a variable C that is a cause of A or has a common cause with A and no connection with B save possibly through A , then the two structures can be distinguished.

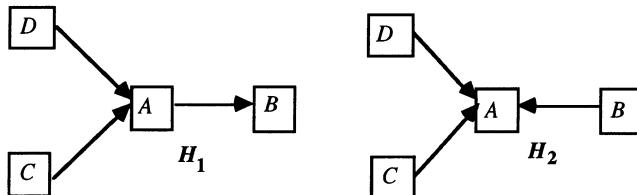


Figure 4

The graphs in figure 4 are not f.i. or s.s.i. It is equally easy to give examples of w.s.i. structures that can be embedded in graphs that are not w.s.i. Which causally sufficient structures can be distinguished by measuring extra variables? To answer the question we require some further definitions.

Let G_1, G_2 be two directed acyclic graphs with common vertex set \mathbf{O} . Let H_1, H_2 be directed graphs having a common set \mathbf{U} of vertices that includes \mathbf{O} and such that

- (i) the subgraph of H_1 over \mathbf{O} is G_1 and the subgraph of H_2 over \mathbf{O} is G_2 ;
- (ii) every directed edge in H_1 but not in G_1 is in H_2 and every directed edge in H_2 but not in G_2 is in H_1 .

We will say then that directed acyclic graphs G_1 and G_2 with common vertex set \mathbf{O} have a **parallel embedding** in H_1 and H_2 over \mathbf{O} and \mathbf{U} . In figures 3 and 4, G_1 and G_2 have a parallel embedding in H_1 and H_2 over $\mathbf{O} = \{A,B\}$ and $\mathbf{U} = \{A,B,C,D\}$. The question of whether two s.s.i. structures can be distinguished by measuring further variables then becomes the following: do the structures have parallel embeddings that are not s.s.i.? If no such embedding exists we will say the structures G_1 and G_2 are **rigidly statistically indistinguishable (r.s.i.)**.

Theorem 4.6: No two distinct s.s.i. directed acyclic graphs with the same vertex set are rigidly statistically indistinguishable.

In other words, provided additional variables with the right causal structure exist and can be measured, the causal structure among a causally sufficient collection of measured variables can in principle be identified. The proof of Theorem 4.6 also demonstrates that a parallel result for faithfully indistinguishable structures. We conjecture that an analog of Theorem 4.6 also holds for weak statistical indistinguishability assuming positivity.

4.5 The Linear Case

Parameter values can force conditional independencies or zero partial correlations that are not linearly implied by a graph. The graphs in figure 2 (reproduced in figure 5 with error variables explicitly included) illustrate the possibility: treat the vertices of the graphs as each attached to an "error" variable, and let the graphs plus error variables determine a set of linear equations. (We

assume that any pair of exogenous variables, including the error terms, have zero covariance.) The result is, up to specification of the joint distribution of the exogenous variables, a structural equation model. A linear coefficient is attached to each directed edge. The correlation matrix, and hence all partial correlations, is determined by the linear coefficients and the variances of the exogenous variables.

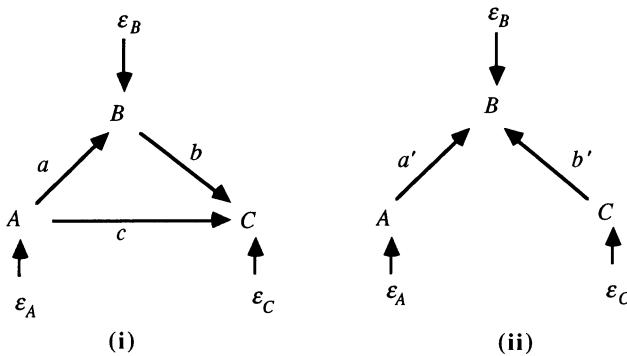


Figure 5

If in the structure on the left $ab = -c$, then A, C will be uncorrelated as in the model on the right. This sort of phenomenon--vanishing partial correlations produced by values of linear coefficients rather than by graphical structure--is bound to mislead any attempt to infer causal structure from correlations. When can it happen? We have already answered that question in the previous chapter, when we considered the conditions under which linear faithfulness might fail. In the linear case, the parameter values--values of the linear coefficients and exogenous variances of a structure with a directed acyclic graph G --form a real space, and the set of points in this space that create vanishing partial correlations not linearly implied by G have Lebesgue measure zero.

Theorem 3.2: Let M be a linear model with directed acyclic graph G and n linear coefficients a_1, \dots, a_n and k positive variances of exogenous variables v_1, \dots, v_k . Let $M(<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>)$ be the distributions consistent with specifying values $<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>$ for a_1, \dots, a_n and v_1, \dots, v_k . Let Π be the set of probability measures P on the space \Re^{n+k} of values of the parameters of M such that for every subset V of \Re^{n+k} having Lebesgue measure zero, $P(V) = 0$. Let Q be the set of vectors of coefficient and variance values such that for all q in Q every probability distribution

in with $M(q)$ has a vanishing partial correlation that is not linearly implied by G . Then for all P in Π , $P(\mathbf{Q}) = 0$.

Measure theoretic arguments of this sort are interesting but may not be entirely convincing. One could, after all, argue that in the general linear model absence of causal connection is marked by linear coefficients with the value zero, and thus form a set of measure zero, so by parity of reasoning everything is causally connected to everything else. In a recent book Nancy Cartwright (1989) objects that since in linear structures independence relations may be produced by special values of the linear coefficients and variances as well as by the causal structure, it is illegitimate to infer causal structure from such relations. In effect, she rejects any inference procedure that is unable to distinguish the true causal structure from w.s.i. alternatives. Such a position may be extreme, but it does serve to focus attention on two interesting questions: when is it impossible for two structures to be w.s.i. but not f.i. or s.s.i., and are there special marks or indicators that a distribution satisfies the Markov and Minimality conditions for two w.s.i. but not s.s.i. or f.i. causal structures? The answers to these questions are essentially just applications to the linear case of the theorems of the preceding sections.

We will assume, with Cartwright, that a time ordering of the variables is known. Pearl and Verma (Pearl 1988) have proved that for a positive distribution P and a given ordering of variables, there is only one directed acyclic graph for which P satisfies the Minimality and Markov Conditions. It follows that for a positive distribution with a given correlation matrix and a given ordering of a causally sufficient set of variables there is a unique directed acyclic graph that linearly represents the distribution and is consistent with the ordering.

In some cases at least, the positivity of a distribution can be tested for. (For example, in a bivariate normal distribution the density function is everywhere non-zero if the correlation is not equal to one.) It follows for those cases that for a given ordering of variables either there is a unique directed acyclic graph for which P satisfies the Markov and Minimality Conditions, or it is detectable that more than one such directed acyclic graph exists. However, even if for a given ordering of variables there is a unique directed acyclic graph for which P satisfies the Markov and Minimality Conditions, algorithms for finding that graph are not feasible for large numbers of variables, because of the number and order of the conditional independence relations that they require be tested.

Suppose that we wrongly assume that a distribution is faithful to the causal graph that generated it. Then Corollary 4.1 applies, which means, informally, that if faithfulness is assumed but not true, then conditional independence relations or vanishing partial correlations due to special

parameter values can only produce erroneous causal inferences in which a true causal connection is *omitted*; no other sorts of error may arise. We will consider when this circumstance is revealed in the correlations.

Recall that a **trek** is an unordered pair of directed acyclic paths having a single common vertex that is the source of both paths (one of the paths in a pair may be the empty path defined in Chapter 2). For standardized models, in which the mean of each variable is zero and non-error variables have unit variance, the correlation of two variables is given by the sum over all treks connecting X, Y of the product for each trek of the linear coefficients associated with the edges in that trek (we call this quantity the **trek sum**). For example, in directed acyclic graph (i) in figure 5, the trek sum between A and C is $ab + c$. We will use standardized systems throughout our examples in this section. The system of correlations determines all partial correlations of every order through the following formula.

$$\rho_{XY.Z \cup \{R\}} = \frac{\rho_{XY.Z} - \rho_{XR.Z} \times \rho_{YR.Z}}{\sqrt{1 - \rho_{XR.Z}^2} \times \sqrt{1 - \rho_{YR.Z}^2}}$$

Since the recursion relations give the same partial correlation between two variables on a set U no matter in what sequence the partials on the members of U are taken, a vanishing partial correlation corresponds to a system of equations in the coefficients of a standardized system.

Suppose now that special values of the linear parameters in a normal, standardized system G produce vanishing partial correlations that are exactly those linearly implied only by some false causal structure, say H . Then the parameter values must generate extra vanishing partial correlations not linearly implied by G . Any partial correlation is a function just of the trek sums connecting pairs of variables, and the trek sums in this case involve just the linear parameters in G . Hence each additional vanishing partial correlation not linearly implied by G determines a system of (non-linear) equations in the parameters of G that must be satisfied in order to produce the coincidental vanishing partial correlation. (For example, in directed acyclic graph (i) of figure 5, the correlation between A and C is 0 only if the single equation $ab = -c$ is satisfied). Now for some G and some H (a sub-graph of G), these systems of equations may have no simultaneous solution. In that case there are no values for the parameters of G that will produce partial correlations that are exactly those linearly implied by H . For other choices of G and a subgraph H , it may be that the system of equations has a solution, but only solutions that allow only a finite number of alternative values for one or more parameters and that require some error variance to vanish. Such a solution must "give itself away" by special correlation constraints

that are not themselves vanishing partial correlation relations. Consider the following choices of G and H , where in each pair G is on the left hand side and H is on the right hand side:

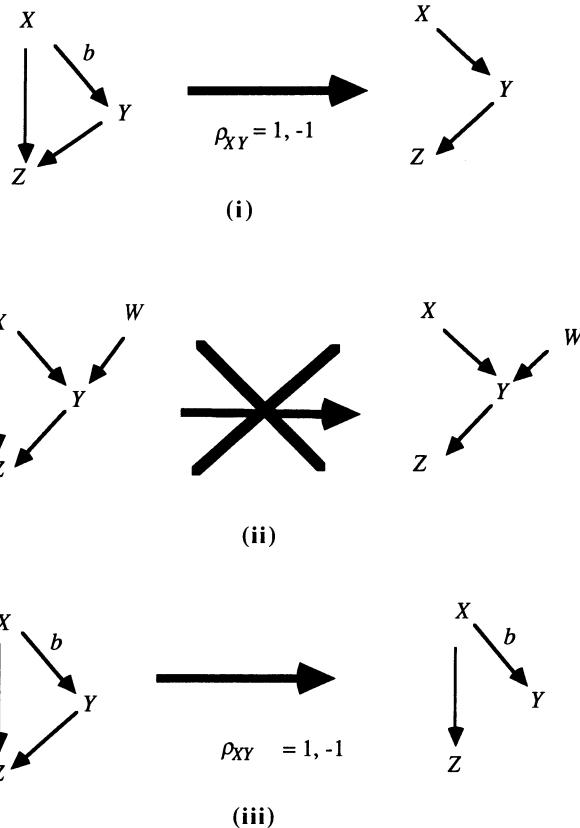


Figure 6

In (i) and (iii), coefficients and variances can be chosen for the graph on the left hand side so that it appears as though an edge does not occur, but only by making the coefficient labeled b equal to either 1 or -1. Since the variables are standardized, this requires that the error term for Y have zero variance and zero mean--i.e., it vanishes. Thus in order for the true graph to be the one on the left hand side and the parameter values to produce vanishing partial correlations that are exactly those linearly implied by the graph on the right hand side, variable Y must be a linear function of variable X and only variable X . The same result obtains if the edges that are not

eliminated in the first and last examples are replaced by directed paths of any length. Clearly in these cases special parameter values that create vanishing partial correlations not linearly implied by the true graph will be revealed by the correlations. In (ii) the edge between variables X and Z cannot be made to appear to be eliminated by any choice of parameter values for the true graph.

We conjecture that even without a prior time order, unless three edges form a triangle in G , if parameter values of G determine exactly the collection of vanishing partial correlations linearly implied by a graph H --whether or not H is a subgraph of G --then there are extra constraints on the correlations not entailed by the vanishing partial correlations.

4.6 Redefining Variables

The indistinguishability results so far considered relate alternative graphs over the same set of vertices. The vertices are interpreted as random variables whose values are subject to some system of measurement. New random variables can always be defined from a given set, for example by taking linear or Boolean combinations. For any specified apparatus of definitions, and any axioms connecting graphs with distributions, questions about indistinguishability classes arise parallel to those we have considered for fixed sets of variables. A distribution P over variable set V may correspond to a graph G , and a distribution P' over variable set V' may correspond to a different graph G' (with P' and V' obtained from P and V by defining new variables, ignoring old ones, and marginalizing). The differences between G and G' may in some cases be unimportant, and one may simply want to say that each graph correctly describes causal relations among its respective set of variables. That is not so, however, when the original variables are ordered by time, and redefinition of variables results in a distribution whose corresponding graphs have later events causing earlier events. Consider the following pair of graphs.

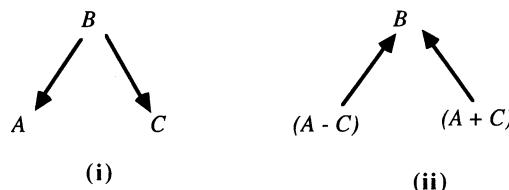


Figure 7

In directed acyclic graph (i), A and C are effects of B ; suppose that B occurs prior to A and C . By the procedure of definition and marginalization, a distribution faithful to graph (i) can be transformed into a distribution faithful to graph (ii). First, standardize A and C to form variables A' and C' with unit variance. Then consider the variables $(A' - C')$ and $(A' + C')$. Their covariance is equal to the expected value of $A'^2 - C'^2$ which is zero. Simple algebra shows that the partial correlation of $(A' - C')$ and $(A' + C')$ given B does not vanish. The marginal of the original distribution is therefore linearly faithful to (ii), and faithful to (ii) if the original distribution is normal.

Note that the transformation just illustrated is unstable; if the variances of A' and C' are unequal in the slightest, or if the transformation gives $(xA' + zC')$ and $(yA' + wC')$ for any values of x , y , z , and w such that $xy + wz + \rho_{A'C'}(yz + xw) \neq 0$ then the marginal on the transformed distribution will be faithful, not to (ii), but to all acyclic orientations of the complete graph on the three variables, a hypothesis that is not inconsistent with the time order.

Viewed from another perspective, a transformation of variables that produces a "coincidental" vanishing partial correlation is just another violation of the Faithfulness Condition. Consider the linear model in figure 8.

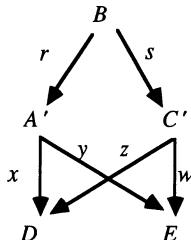


Figure 8

Let $A' = rB + \varepsilon_{A'}$, $C' = sB + \varepsilon_{C'}$, $D = xA' + zC' + \varepsilon_D$, and $E = yA' + wC' + \varepsilon_E$. If the variables are standardized, ρ_{DE} is equal to $xy + zw + rysz + rxsw = xy + zw + rs(yz + xw)$, which, since $rs = \rho_{A'C'}$, is the formula of the previous paragraph. If $\rho_{DE} = 0$, the Faithfulness Condition is violated. Hence the conditions under which we obtain a linear transformation of A and C that produces a "coincidental" zero correlation are identical to the conditions under which the treks between A and C exactly cancel each other in a violation of the Faithfulness Condition.

We get the example of figure 7 when $D = A' + C'$ (i.e. $x = z = 1$), and $E = A' - C'$ (i.e. $y = -w = 1$) where the variances and means of the error terms have been set to zero. Since the set of parameter values that violate Faithfulness in this example has Lebesgue measure zero, so does the set of linear transformations of A and C that produce a "coincidental" zero correlation.

4.7 Background Notes

The underdetermination of linear statistical models by values of measured variables has been extensively discussed as the "identification problem," especially in econometrics (Fisher, 1966) where the discussion has focused on the estimation of free parameters. The device of "instrumental variables," widely used for linear models, is in the spirit of Theorem 4.6 on rigid distinguishability, although instrumental variables are used to identify parameters in cyclic graphs or in systems with latent variables. The possibility of "rewriting" a pure linear regression model so that the outcome variable is treated as a cause seems to have been familiar for a long while, and we do not know the original source of the observation, which was brought to our attention by Judea Pearl.

Accounts of statistical indistinguishability in something like one or another of the senses investigated in this chapter have been proposed by Basman (1965), Stetzl (1986) and Lee (1987). Basman argued, in our terms, that for every simultaneous equation model with a cyclic graph (i.e., "non-recursive") there exists a statistically indistinguishable model with an acyclic graph. The result is a weak indistinguishability theorem (see Chapter 12). Stetzl and Lee focus exclusively on linear structural equation models with free parameters for linear coefficients and variances, and they define equivalence in terms of maximum likelihood estimates of the parameters and hence of the covariance matrix. No general graph theoretic characterizations are provided, although interesting attempts were made in Lee's thesis.

The notion of a pattern and Theorem 4.2 are due to Verma and Pearl (1990b). We state some results about indistinguishability relations for causally insufficient graphs in Chapter 6. A well-known result due to Suppes and Zanotti (1981) asserts that every joint distribution P on a set X of discrete variables is the marginal of some joint distribution P^* on $X \cup \{T\}$ satisfying the Markov Condition for a graph G in which T is the common cause of all variables in X and there are no other directed edges. The result can be viewed as a weak indistinguishability theorem

when causally insufficient structures are admitted. Except in special cases, P^* cannot be faithful to G .

Chapter 5

Discovery Algorithms for Causally Sufficient Structures

5.1 Discovery Problems

A discovery problem is composed of a set of alternative structures, one of which is the source of data, but any of which, for all the investigator knows before the inquiry, could be the structure from which the data are obtained. There is something to be found out about the actual structure, whichever it is. It may be that we want to settle a particular hypothesis that is true in some of the possible structures and false in others, or it may be that we want to know the complete theory of a certain sort of phenomenon. In this book, and in much of the social sciences and epidemiology, the alternative structures in a discovery problem are typically directed acyclic graphs paired with joint probability distributions on their vertices. We usually want to know something about the structure of the graph that represents causal influences, and we may also want to know about the distribution of values of variables in the graph for a given population.

A discovery problem also includes a characterization of a kind of evidence; for example, data may be available for some of the variables but not others, and the data may include the actual probability or conditional independence relations or, more realistically, simply the values of the variables for random samples. Our theoretical discussions will usually consider discovery problems in which the data include the true conditional independence relations among the measured variables, but our examples and applications will always involve inferences from statistical samples.

A method solves a discovery problem in the limit if as the sample size increases without bound the method converges to the true answer to the question or to the true theory, *whatever*

(consistent with prior knowledge) the truth might be. A procedure for inferring causes does not solve the problem posed if for some of the alternative possibilities it gives no answer or the wrong answer, although it may solve another, easier problem that arises when some of the alternative structures are excluded. Which causal discovery problems are solvable in the limit, and by what methods, are determinate, mathematical questions. The metaphysical wrangling lies entirely in motivating the problems, not in solving them. The remainder of this book is an introduction to the study of these formal questions and to the practical applications of particular answers.

5.2 Search Strategies in Statistics

The statistical literature is replete with procedures that use data to guide a search for some restricted parametrization of alternative distributions. When the representation of the statistical hypothesis is used to guide policy or practice, to predict what will happen if some of the variables are manipulated or to retrodict what would have happened if some of the variables had in the past been manipulated, then the statistical hypotheses are usually also causal hypotheses. In that case the first question is whether the search procedures are any good at finding causal structure.

Many of the search strategies proposed in the statistical literature are best-only beam searches, beginning either with an arbitrary model, or with a complete (or almost complete) structure that entails no constraints, or with a completely (or almost completely) constrained structure in which all variables are independent. Statisticians sometimes refer to the latter procedure as "forwards" search, and the former procedures as "backwards" search. Depending on which order is followed, the procedures iteratively apply a fit measure of some kind to determine which fixed parameter in the parametrization will most improve fit when freed--or which free parameter should be fixed. They then reestimate the modified structure to determine if a stopping criterion is satisfied. A "forward" procedure of this kind was proposed by Arthur Dempster (1972) for covariance structures, and a "backward" procedure was proposed by his student, Nanny Wermuth (1976), for both log-linear and linear systems whose distributions are "multiplicative"--in our terms, satisfy the Markov condition for some directed acyclic graph. Forward search algorithms using goodness of fit statistics were proposed for multinormal linear systems by Byron (1972) and by Sorbom (1975) and versions of them have been automated in the LISREL (Joreskog and Sorbom 1984) and EQS (Bentler 1985) estimation packages. The

latter program also contains a backwards search procedure. Versions of the general strategy for log-linear parametrizations are described by Bishop, Fienberg and Holland (1975), by Fienberg (1977) by Aitkin (1979), by Christensen (1990) and many others. The same representations and search strategies have been used in the systems science literature by Klir and Parviz (1976) and others under the title of "reconstructability" analysis. Stepwise regression procedures in logistic regression can be viewed as versions of the same strategies. The same strategies have been applied to undirected and directed graph representations. They are illustrated for a variety of examples by Whittaker (1990).

In each of these cases the general statistical search strategy is unsatisfactory if the goal is not just to estimate the distribution but also to identify the causal structure or to predict the results of manipulations of some of the variables. When used to these ends, these searches are inefficient and unreliable for at least three reasons: (i) they often search a hypothesis space that excludes many causal hypotheses and includes many hypotheses of no causal significance; (ii) the specifications of distributions typically force the use of numerical procedures that for statistical or computational reasons unnecessarily limit search; (iii) restrictions requiring the search to output a single hypothesis entail that the search fails to output alternative hypotheses that may be indistinguishable given the evidence. We will consider each of these points in more detail.

5.2.1 The Wrong Hypothesis Space

In searching for the correct causal hypothesis the space of alternatives should, insofar as possible, include all causal hypotheses that have not been ruled out by background knowledge and no hypotheses that do not have a causal interpretation. The log-linear formalism, introduced by Birch in 1963, provides an important example of a search space poorly adapted to the goal of finding correct causal hypotheses. For discrete data a more appropriate search space turns out to be a sub-class of conjunctions of log-linear hypotheses.

The log-linear formalism provides a general framework for the analysis of contingency tables of any dimension. In the discrete case we are concerned with variables that take a finite number of values, whether ordered or not. For a system with four variables, for example, we will let i range over the values of the first variable, j the second, k the third and l the fourth. In a particular sample or population, x_{ijkl} will then denote the number of units that have value i for the first variable, value j for the second variable, k for the third and l for the fourth. We will refer to a particular vector of values for the four (or other number of) variables as a "cell." In the formalism the joint distribution over the cells is given by an equation for the logarithm of the

expected value of each cell, expressed as the sum of a number of parameters. For example, in Birch's notation in which m_{ijk} denotes the expected number in cell i, j, k ,

$$\ln(m_{ijk}) = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}$$

The various u 's are arbitrary parameters with an associated set of indices; only seven of the u terms can be independent for a system of three binary variables. The power of Birch's parametrization lies in at least two features. First, associations in multi-dimensional contingency tables that had long been studied in statistics can be represented as hypotheses that certain of the parameters are zero. For example Bartlett's representation of the hypothesis of no "three factor interaction" among three binary variables is given by the following relation among the cell probabilities:

$$p_{111}p_{122}p_{212}p_{221} = p_{222}p_{211}p_{121}p_{112}$$

Birch shows that a generalization of this condition to variables of any finite number of categories obtains if and only if various of the u terms are zero. Second, for each hypothesis obtained by setting some of the u terms to zero, there exist iterative methods for obtaining maximum likelihood estimates for a variety of sampling procedures.

Birch's results were extended by several researchers. A hypothesis in the log-linear parametrization has come to be treated as a specification that particular u terms vanish. There are direct maximum likelihood estimates of the expected cell counts for certain forms of such specifications, and for other specifications iterative algorithms have been developed that converge to the maximum likelihood estimates. Various formal motivations have been developed for focusing on particular classes of log-linear parametrizations. Using his information-based distance measure, for example, Kullback (1959) derived a class of log-linear relations that could be obtained in the same way from a slightly different perspective, the maximum entropy principle. Fienberg (1977) and others have urged restricting attention to "hierarchical models"--log linear parametrizations in which if a u term with a set of indices is put to zero so are all other u terms whose indices contain the first set. The motivation for the restriction is that these parametrizations bear a formal analogy to analysis of variance, so that the u_1 term, for example, may be thought of as the variation from the grand mean due to the action of the first variable.

To see the difficulties in representing causal structure in the log-linear formalism, consider the most fundamental causal relation of the preceding chapters, namely any collider $A \rightarrow B \leftarrow C$ in

which A and C are not adjacent. Such a structure corresponds (assuming faithfulness) to two facts about conditional independence: first, A and C are independent conditional on some set of variables that does not contain B ; second, A and C are dependent conditional on every set that does contain B but not A or C . In the very simplest case of this kind, in which A , B and C are the only variables, A and B are independent, but dependent conditional on C . The hypothesis that these relations obtain cannot be expressed in the log-linear formalism by vanishing u terms. Birch himself observed that in a three variable system the hypothesis that in the marginal distribution two of the variables are independent cannot be expressed by the vanishing of any subset of parameters in the general log-linear expansion for the three variables. There are of course log-linear hypotheses that are *consistent* with marginal independence hypotheses, but do not entail them.

Another inappropriate search space is provided by the LISREL program. The LISREL formalism--at least as intended by its authors, Joreskog and Sorbom--allows search for structures corresponding to causal relations among measured variables when there are no unmeasured common causes, but when the search includes structures with unmeasured common causes, causal relations among measured variables are forbidden. Users have found ways around these restrictions (Glymour, et al., 1987; Bollen, 1989), rather to the dissatisfaction of the authors of the program (Joreskog and Sorbom, 1990). LISREL owes these peculiarities to its ancestry in factor analysis, which provides still another example of an artificially contracted search space. Thurstone (1935) carefully and repeatedly emphasized that his "factors" were not to be taken as real causes but only as a mathematical simplifications of the measured correlations. Of course factors were immediately treated as hypothetical causes. But so applied, Thurstone's methods exclude *a priori* any causal relations among measured variables themselves, they exclude the possibility that measured variables are causes of unmeasured variables, and they cannot determine causal structure--only correlations--among the latent variables.

5.2.2 Computational and Statistical Limitations

Some searches examine only a small portion of the possible space of hypotheses because they require computationally intensive iterative algorithms in order to test each hypothesis. For example, the automatic model respecification procedure in LISREL re-estimates the entire model every time it examines a new hypothesis. One consequence is that the slowness of the search prohibits the procedure from examining large portions of the hypothesis space where the truth may be hiding.

Another common problem is that many searches require the determination of conditional independence relations that cannot be reliably tested. Many log-linear search procedures implicitly require the estimation of probabilities conditional on a set of variables whose size equals the total number of variables minus two, no matter what the true structure turns out to be. Estimates of higher order conditional probabilities and tests of higher order conditional independencies tend to be unreliable (especially with variables taking several discrete values) because at reasonable sample sizes most cells corresponding to an array of values of the variables will be empty or nearly empty. This disadvantage is not inherent in the log-linear formalism. A recent algorithm proposed by Fung and Crawford (1990) for searching the set of graphical models (the subset of the hierarchical log-linear models that can be represented by undirected independence graphs) reduces the need for testing high order conditional independencies. A version of the same problem arises for linear regression with a large number of regressors and small sample size, since in tests of the hypothesis that a regression coefficient vanishes, the sample size is effectively reduced by the number of other regressors, or the degrees of freedom are altered, so that the test may have little power against reasonable alternatives.

A related but equally fundamental difficulty is that searches for models of discrete data that use some measure of fit requiring model estimation at each (or any) stage are subject to an exponential increase in the number of cells that must be estimated as the number of variables increases. If, to take the simplest case, the variables are binary, then the number of cells for which an expected value must be computed is 2^n . When $n = 50$, say, the number of cells is astronomical.

One might think that these difficulties will beset any possible reliable search procedure. As we will see in this chapter and the next, that is not the case.

5.2.3 Generating a Single Hypothesis

If a kind of evidence is incapable of reliably distinguishing when one rather than another of several alternative hypotheses is correct, then an adequate search procedure should reflect this fact by outputting all of them. Producing only a single hypothesis in such circumstances misleads the user, and denies her information that may be vital in making decisions.

An example of this sort of flaw is illustrated by the LISREL and EQS programs. Beginning with a structure constructed from background knowledge, each of these programs searches for causal models among linear structures using a best-only beam search. At each stage they free the fixed parameter that is judged will most increase the fit of the model to the data. Since freeing a number of different fixed parameters may result in the very same improvement in fit, the programs employ an arbitrary tie-breaking procedure. The output of the search is a single linear model and any alternative statistically indistinguishable models are ignored.

In a later chapter we will describe a large simulation study of the reliabilities of the statistical search procedures implemented in the LISREL and EQS programs for linear models. Because of the computational problems and arbitrary choices from among indistinguishable models at various stages of search, we find that the procedures are of little value in discovering dependencies in the structures from which the data are generated, even when the programs are given most of the structure correctly to start with, including even correct linear coefficients and variances. The study involves systems with unmeasured variables, but we expect that similar results would be obtained in studies with causally sufficient systems.

5.2.4 Other Approaches

There are several exceptions to the generalization that statistical search strategies have been confined to generate-and-test-best-only procedures. Edwards and Havranek (1987) describe a form of procedure that tests models in sequence, under the assumption that if a model passes the test so will any more general model and if a model fails the test so will any more restricted model. Their proposal is to keep track of a bounding set of rejected hypotheses and a bounding set of accepted hypotheses until all possible hypotheses (in some parametrization) are classified. Apparently unknown to Edwards and Havranek, the same idea was earlier developed at length in the artificial intelligence literature under the name of "version spaces" (Mitchell, 1977). For the applications they have in mind, no analysis of complexity or reliability is available.

5.2.5 Bayesian Methods

The best known discussion of search problems in statistics from a Bayesian perspective is Leamer's (1978). Leamer's book contains a number of interesting points, including a consideration of what a Bayesian should do upon meeting a novel hypothesis, but it does not contain a method for reliable search. Considering the use of regression methods in causal

inference, for example, Leamer subsequently recommended analyzing separately the sets of relevant regressors endorsed by any opinion, and giving separate Bayesian updates of distributions of parameters for each of these sets of regressors. The problem of deciding which variables actually influence an outcome of interest is effectively ignored.

A much more promising Bayesian approach to search has been developed by Cooper and Herskovits (1991, 1992). At present, their procedure is restricted to discrete variables and requires a total ordering such that no later variable can cause an earlier variable. Each directed graph compatible with the order is assigned a prior probability. The joint distribution of the variables assigns each vertex in the graph a distribution conditional on its parents, and these conditional probabilities parametrize the distributions for each graph. Using Dirchelet priors, a density function is imposed on the parameters for each graph. The data are used to update the density function by Bayes' rule. The probability of a graph is then just the integral of the density function over the distributions compatible with the graph. The probability of an edge is the sum of the probabilities of all graphs that contain it. Cooper and Herskovits use a greedy algorithm to construct the output graph in stages. For each vertex X in the graph, the algorithm considers the effect of adding to the parent set of X each individual predecessor of X that is not already a parent of X ; it chooses the vertex whose addition to the parent set of X most increases the posterior probability of the local structure consisting of X and its parents. Parents are added to X in this fashion until there is no single vertex that can be added to the parent set of X that will increase the posterior probability of the local structure. The program runs very well even on quite large sets of variables provided the true graph is sparse, and on discrete data with a prior ordering appears to determine adjacencies with remarkable accuracy. Its accuracy on dense graphs is not known at this time.

The Bayesian approach developed by Cooper and Herskovits has the advantages that appropriate prior degrees of belief can be used in search, that models are output with ratios of posterior distributions consistent with the specified prior distribution and the data, and that under appropriate assumptions¹ the method converges to the correct graph. Because the method can calculate the ratio of the posterior probabilities of any pair of graphs, it is possible to make inferences over multiple graphs weighted by the probability of the graph (although generally some heuristic to consider only the most probable graphs must be used because of the sheer

¹ In particular, when the method is idealized to give up the greedy algorithm. Because of the greedy algorithm, we would expect the specific search procedure to be asymptotically unreliable when there are two or more treks between a pair of non-adjacent variables, say X and Y , that result in a close statistical association between those variables. This is the circumstance in the case of the one edge the procedure erroneously introduces in the ALARM network. In practice, such structures may be sufficiently uncommon for the error to be tolerable., and Cooper and his colleagues are investigating techniques to ameliorate the problem.

number of possibilities.) The method works with Dirchelet priors because the relevant integrals are available analytically and posterior densities can therefore be rapidly evaluated without any numerical analysis. In view of the combinatorics of graphs, any other application of the search architecture must have the same feature. One problem is to extend the method to continuous variables, which depends on finding a family of conjugate priors that can be rapidly updated for parameters that describe graphs. Another, more fundamental, problem concerns whether the requirement of a prior ordering of the variables can be relaxed while preserving computational feasibility. Using a fixed ordering of the variables reduces the combinatorics enormously, but in many applied cases any such ordering may be uncertain. Since the procedure is reasonably fast, requiring about 15 minutes (on a Macintosh II) to analyze data from the ALARM network described in Chapter 1, Cooper and his colleagues are investigating procedures that use a number of orderings and compare the posterior probabilities of the graphs obtained.

5.3 The Wermuth-Lauritzen Algorithm

In 1983 Wermuth and Lauritzen defined what they called a *recursive diagram*. A recursive diagram is a directed acyclic graph G together with a total ordering of the vertices of the graph such that $V_1 \rightarrow V_2$ occurs only if $V_1 < V_2$ in the ordering. In addition there is a probability distribution P on the vertices such that V_i is a parent of V_k if and only if $V_i < V_k$ and V_i and V_k are dependent conditional on the set of all other variables previous to V_k in the ordering. Following Whittaker (1990), we call such systems *directed independence graphs*.

We can view this definition as an algorithm for constructing causal graphs from conditional independence relations and a time ordering of the variables. It has in fact been used in this way by some authors (Whittaker, 1990). Given an ordering of the variables and a list of the conditional independence relations, proceed through the variables in their time order, and for each variable V_k to each variable V_i such that $V_i < V_k$ apply the dependence test in the definition, and add $V_i \rightarrow V_k$ if the test is passed. The procedure will correctly recover the directed graph from the order and the independence relations of a faithful distribution in which, for discrete variables, every combination of variable values has positive probability. In a sense, the discovery problem for causally sufficient faithful systems is solved. In practice, however, the Wermuth-Lauritzen algorithm is not feasible save for very small variable sets. The remaining issues are therefore these:

- (i) how to remove the requirement that an ordering of the variables be known beforehand;

- (ii) how to improve on the computational efficiency and statistical requirements of the Wermuth-Lauritzen procedure;
- (iii) how to remove the tacit restriction to causally sufficient systems of variables.

In this chapter we will address the first two of these problems. The problem of causal inference when unmeasured common causes, or "latent variables," may be acting will be taken up in Chapter 6.

5.4 New Algorithms

We will describe several algorithms for discovering causal structure (assuming causal sufficiency); they eliminate the need for a prior ordering of the variables and all but two of them improve computational efficiency and reduce the difficulty of statistical decisions in comparison with the Wermuth-Lauritzen algorithm. Some of the improvements are dramatic, others less so. Each of the search procedures described can also be used on discrete data to search for graphical log-linear models. (For each triple of variables, if $X \rightarrow Y \leftarrow Z$ occurs in the directed graph, and X is not adjacent to Z , add an undirected edge between X and Z ; then remove all arrowheads from the graph. The result is an undirected independence graph.)

Under the following assumptions all of the algorithms presented in this section provably recover features of graphs faithful to the population distribution:

- (i) The set of observed variables is causally sufficient.
- (ii) Every unit in the population has the same causal relations among the variables.
- (iii) The distribution of the observed variables is faithful to an acyclic directed graph of the causal structure (in the discrete case) or linearly faithful to such a graph (in the linear case).
- (iv) The statistical decisions required by the algorithms are correct for the population.

The fourth requirement is unnecessarily strong, since the algorithms will in many cases succeed even if some statistical decisions are in error. Nonetheless, this is a strong set of assumptions that is often not met in practice, but it is no stronger than the assumptions that would be required to warrant most of the particular statistical models with a causal interpretation found in the medical, behavioral, and social scientific literature. In subsequent chapters we will examine the consequences of weakening some of these assumptions.

In practice, the algorithms take as input either a covariance matrix or cell counts. Where d-separation facts are needed by an algorithm, in the discrete case the procedure performs tests of conditional independence and in the linear continuous case tests for vanishing partial correlations. (Recall that if P is a discrete distribution faithful to a graph G , then A and B are d-separated given a set of variables C if and only A and B are conditionally independent given C , and if P is a distribution linearly faithful to a graph G , then A and B are d-separated given C if and only if $\rho_{AB|C} = 0$.) The algorithms construct the set of directed acyclic graphs that satisfy the given set of d-separability relations, if any such graph exists. Since the results of either kind of test are used only to determine the d-separation relations among the variables, we will speak as if the input to the algorithms is simply the d-separation relations themselves.²

Let us say that a graph G **faithfully represents a list of d-separations L** if and only if all and only the d-separations in L are true of G . A **list L of d-separations is faithful** if and only some acyclic directed graph faithfully represents L . In practice, even if a distribution is faithful to the causal structure that generates it, sampling error or minor violations of the assumptions of the statistical tests employed can lead to errors in judgment about the properties of the population. The robustness of the procedures to erroneous specification of the distribution family or to sampling variation can be investigated by Monte Carlo simulation methods.

Each of the following algorithms can have as output either a class of directed acyclic graphs, or else a single mixed object with both directed and undirected edges-- the pattern that represents a class of graphs. Recall that pattern Π represents a set of directed acyclic graphs. A graph G is in the set of graphs represented by Π if and only if:

- (i) G has the same adjacency relations as Π ;
- (ii) if the edge between A and B is oriented $A \rightarrow B$ in Π , then it is oriented $A \rightarrow B$ in G ;
- (iii) if Y is an unshielded collider on the path $\langle X, Y, Z \rangle$ in G then Y is an unshielded collider on $\langle X, Y, Z \rangle$ in Π .

If any of the algorithms use as input a covariance matrix from a distribution linearly faithful to G , or cell counts from a distribution faithful to G , we will say the input is **data faithful to G** . All of the algorithms we will discuss in this section have the following correctness property:

²Indeed, any statistical constraint can be used as input for the algorithms for any pairing of distributions with graphs such that the constraint is satisfied in the distribution if and only if the corresponding d-separation relation holds in the graph.

Theorem 5.1: If the input to any of the algorithms is data faithful to G , the output of each of the algorithms is a pattern that represents G .

The algorithms do not, however, always provide a pattern that explicitly characterizes all of the orientation information implicit in the d-separation facts; a pattern may be produced that is consistent only with one orientation of an edge but does not explicitly contain the corresponding arrowhead.

5.4.1 The SGS Algorithm

The correctness of the SGS algorithm (Spirtes, Glymour and Scheines, 1990c) follows from Theorem 3.4:

Theorem 3.4: If P is faithful to some directed acyclic graph, then P is faithful to G if and only if

- (i) for all vertices, X, Y of G , X and Y are adjacent if and only if X and Y are dependent conditional on every set of vertices of G that does not include X or Y ; and
- (ii) for all vertices X, Y, Z such that X is adjacent to Y and Y is adjacent to Z and X and Z are not adjacent, $X \rightarrow Y \leftarrow Z$ is a subgraph of G if and only if X, Z are dependent conditional on every set containing Y but not X or Z .

SGS Algorithm:

A.) Form the complete undirected graph H on the vertex set V .

B.) For each pair of vertices A and B , if there exists a subset S of $V \setminus \{A, B\}$ such that A and B are d-separated given S , remove the edge between A and B from H .

C.) Let K be the undirected graph resulting from step B). For each triple of vertices A, B , and C such that the pair A and B and the pair B and C are each adjacent in K (written as $A - B - C$) but the pair A and C are not adjacent in K , orient $A - B - C$ as $A \rightarrow B \leftarrow C$ if and only if there is no subset S of $\{B\} \cup V \setminus \{A, C\}$ that d-separates A and C .

D.) repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B , then orient $B - C$ as $B \rightarrow C$.

If there is a directed path from A to B , and an edge between A and B , then orient $A - B$ as $A \rightarrow B$.

until no more edges can be oriented.

5.4.1.1 Complexity

Reliability is one thing, efficiency another. Step B) of the SGS algorithm requires that for each pair of variables adjacent in G we look at all possible subsets of the remaining variables, and that, of course, is an exponential search. In the worst case that complexity is unavoidable if reliability is to be maintained. Two variables can be dependent conditional on a set \mathbf{U} but independent on a superset or subset of \mathbf{U} . Any procedure that in the worst case does not examine the conditional independence relations of variables X, Y on all subsets of vertices not containing that pair will fail--there will be some structure the procedure does not get correctly.

5.4.1.2 Stability of SGS

We need to consider whether an algorithm remains reasonably reliable when the data are imperfect. We will use the notion of stability informally: If intuitively small errors of input produce intuitively large errors of output, the algorithm is not stable. For the SGS algorithm, an intuitively small error in input consists of a few d-separation relations that are falsely included or falsely excluded from the input. An intuitively small error for Step B is a few undirected edges erroneously included in or omitted from the output. An intuitively small error for Step C is a few edges misoriented.

Step B) of the SGS algorithm is stable. If, for example, a single correct d-separation relation is omitted from the input, the algorithm will nonetheless produce the correct undirected graph unless there is no other set besides \mathbf{U} on which X, Y are d-separated. Even in that case Step B will make an error in postulating an $X - Y$ connection, but no other errors. If X and Y are adjacent in the true graph, but it is incorrectly judged that X and Y are d-separated given \mathbf{U} , the algorithm will fail to connect X and Y but no other error will be made.

Step C) of the SGS algorithm is less stable. A small error in either component of the input, either the undirected graph or the list of d-separation relations, can (and often will) produce large errors in the output. That is because the edges that occur in collisions determine the orientations of other edges in the graph, and if input errors lead the algorithm erroneously to include or exclude a collision, the error may affect the orientations of many other edges in the graph.

Suppose, for example, an edge connecting X, Z is erroneously omitted in the undirected graph input to step C), and $X - Y - Z$ correctly occurs in the input. Then if X and Z are not d-separated

by any subset of variables containing Y but not X, Z , the algorithm will mistakenly require a collision at Y , and this requirement will ramify through orientations of other edges. Or, if the true structure contains a collision at Y but $X - Y$ is omitted in the input to step C), no unique orientation will be given to $Y - Z$, and this uncertainty may ramify through the orientations of other edges on paths including Z .

Instabilities may also arise in Step C) because of errors in the list of d-separation relations input, even when the underlying undirected graph is correct. If in the input to C), X is adjacent to Y and Y to Z but not X to Z and a d-separation relation between X and Z given S containing Y is omitted from the input, no orientation error will result unless no other set containing Y d-separates X and Z . But if in the true directed graph, the edges between X and Y and between Y and Z collide at Y , and a d-separation relation involving X and Z and some set U containing Y but not X or Z is erroneously included in the input, the algorithm will conclude that there is no collision at Y , and this error may be ramified to other edges.

A little reflection on Step C) reveals that its output may not be a collection of directed acyclic graphs if one of the four assumptions listed at the beginning of this section is violated. This is not necessarily a defect of the algorithm. If the algorithm finds that the edges $X - Y - Z$ collide at Y , and $Y - Z - W$ collide at Z , it will create a pattern with an edge $Y <-> Z$. Double headed edges can occur when the causal structure is not causally sufficient, or when there is an error in input (as from sampling variation). They have a theoretical role in identifying the presence of unmeasured common causes, an issue discussed further in the next chapter.

5.4.2 The PC Algorithm

In the worst case, the SGS algorithm requires a number of d-separation tests that increases exponentially with the number of vertices, as must any algorithm based on conditional independence relations or vanishing partial correlations. But the SGS algorithm is very inefficient because for edges in the true graph the worst case is also the expected case. For any undirected edge that is in the graph G , the number of d-separation tests that must be conducted in stage B) of the algorithm is unaffected by the connectivity of the true graph, and therefore even for sparse graphs the algorithm rapidly becomes infeasible as the number of vertices increases. Besides problems of computational feasibility, the algorithm has problems of reliability when applied to sample data. The determination of higher order conditional independence relations from sample distributions is generally less reliable than is the determination of lower order independence relations. With, say, 37 variables taking three values

each, to determine the conditional independence of two variables on the set of all remaining variables requires considering the relations among the frequencies of 3^{35} distinct states, only a fraction of which will be instantiated even in very large samples.

We should like an algorithm that has the same input/output relations as the SGS procedure for faithful distributions but which for sparse graphs does not require the testing of higher order independence relations in the discrete case, and in any case requires testing as few d-separation relations as possible. The following procedure (Spirtes, Glymour, and Scheines, 1991) starts by forming the complete undirected graph, then "thins" that graph by removing edges with zero order conditional independence relations, thins again with first order conditional independence relations, and so on. The set of variables conditioned on need only be a subset of the set of variables adjacent to one or the other of the variables conditioned.

Let **Adjacencies**(C, A) be the set of vertices adjacent to A in directed acyclic graph C . (In the algorithm, the graph C is continually updated, so **Adjacencies**(C, A) is constantly changing as the algorithm progresses.)

PC Algorithm:

A.) Form the complete undirected graph C on the vertex set \mathbf{V} .

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that **Adjacencies**($C, X \setminus \{Y\}$) has cardinality greater than or equal to n , and a subset S of **Adjacencies**($C, X \setminus \{Y\}$) of cardinality n , and if X and Y are d-separated given S delete edge $X - Y$ from C and record S in **Sepset**(X, Y) and **Sepset**(Y, X);

until all ordered pairs of adjacent variables X and Y such that **Adjacencies**($C, X \setminus \{Y\}$) has cardinality greater than or equal to n and all subsets S of **Adjacencies**($C, X \setminus \{Y\}$) of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y , **Adjacencies**($C, X \setminus \{Y\}$) is of cardinality less than n .

C.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in C but the pair X, Z are not adjacent in C , orient $X - Y - Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in $\text{Sepset}(X, Z)$.

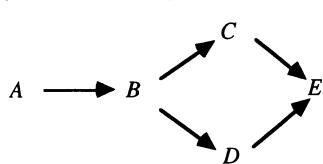
D. repeat

If $A \rightarrow B$, B and C are adjacent, A and C are not adjacent, and there is no arrowhead at B , then orient $B - C$ as $B \rightarrow C$.

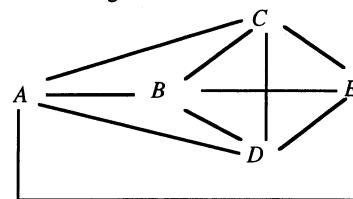
If there is a directed path from A to B , and an edge between A and B , then orient $A - B$ as $A \rightarrow B$.

until no more edges can be oriented.

Figure 1 traces the operation of the first two parts of the PC algorithm:



True Graph



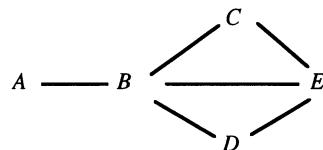
Complete Undirected Graph

$n = 0$ No zero order independencies

$n = 1$ First order independencies

Resulting Adjacencies

$$\begin{array}{ll} A \perp\!\!\!\perp C \mid B & A \perp\!\!\!\perp D \mid B \\ A \perp\!\!\!\perp E \mid B & C \perp\!\!\!\perp D \mid B \end{array}$$



$n = 2$: Second order independencies

Resulting Adjacencies

$$B \perp\!\!\!\perp E \mid \{C, D\}$$

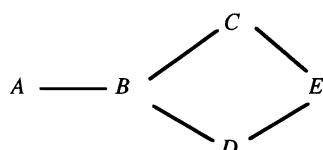


Figure 1

Although it does not in this case, stage B) of the algorithm may continue testing for some steps after the set of adjacencies in the true directed graph has been identified. The undirected graph at the bottom of figure 1 is now partially oriented in step C). The triples of variables with only two adjacencies among them are:

$$\begin{array}{ll} A - B - C; & A - B - D; \\ C - B - D; & B - C - E; \\ B - D - E; & C - E - D \end{array}$$

E is not in $\text{Sepset}(C,D)$ so $C - E$ and $E - D$ collide at E . None of the other triples form colliders. The final pattern produced by the algorithm is shown in figure 2.

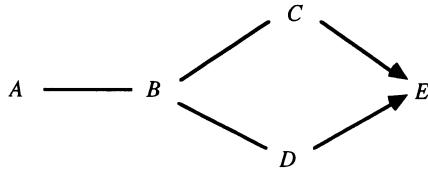


Figure 2

The pattern in figure 2 characterizes a faithful indistinguishability class. Every orientation of the undirected edges in figure 2 is permissible that does not include a collision at B .

5.4.2.1 Complexity

The complexity of the algorithm for a graph G is bounded by the largest degree in G . Let k be the maximal degree of any vertex and let n be the number of vertices. Then in the worst case the number of conditional independence tests required by the algorithm is bounded by

$$2 \binom{n}{2} \sum_{i=0}^k \binom{n-1}{i}$$

which is bounded by

$$\frac{n^2(n-1)^{k-1}}{(k-1)!}$$

This is a loose upper bound even in the worst case; it assumes that in the worst case for n and k , no two variables are d-separated by a set of less than cardinality k , and for many values of n and k we have been unable to find graphs with that property. While we have no formal expected complexity analysis of the problem, the worst case is clearly rare, and the average number of conditional independence tests required for graphs of maximal degree k is much smaller. In practice it is possible to recover sparse graphs with as many as a hundred variables. Of course the computational requirements increase exponentially with k .

The structure of the algorithm and the fact that it continues to test even after having found the correct graph suggest a natural heuristic for very large variable sets whose causal connections are expected to be sparse, namely to fix a bound on the order of conditional independence relations that will be tested.

5.4.2.2 Stability of PC

In theory, the PC Algorithm is unstable in both steps B) and C) although in practice step B) has proved to be much more reliable than step C).

If an edge is mistakenly removed from the true graph at an early stage of step B) of the algorithm, then other edges which are not in the true graph may be included in the output. Consider the following example.

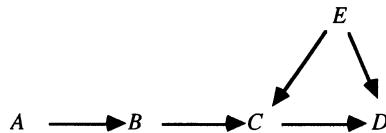


Figure 3

If the edge $E - D$ is mistakenly removed from the initial complete graph then at a subsequent stage of the search the edge $B - D$ will not be removed, because E will no longer be in the adjacency set for D , and B and D are dependent on every subset of A and C . The omission of an edge can also lead to orientation errors. If an edge is mistakenly left in the graph and there

are no additional errors in the list of d-separations in the input, the only further errors that result are that some edges which theoretically could be oriented, will not be oriented.

Step C) of the algorithm is unstable for the same reasons as in step C) of the SGS algorithm.

The PC algorithm is faster than the SGS algorithm because it tests fewer d-separation relations. Given a faithful list of d-separability relations, the two algorithms output the same set of pattern graphs. But if the list of d-separability relations is not faithful, due to sampling error for example, the two algorithms can output different pattern graphs. Consider the following example.



Figure 4

According to this graph, A and E are d-separated from each other given any non-empty subset of B , C , and D . If, after the $A - C$ and $E - C$ edges have been removed from the initial undirected graph, the procedure incorrectly judges that A and E are not d-separated given any non-empty subset of B and D , the PC algorithm will incorrectly include an edge between A and E , because it only tests whether A and E are d-separated given subsets of the adjacencies of A and E . On the other hand, because the SGS algorithm tests whether A and E are d-separated given any subset of $V \setminus \{A, B\}$, it would properly recognize that there is not an edge between A and E because A and E are d-separated given C .

In contrast, if after the $A - E$ and $B - E$ edges are removed from the initial undirected graph, it is mistakenly judged that A and B are d-separated given E , the SGS algorithm will mistakenly remove the $A -> B$ edge. If the $A - E$ and $B - E$ edges are removed first, the PC algorithm, would correctly leave the $A -> B$ edge in, because it would not test whether A and B are d-separated given E .

Because the PC algorithm attempts to use "local" information to judge whether an edge exists or not, it is not guaranteed to produce a graph that is in some sense "closest" to an unfaithful distribution. Consider the following example.

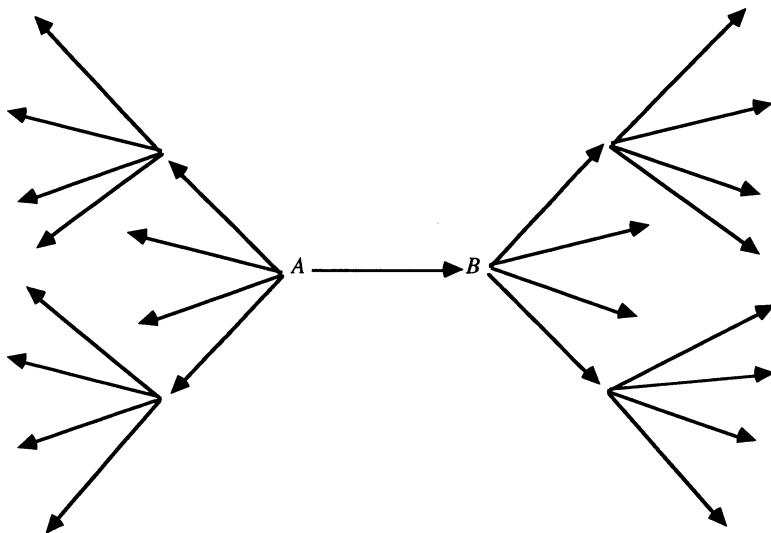


Figure 5

In a distribution faithful to this graph, every variable is dependent on every other variable. Suppose a test determines that A and B are independent conditional on some other variable, either because of some coincidental parameter values, or because of sampling error. The PC algorithm would then remove the $A - B$ edge in order to satisfy that constraint. In doing so, however, it would disconnect the graph. The resulting graph would entail that A and all of its descendants to the left are independent of B and all of its descendants. So, in order to satisfy one conditional independence constraint, the PC algorithm may produce a graph that violates a great many independence constraints. In a number of data sets the correlations between two variables do not vanish but the output pattern disconnects them. For greater reliability the procedure should be supplemented with a repair algorithm, for which the Cooper and Herskovits Bayesian procedure might suffice in the discrete case; alternatively, a variation of the procedure described in Chapter 11 could be applied.

5.4.2.3 The PC* Algorithm

The PC algorithm is computationally efficient and asymptotically reliable, but on sample data the procedure takes unnecessary risks. In determining whether to eliminate an undirected edge between variables A and B , the procedure may test every subset of the adjacency set of A and of

the adjacency set of B . But the independence or dependence of A and B on many of these subsets of variables may be entirely irrelevant to the causal relations between A and B . For a distribution faithful to a directed acyclic graph, if variables A and B are independent conditional given **Parents**(A) or given **Parents**(B) then they are independent given a subset of **Parents**(A) or given a subset of **Parents**(B) consisting only of vertices lying on undirected paths between A and B . It is sufficient, then, to test for the conditional independence of A and B given subsets of variables adjacent to A and subsets of variables adjacent to B that are on undirected paths between A and B . Call the modified algorithm PC*.

The PC and PC* algorithms yield the same output given a faithful list of conditional independence relations or correlations as input, but they may differ given conditional independence relations determined from tests on sample data. The PC* algorithm avoids one kind of error made by the PC algorithm. If, however, at an early stage the PC* algorithm mistakenly disconnects a path between X and Y it may then mistakenly leave the $X - Y$ edge in the undirected graph, while the PC algorithm, given the same data, might avoid that error. Moreover, whatever increased reliability the PC* algorithm may have is bought at great cost, since the algorithm must at each stage of step B) keep track of all of the undirected paths in the graph it considers at that stage. The number of undirected paths is typically very large, and the memory requirements of the PC* algorithm are not feasible save for relatively small numbers of variables, in which case it may be the algorithm of choice. For large numbers of variables the PC algorithm must be used instead, although if the true graph is sparse, the PC algorithm can be used until the average degree of the undirected graph C is small, after which stage the PC* algorithm may be used. Later in this chapter we will describe the performance of the two algorithms on discrete data taken from Christensen (1990).

5.4.2.4 Speed-Up Heuristics for Ordering Tests

Step B of the PC algorithm selects some variable pair and some subset S of a given size to test for d-separation. The faster edges are eliminated from the complete graph, the smaller the search that has to be conducted at later stages of the algorithm, and the faster the algorithm runs. Hence, it is best to select first for testing those variable pairs A and B and subsets S for which A and B are most likely to be d-separated by S . We have considered three variants of the PC algorithm that use different methods of selecting the order of tests.

Heuristic 1: Test the variable pairs and subsets S in lexicographic order. (We will call this PC-1.)

Heuristic 2: First test those variables pairs that are least dependent³ in probability. The conditioning subsets are selected by lexicographic order. (We will call this PC-2.)

Heuristic 3: For a given variable A , first test those variables B that are least probabilistically dependent on A , conditional on those subsets of variables that are most probabilistically dependent on A . (We will call this PC-3.)

The intuition behind heuristic 2 is that variables with the highest probabilistic dependence are most likely to be adjacent in the true graph, and hence not ever eliminated from the graph being constructed, while those with the smallest probabilistic dependence are least likely to be adjacent in the true graph. Of course, no such relation strictly holds.

The intuition behind heuristic 3 is similar. A variable B that is not genuinely adjacent to a variable A is d-separated from A given some subset of the variables that are adjacent to A or given some subset of the variables that are adjacent to B in the true graph. Assuming that variables with the highest probabilistic dependence upon A are most likely to be adjacent to A in the true graph, this suggests testing whether A is d-separated from variables with a low probabilistic dependence on A , conditional on variables with a high probabilistic dependence upon A .

5.4.3 The IG (Independence Graph) Algorithm

Verma and Pearl (1990) have suggested a variation of the SGS algorithm. In their alternative, the first step in searching for the directed acyclic graph is to construct the undirected independence graph N , i.e. for each pair of variables A, B introduce an undirected edge between them if they are dependent conditional on the set of all other variables. In the undirected independence graph for a distribution faithful to a directed acyclic graph the parents of any variable form a maximal complete subgraph--a clique. Again for each pair of variables A, B adjacent in N , determine if A and B are d-separated given any subsets of variables in the cliques in N containing A or B . If so A is not adjacent to B in G . The complexity is thus a function of the size of the largest clique in N .

³In the following heuristics, "high probabilistic dependence" means high partial correlation in the linear case, and high G^2 statistic in the discrete case.

Determining the cliques in a graph would appear to require unnecessary computation, and in other than the worst case, testing for conditional independence of two variables conditional on all members of the maximal clique of one or the other will involve a test of unnecessarily high order. A better idea might be to blend the procedure with the PC algorithm: modify step A of the PC algorithm by setting the initial graph in the PC procedure to the *undirected independence graph*, rather than the complete undirected graph, and then proceed in the same way. We will call this algorithm IG (independence graph.)

The efficiency of these algorithms obviously depends upon how easily the independence graph can be constructed. The off-diagonal elements of the standardized inverse of the correlation matrix are the negatives of the partial correlation coefficients between the corresponding variables given the remaining variables (see e.g. Whittaker, 1990). Hence in the linear case, the independence graph can be efficiently constructed by placing an edge between A and B if and only if the entry in the standardized inverse correlation matrix is non-zero. In the discrete case, Fung and Crawford (1990) have recently proposed a fast algorithm for constructing an independence graph from discrete data. We have not tested their procedure as a preprocessor for the PC algorithm.

5.4.4 Variable Selection

While prior knowledge of causal structure can sometimes make the results of the algorithms we have described more informative on real samples, correct selection of variables is essential for reliable inference, and for that algorithms (at least these algorithms) provide no help.

We can aggregate variables or we can aggregate distinct values of a variable. As in Salmon's imaginary example discussed in Chapter 3, we sometimes measure a variable that is an imprecise version of a more precise natural variable; we fail, in other words, to distinguish values that have differing effects on other variables. Continuous variables are often deliberately collapsed into a few discrete categories, sometimes because contingency table methods offer the promise of statistical analysis free of the substantive assumptions that would otherwise be required about the form of the functional dependencies--e.g., linear or otherwise--and sometimes because some of the variables to be analyzed are necessarily discrete and there are few methods available for problems with mixtures of discrete and continuous variables. Sometimes, whether through ignorance or even deliberately, we may aggregate two or more distinct variables with distinct causal structures into a single scale. What effects can aggregation and collapse have on the reliability of causal inference?

We have already observed that if C is a cause of A and B and some proxy C' for C is used that is not so precise as C and not perfectly correlated with C , it may be that A and B are statistically dependent conditional on C' . Examples of this sort appear whenever a theory postulates a cause that is measured by proxies. Friedman (1957), for example, advocated a much discussed theory in which consumption is caused by "permanent" income which can only be measured by proxies; if Friedman's theory were true, regression of consumption on measured income would provide a biased estimate of the regression coefficient of consumption on permanent income and might leave unexplained correlations between consumption and other variables. Klepper (1988) has shown how, in the linear normal case, such errors may be bounded.

Suppose we are given variables A, B, C such that A and B are independent conditional on C . Let $C' = \text{PROJ}(C)$ where $\text{PROJ}(C)$ is a projection mapping the set of n values of C to a set of $m < n$ values. If there exist values c_1, c_2 for C such that $P(A, B | C=c_1) \neq P(A, B | C=c_2)$ and $\text{PROJ}(C=c_1) = \text{PROJ}(C=c_2)$, then A and B are not independent conditional on C' . Independence relations can be made to appear rather than disappear by collapsing values of a variable. Suppose that variable B, C are dependent. Let $C' = \text{PROJ}(C)$ where $\text{PROJ}(C)$ is a projection mapping the set of n values of C to a set of $m < n$ values. If there exists a value c_1 of C such that $P(C = c_1 | B) = P(C = c_1)$ and $\text{PROJ}(c_1)$ has a unique inverse and $\text{PROJ}(c_k) = \text{PROJ}(c_j)$ for all k, j not equal to 1, then B and C' are independent.

Pearl (personal communication) has pointed out that a very simple sort of aggregation can produce an unfaithful distribution. Suppose A causes C_1 and B causes C_2 , and C_1 and C_2 are each binary, and there is no other causal connection among the variables. So $\{A, C_1\}$ is independent of the set $\{B, C_2\}$, but A and C_1 are dependent and so are B and C_2 . Introduce variable C taking values 0, 1, 2, 3 coding the different value pairs for C_1 and C_2 . Then the actual causal structure among A, B and C is shown in figure 6.

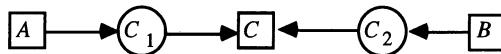


Figure 6

But in the joint distribution A and B are independent conditional on C , and so the joint distribution is not faithful to any causal structure whatsoever. In this case the unfaithfulness of the distribution is due to the fact that it is the marginal of a distribution that is unfaithful because of deterministic relationships among the variables: the independence of A and B given C

follows directly from the application of D-separability (see Chapter 3) to figure 6. This sort of thing may sometimes happen in practice, but it could always be tested for and in principle identified: Conditioning on A divides the values of C into two equivalence classes each containing values of C with the same conditional probability, and conditioning on B divides the values of C into a distinct pair of equivalence classes. Letting the equivalence classes induced by A be values of one variable and the equivalence classes induced by B be values of another variable recovers C_1 and C_2 .

5.4.5 Incorporating Background Knowledge

A user of any of these algorithms may have a great deal of background knowledge--or at least belief--that could constrain the search. This knowledge might be about the existence or non-existence of certain edges in the graph, or it might be about the orientation of some of the edges, or it might be about the time order of the variables. How can this background knowledge be used by the algorithms?

The most common sort of reliable prior belief orders or partially orders the variables by time of occurrence: either measurements of A were taken before measurements of B , or A and B are believed to be exact measures of events that are so ordered. Any of the algorithms of this section can be easily modified to make two uses of such knowledge:

- (i) In determining whether A and B are adjacent in the true graph by testing whether B is independent of A conditional on some subset of the current adjacencies of A , do not test for independence conditional on any set of variables that includes a variable that is later than A .
- (ii) If A and B are adjacent and B is later than A , orient the edge as $A \rightarrow B$.

In the examples we give throughout this book the algorithms have been so modified, and we sometimes make use of common sense time order, always noting where such assumptions have been made.

Prior belief about whether one variable directly influences another can also be incorporated in these algorithms: if prior belief forbids an adjacency, for example, the algorithms need not bother to test for that adjacency; if prior belief requires that there be a direct influence of one variable on another, the corresponding directed edge is imposed and assumed in the orientation procedures for other edges. These procedures assume that prior belief should override the results of unconstrained search, a preference that may not always be judicious; they are nonetheless incorporated in versions of the TETRAD II program with the PC algorithm.

5.5 Statistical Decisions

The algorithms we have described are completely modular, and can be applied given any procedures for making the requisite statistical decisions about conditional independence or vanishing partial correlations. The better the decisions the better the performance to be expected from the algorithms. While tests of conditional independence relations form the most obvious class of such decisions, any statistical constraints that give d-separability relations for graphical structure will suffice. For example, in the linear normal case, vanishing partial correlation is equivalent to conditional independence, and the statistical decisions required by the algorithms could be provided by t -tests of the hypotheses that partial correlations vanish. But vanishing partial correlation marks d-separability whether or not the distribution is normal, so long as linearity and linear Faithfulness hold.⁴ Hence under these assumptions the test of any statistic that vanishes when partial correlations vanish would suffice; one might, for example, use an F test for the square of the semi-partial correlation coefficient, which equals the square of the t -test for a corresponding regression coefficient (Edwards, 1976).

In the examples in this book we test whether $\rho_{XY,C} = 0$ using Fisher's z :

$$z(\rho_{XY,C}, n) = \frac{1}{2} \sqrt{n - |C| - 3} \ln \left[\frac{(1 + \rho_{XY,C})}{(1 - \rho_{XY,C})} \right]$$

$\rho_{XY,C}$ = population partial correlation of X and Y given C , and $|C|$ equals the number of variables in C . If X , Y , and C are normally distributed and $r_{XY,C}$ denotes the sample partial correlation of X and Y given C , the distribution of $z(\rho_{XY,C}, n) - z(r_{XY,C}, n)$ is standard normal (Anderson, 1984).

In the discrete case, for simplicity consider two variables. Recall that we view the count in a particular cell, x_{ij} , as the value of a random variable obtained from sampling N units from a

⁴ For causally sufficient structures, if a distribution P , obtained by imposing a linear distribution compatible with a graph G , implies some vanishing partial correlation not linearly implied by G , is then P not faithful to G ? If P is not faithful to G , does P necessarily imply some vanishing partial correlation not linearly implied by G ? We don't know the answer to either question.

multinomial distribution. Let x_{i+} denote the sum of the counts in all cells in which the first variable has the value i , and similarly let x_{+j} denote the sum of the counts in all cells in which the second variable has the value j . On the hypothesis that the first and second variables are independent, the expected value of the random variable x_{ij} is:

$$E(x_{ij}) = \frac{x_{i+}x_{+j}}{N}$$

Analogously, we can compute the expected values of cells on any hypothesis of conditional independence from appropriate marginals. For example, on the hypothesis that the first variable is independent of the second conditional on the third, the expected value of the cell x_{ijk} is

$$E(x_{ijk}) = \frac{x_{i+k}x_{+jk}}{x_{++k}}$$

If there are more than three variables this formula applies to the expected value of the marginal count of the i, j, k values of the first three variables, obtained by summing over all other variables. The number of independent constraints that a conditional independence hypothesis places on a distribution is an exponential function of the order of the conditional independence relation and also depends on the number of distinct values each variable can assume.

Tests of such independence hypotheses have used--among others--two statistics:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$G^2 = 2 \sum (Observed) \ln \left(\frac{Observed}{Expected} \right)$$

each asymptotically distributed as χ^2 with appropriate degrees of freedom. In the examples in this book we calculate the degrees of freedom for a test of the independence of A and B conditional on C in the following way. Let $Cat(X)$ be a function which returns the number of categories of the variable X , and n be the number of variables in C . Then the number of degrees of freedom (df) in the test is:

$$df = (Cat(A) - 1) \times (Cat(B) - 1) \times \prod_{i=1}^n Cat(C_i)$$

We assume that there are no structural zeroes. As a heuristic, for each cell of the distribution that has a zero entry, we reduce the number of degrees of freedom by one⁵.

Because the number of cells grows exponentially with the number of variables, it is easy to construct cases with far more cells than there are data points. In that event most cells in the full joint distribution will be empty, and even non-empty cells may have only small counts. Indeed, it can readily happen that some of the marginal totals are zero and in these cases the number of degrees of freedom must be reduced in the test. For reliable estimation and testing, Fienberg recommends that the sample size be at least five times the number of cells whose expected values are determined by the hypothesis under test.

For discrete data we fill out the PC algorithm with tests for independence using G^2 which in simulations we have found more often leads to the correct graph than does X^2 . In testing the conditional independence of two variables given a set of other variables, if the sample size is less than ten times the number of cells to be fitted we assume the variables are conditionally dependent.

5.6 Reliability and Probabilities of Error

Most of the algorithms we have described require statistical decisions which, as we have just noted, can be implemented in the form of hypothesis tests. But the parameters of the tests cannot be given their ordinary significance. The usual comforts of a statistical test are the significance level, which offers assurance as to the limiting frequency with which a true null hypothesis would erroneously be failed by the test, and the power against an alternative, which is a function of the limiting frequency with which a false null hypothesis would not be rejected when a specified alternative hypothesis is true. Except in very large samples, neither the significance level nor the power of tests used within the search algorithms to decide statistical

⁵An exact general rule for calculating the reduction of degrees of freedom given cells with zero entries seems not to be known. See Bishop, Fienberg, and Holland (1975).

dependence measures the long run frequency of anything interesting about the search. What does?

The error probabilities one might naturally want to know for a search procedure include:

1. Given that model M is true, what is the probability that the procedure will return a conclusion inconsistent with M on sample size n ?
2. Given that model M^* is true, what is the probability that the procedure will return a conclusion inconsistent with M^* but consistent with M on sample size n ?
3. Given that model M is true, for samples of size n what is the probability that a search procedure will specify an adjacency not in M ? What is the probability that a search procedure will omit an adjacency in M ? What is the probability that a search procedure will add an arrowhead not in M to an edge that is in M ? What is the probability that a search procedure will omit an arrowhead in M ? What are these probabilities for any particular variable pair, A, B ?

For large models, where we expect some errors of specification from most samples, questions of kind 3 are the most important.

There is little hope of obtaining analytic answers to these questions. In repeated tests of independence hypotheses in a sample, each using the same significance level, the probability that *some* true hypothesis will be rejected is not given by the significance level; depending on the number of hypotheses and the sample size, that probability may in fact be much higher than the significance level, but in any case the probability of some erroneous decision depends on which hypotheses are tested, and for all of the algorithms considered that in turn depends in a complex way on the actual structure. Further, each of the algorithms can produce correct output even though some required statistical decisions are made incorrectly. For example, suppose in graph G , vertices A and B are not adjacent. Suppose in fact A and B are independent conditional on C , on D , on C and D , and so on. If the hypothesis that A and B are independent conditional on C is rejected in the search procedure, and the algorithm goes on to test whether A and B are independent conditional on D , and decides in favor of the latter independence, then despite the earlier error the procedure will correctly conclude that A and B are not adjacent.

For any particular M and M^* estimates of the answers to questions 1, 2, and 3 can be found empirically by Monte Carlo methods. Simulation packages for linear normal models are now

common in commercial statistical packages, and the TETRAD II program contains a simulation package for linear and for discrete variable models with a variety of distributions. For small models it takes only a few minutes to generate a hundred or more samples and run the samples through the search procedures. Most of the time required is in counting the outcomes, a process that we have automated *ad hoc* for our simulations, and that can and should be automated in a general way for testing the reliabilities of particular search outcomes.

5.7 Estimation

There are well known methods for obtaining maximum likelihood estimates subject to a causal hypothesis under the assumption of normality, even with unmeasured variables, (Joreskog, 1981; Lohmoller, 1989)). A variety of computerized estimation methods, including ordinary and generalized least squares, are also available when the normality assumption is given up. In the discrete case, for a positive multinomial distribution, the maximum likelihood estimates (when they exist) for a cell subject to the independence constraints of the graph over a set of variables V can be obtained by substituting the marginal frequencies for probabilities in the factorization formula of Chapter 3 (Kiiveri and Speed, 1982).

$$P(V) = \prod_{V \in V} P(V | \text{Parents}(V))$$

When there are unmeasured variables that act as common causes of measured variables, the pattern obtained from the procedures we have described can have edges with arrows at each end. In that rather common circumstance we do not know how to obtain a maximum likelihood estimate for the joint distribution of discrete measured variables, but work on estimating log-linear models with latent variables (Haberman, 1979) may serve as a guide.

5.8 Examples and Applications

We illustrate the algorithms for simulated and real data sets. With simulated data the examples illustrate the properties of the algorithms on samples of realistic sizes. In the empirical cases we

often do not know whether an algorithm produces the truth. But it is at the very least interesting that in cases in which investigators have given some care to the treatment and explanation of their data, the algorithm reproduces or nearly reproduces the published accounts of causal relations. It is also interesting that in cases without these virtues the algorithm suggests quite different explanations from those advocated in published reports.

Studies of regression models and alternatives produced by the PC algorithm and by another procedure, the Fast Causal Inference (FCI) algorithm, are postponed until Chapter 8, after latent variables and prediction have been considered in Chapters 6 and 7, respectively.

5.8.1 The Causes of Publishing Productivity

In the social sciences there is a great deal of talk about the importance of "theory" in constructing causal explanations of bodies of data. Of course in explaining a data set one will always eliminate causal graphs that contradict common sense or that violate the time order of variables. But in addition, many practitioners require that every attempt to provide a causal explanation of observational data in the social sciences proceed through the particulars of principles in sociology, psychology, economics, political science, or whatever, and come accompanied with a denial of the possibility of determining a correct explanation from the statistical dependencies alone. In many of these cases the necessity of theory is badly exaggerated. Indeed, for every "recursive" structural equation model in the entire scientific literature, if the assumptions of the model are correct and no unmeasured common causes are postulated, then if the distribution is faithful the statistical dependencies in the population uniquely determine the undirected graph underlying the directed graph of causal relations. And in many cases the population statistics alone determine a direction of some, or even all, edges. When the variables are linearly ordered by time, so that variable A can be a cause of variable B only if A occurs later than B , the statistical dependencies and the time order determine a *unique* directed graph assuming only that the distribution is positive and the Markov and Minimality Conditions are satisfied. The efforts spent citing literature to justify specifications of causal dependencies are not misplaced, but in many cases effort would be better directed towards establishing the fundamental statistical assumptions, including the approximate homogeneity of the units, the correctness of the sampling assumptions, and sometimes the linearity of dependencies.

Here is a recent and rather vivid example. There is a considerable literature on causes of academic success, including publication and citation rates. A recent paper by Rodgers and

Maranto (1989) considers hypotheses about the causes of academic productivity drawn from sociology, economics, and psychology, and produces a combined "theoretically based" model.

Their data were obtained in the following way: solicitations and questionnaires were sent to 932 members of the American Psychological Association who obtained doctoral degrees between 1966 and 1976 and were currently working academic psychologists. Equal numbers of male and female psychologists were sampled, and after deleting respondents who did not have degrees in psychology, did not take their first job in psychology, etc. a sample of 86 men and 76 women was obtained.

The response items were clustered into groups. For example, the *ABILITY* group consisted of measures of the mean *ACT*, *NMSQT* and selectivity scores of the subject's undergraduate institution, together with membership in Phi Beta Kappa and undergraduate honors at graduation. Graduate Program Quality (*GPQ*) consisted of the scholarly quality of department faculty and program effectiveness using national rankings, the fraction of faculty with publications between 1978 and 1980, and whether an editor of a journal was on the department faculty. These response items were treated as indicators--i.e., as effects--of the unmeasured variables *GPQ*, and *ABILITY*. Other measures were quality of first job (*QFJ*), *SEX*, citation rate (*CITES*) and publication rate (*PUBS*). In preliminary analyses they also used an aggregated measure of productivity (*PROD*). The various hypotheses Rodgers and Maranto considered were then treated as linear "structural equation models"⁶ They report the following correlations among the cluster variables

<i>ABILITY</i>	<i>GPQ</i>	<i>PREPROD</i>	<i>QFJ</i>	<i>SEX</i>	<i>CITES</i>	<i>PUBS</i>
1.0						
.62	1.0					
.25	.09	1.0				
.16	.28	.07	1.0			
-.10	.00	.03	.10	1.0		
.29	.25	.34	.37	.13	1.0	
.18	.15	.19	.41	.43	.55	1.0

⁶It is not clear from the article how the correlations of the latent variables, *GPQ* and *ABILITY*, with other variables such as publishing productivity and *QFJ* were obtained. They can be obtained, for example, by using the factor structure as a regression model to calculate estimated factor scores for each subject, or by including the covariances of the latents among the free parameters in a set of structural equations and letting a program such as LISREL estimate their values. In general the results of these procedures will be different.

There follows a very elaborate explanation of causal theories suggested by the pieces of sociological, economic and psychological literature. Rodgers and Maranto estimate no fewer than six different sets of structural equations and corresponding causal theories. The six structures they consider are as follows:

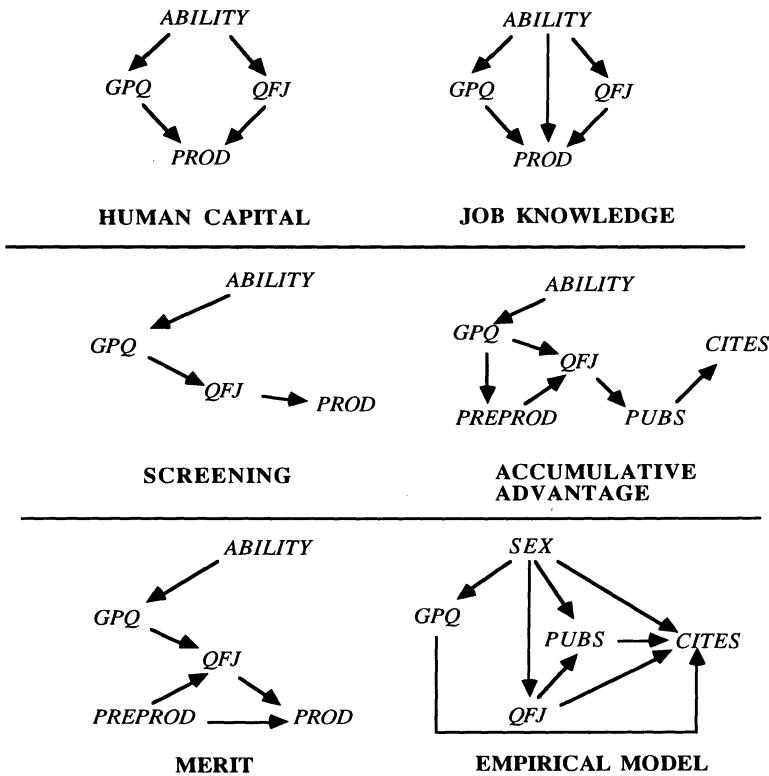


Figure 7

The labels on the graphs indicate simply the social scientific theory from which Rodgers and Maranto derived the causal graph. For example, the "Human Capital" and the "Screening" graphs were obtained from economic theory in the following way:

In the human capital model (Becker, 1964) education has a direct effect on productivity because it conveys relevant knowledge. People invest in education

until its marginal cost (the extra expenses and foregone earnings for an additional year of education) is equal to its marginal benefit (the increase in lifetime earnings caused by another year of education). More able individuals are more productive in both work and the acquisition of skills than their less able counterparts. Thus, ability has a direct effect on productivity and an indirect effect through education, because more able individuals gain more from school. Work experience also increases productivity by providing on-the-job training. The quality as well as the quantity of education is relevant to the human capital framework.

The screening hypothesis implicitly views ability as the primary determinant of productivity. Employers wish to hire the most productive applicants, but ability is not directly observable. Individuals invest in education as a means of signaling their ability to employers. The marginal cost of education is inversely related to ability, inducing a positive correlation between ability and the level of education. Therefore, by selecting applicants based on their education, employers hire by ability (Spence, 1973). In this model, education does not affect productivity directly, but only through its association with ability. Variations in the quality of education are consistent with the screening model (Wise, 1975).

The "empirical model" was obtained from a previous study that did not appeal to social theory.

None of the structural equation systems based on these models save the phenomena. But combining all of the edges in the "theoretical" models, adding two more that seem plausible, and then throwing out statistically insignificant (at .05) dependencies, leads Rodgers and Maranto instead to propose a different causal structure which fits the data quite well:

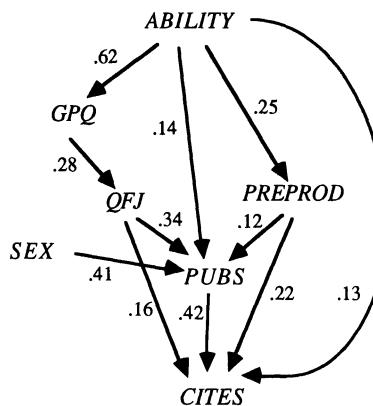


Figure 8

It would appear that the tour through "theory" was nearly useless, but Rodgers and Maranto say otherwise:

Causal models based solely on the pattern of observed correlations are highly suspect. Any data can be fitted by several alternative models. The construction of the best-fit model was thus guided by theory-based expectations. By using the two measures of productivity, *PUBS* and *CITES*, and the five causal antecedents, we initially estimated a composite model with all of the paths identified by the six theories. This model produced a large positive deviation between the observed and predicted correlation of *ABILITY* with *PREPROD*, suggesting that we omitted one or more important paths. Reexamination of our initial interpretation of the six theories led us to conclude that two paths had been overlooked. One such path is from *ABILITY* to *PREPROD*...The other previously unspecified path is from *ABILITY* to *PUBS*. These two paths were added and all nonsignificant paths were deleted from the composite model to arrive at the best-fit model.

If the Rodgers and Maranto theory were completely correct, the undirected graph underlying their directed graph would be uniquely determined by the conditional independence relations, and the orientation would be almost uniquely determined; only the directions of the *GPQ* \rightarrow *QFJ*, *ABILITY* \rightarrow *GPQ* and *ABILITY* \rightarrow *PREPROD* edges could be changed, and only in a way that does not create a new collision.

When the PC algorithm is applied to their correlations with the common sense time order using a significance level of .1 for tests of zero partial correlations, the output is the graph on the left side of figure 9, which we show alongside the Rodgers and Maranto model.

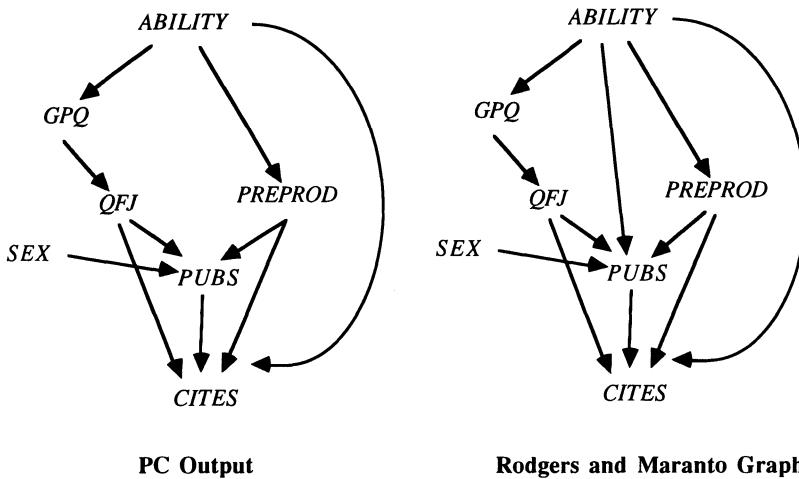


Figure 9

All but one of the edges in the Rodgers and Maranto model is produced instantaneously from the data and common sense knowledge of the domain—the time order of the variables. EQS gives this model a χ^2 of 13.58 with 11 degrees of freedom and $p = .257$. If the search procedure is repeated using .05 as the significance level, the program deletes the $PREPROD \rightarrow PUBS$ edge. When that model is estimated and tested with the EQS program we find that χ^2 is 19.2 with 12 degrees of freedom and a p value of .08, figures that should be taken as estimates of fit rather than of the probability of error.

Any claim that social scientific theory--other than common sense--is required to find the essentials of the Rodgers and Maranto model is clearly false. Nor do the preliminary results of Rodgers and Maranto's search afford any reason for confidence in social scientific theory. In contrast, we know a good deal about the reliability and limitations of the PC algorithm. The entire study with TETRAD and EQS takes a few minutes. A slight variant of the model is obtained using the SGS algorithm rather than the PC algorithm.

5.8.2 Education and Fertility

Rindfuss, Bumpass and St. John (1980) were interested in the mutual influence in married women of education at time of marriage (*ED*) and age at which a first child is born (*AGE*). On theoretical grounds they argue at length for the model on the left in figure 10, where the regressors from top to bottom are as follows:

<i>DADSO</i> =	father's occupation
<i>RACE</i> =	race
<i>NOSIB</i> =	absence of siblings
<i>FARM</i> =	farm background
<i>REGN</i> =	region of the United States
<i>ADOLF</i> =	presence of two adults in the subject's childhood family
<i>REL</i> =	religion
<i>YCIG</i> =	cigarette smoking
<i>FEC</i> =	whether the subject had a miscarriage.

Regressors are correlated. The sample size is 1766, and the covariances are given below.

<i>DADSO</i>	<i>RACE</i>	<i>NOSIB</i>	<i>FARM</i>	<i>REGN</i>	<i>ADOLF</i>	<i>REL</i>	<i>YCIG</i>	<i>FEC</i>	<i>ED</i>	<i>AGE</i>
456.676										
-.9201	.089									
-15.825	.1416	9.212								
-3.2442	.0124	.3908	.2209							
-1.3205	.0451	.2181	.0491	.2294						
-.4631	.0174	-.0458	-.0055	.0132	.1498					
.4768	-.0191	.0179	-.0295	-.0489	-.0085	.1772				
-0.3143	.0031	.0291	.0096	.0018	.0089	-.0014	.1170			
.2356	.0031	.0018	-.0045	-.0039	.0021	-.0003	.0009	.0888		
18.66	-.1567	-2.349	-.2052	-.2385	-.1434	-.0119	-.1380	.0267	5.5696	
16.213	-.2305	-1.4237	-.2262	-.3458	.1752	.1683	.1702	.2626	3.6580	16.6832

Apparently to their surprise, the investigators found on estimating coefficients that the *AGE* \rightarrow *ED* parameter is zero. Given the prior information that *ED* and *AGE* are not causes of the other variables, the PC algorithm (using .05 significance level for tests) directly finds the model on the right in figure 10, where connections among the regressors are not pictured.

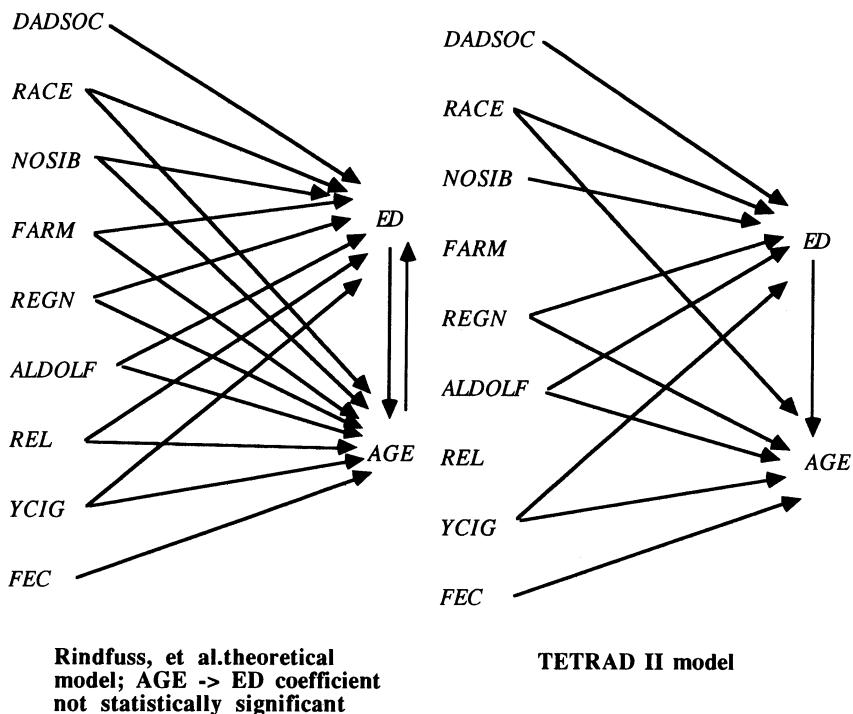


Figure 10

5.8.3 The Female Orgasm

Bentler and Peeler (1979) obtained data from 281 female university undergraduates regarding personality and sexual response. They include the Eysenck Personality Inventory which measured neuroticism (*N*) and extraversion (*E*); a heterosexual behavior inventory (*HET*), a monosexual behavior inventory (*MONO*); a scale of negative attitudes towards masturbation (*ATM*) and an inventory of subjective assessments of coital and masturbatory experiences. Using factor analysis the investigators formed scales, thought to be unidimensional, from these responses, including two scales (*SCOR*) and (*SMOR*) from the subjective assessments of coital and masturbatory experiences.

The investigators were interested in two hypotheses: (1) subjective orgasm responses in masturbation and coitus are due to distinct internal processes; (2) extraversion, neuroticism and

attitudes toward masturbation have no direct effect on orgasmic responsiveness, measured by *SCOR* and *SMOR*, but influence that phenomenon only through the history of the individual's sexual experience measured by *HET* and *MONO*.

We will not discuss the formation of the scales in this case, since the only data presented are the correlations of the scales and inventory scores, which are:

<i>E</i>	<i>N</i>	<i>ATM</i>	<i>HET</i>	<i>MONO</i>	<i>SCOR</i>	<i>SMOR</i>
1.0						
-.132	1.0					
.009	-.136	1.0				
.22	-.166	.403	1.0			
-.008	.008	.598	.282	1.0		
.119	-.076	.264	.514	.176	1.0	
.118	-.137	.368	.414	.336	.338	1.0

Bentler and Peeler offer two linear models to account for the correlations. The models and the probability values for the associated asymptotic χ^2 are shown in figure 11.

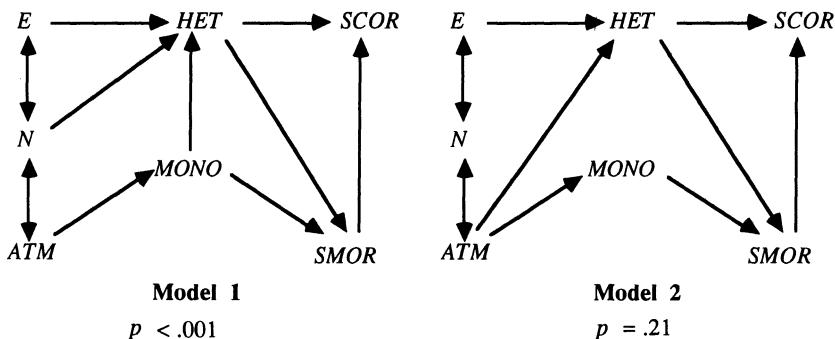


Figure 11

Only the second model saves the phenomena. The authors write that

...it proved possible to develop a model of orgasmic responsiveness consistent with the hypothesis that extraversion (*e*), neuroticism (*n*), and attitudes toward masturbation (*atm*) influence orgasmic responsiveness only through the effect

these variables have on heterosexual (*het*) and masturbatory (*mono*) experience. Consequently hypothesis 2 appears to be accepted (p.419).

The logic of the argument is not apparent. As the authors note "it must be remembered that other modes (sic) could conceivably also be developed that would equally well describe the data." (p. 419). But if the data could equally well be described, for example, by a model in which *ATM* has a direct effect on *SCOR* or on *SMOR*, there is no reason why hypothesis 2 should be accepted. Using the PC algorithm, one readily finds such a model.

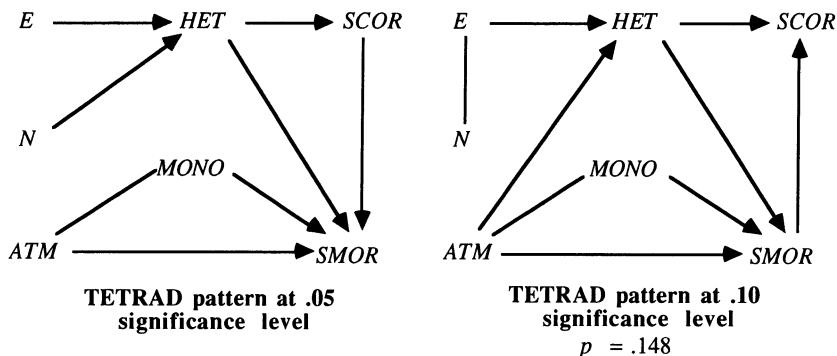


Figure 12

The model on the right of figure 12 has an asymptotic χ^2 value of 17 with 12 degrees of freedom, with $p(\chi^2) = .148$.

The PC algorithm finds a model that cannot be rejected on the basis of the data and that postulates a direct effect of attitude toward masturbation on orgasmic experience during masturbation, contrary to Bentler and Peeler.

5.8.4 The American Occupational Structure

Blau and Duncan's (1967) study of the American occupational structure has been praised by the National Academy of Sciences as an exemplary piece of social research and criticized by one statistician (Freedman 1983a) as an abuse of science. Using a sample of 20,700 subjects, Blau and Duncan offered a preliminary theory of the role of education (*ED*), first job (*J₁*), father's education (*FE*), and father's occupation (*FO*) in determining one's occupation (*OCC*) in 1962.

They present their theory in the following graph, in which the undirected edge represents an unexplained correlation:

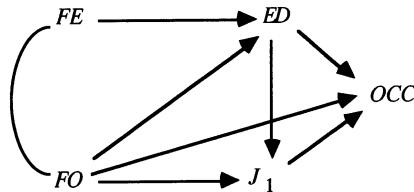


Figure 13

Blau and Duncan argue that the dependencies are linear. Their salient conclusions are that father's education affects occupation and first job only through the father's occupation and the subject's education.

Blau and Duncan's theory was criticized by Freedman as arbitrary, unjustified, and statistically inadequate (Freedman 1983a). Indeed, if the theory is subjected to the asymptotic χ^2 likelihood ratio test of the EQS (Bentler 1985) or LISREL (Joreskog and Sorbom, 1984) programs the model is decisively rejected ($p < .001$), and Freedman reports it is also rejected by a bootstrap test.

If the conventional .05 significance level is used to test for vanishing partial correlations, given a common sense ordering of the variables by time, from Blau and Duncan's covariances the PC algorithm produces the following graph:

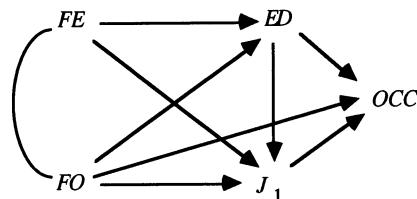


Figure 14

In this case every collider occurs in a triangle and there are no unshielded colliders. The data therefore do not determine the directions of the causal connections, but the time order of course determines the direction of each edge. We emphasize that the adjacencies are produced by the

program entirely from the data, without any prior constraints. The model shown passes the same likelihood ratio test with $p > .3$.

The algorithm adds to Blau and Duncan's theory a direct connection between FE and J_1 . The connection between FE and J_1 would only disappear if the significance level used to test for vanishing partial correlations were .0002. To determine a collection of vanishing partial correlations that are consistent with a directed edge from FE to OCC in 1962 one would have to reject hypotheses of vanishing partial correlations at a significance level greater than .3. The conditional independence relations found in the data at a significance level of .0001 are faithful to Blau and Duncan's directed graph.

Freedman argues that in the American population we should expect that the influences among these variables differ from family to family, and therefore that the assumption that all units in the population have the same structural coefficients is unwarranted. A similar conclusion can be reached in another way. We noted in Chapter 3 that if a population consists of a mixture of subpopulations of linear systems with the same causal structure but different variances and linear coefficients, then unless the coefficients are independently distributed or the mixture is in special proportions, the population correlations will be different from those of any of the subpopulations, and variables independent in each subpopulation may be correlated in the whole. When subpopulations with distinct linear structures are mixed and these special conditions do not obtain, the directed graph found from the correlations will typically be complete. We see that in order to fit Blau and Duncan's data we need a graph that is only one edge short of being complete.

The same moral is if anything more vivid in another linear model built from the same empirical study by Duncan, Featherman and Duncan (1972). They developed the following model of socioeconomic background and occupational achievement, where FE signifies father's education, FO father's occupational status, SIB the number of the respondent's siblings, ED the respondent's education, OCC the respondent's occupational status and INC the respondent's income.

In this case the double headed arrows merely indicate a residual correlation. The model has four degrees of freedom, and entirely fails the EQS likelihood ratio test (χ^2 is 165). When the correlation matrix is given to the TETRAD II program along with an obvious time ordering of the variables, the PC algorithm produces a complete graph.

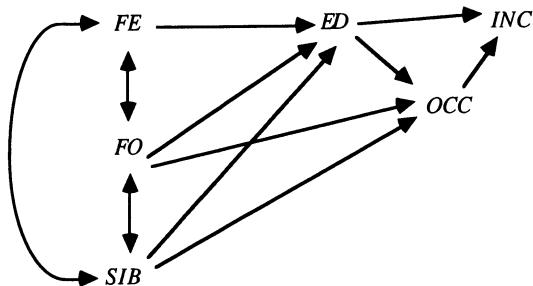


Figure 15

5.8.5 The ALARM Network

Recall the ALARM network developed to simulate causal relations in emergency medicine:

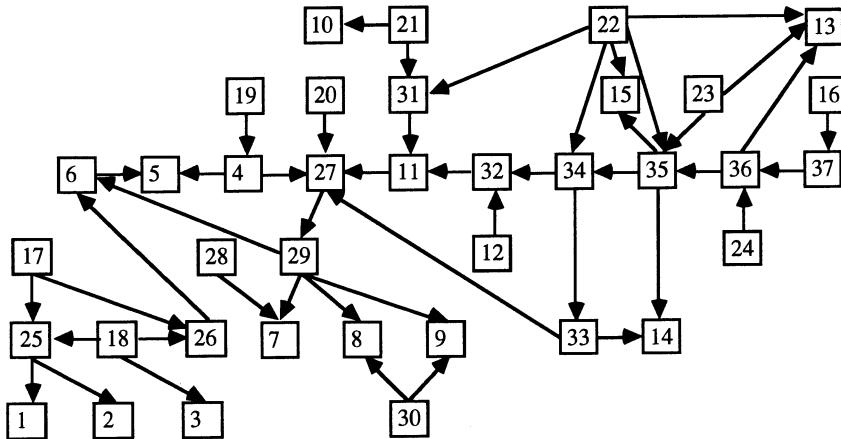


Figure 16

The SGS and PC* algorithms will not run on a problem this large. We have applied the PC algorithm to a linear version of the ALARM network. Using the same directed graph, linear coefficients with values between .1 and .9 were randomly assigned to each directed edge in the graph. Using a joint normal distribution on the variables of zero indegree, three sets of simulated data were generated, each with a sample size of 2,000. The covariance matrix and

sample size were given to a version of the TETRAD II program with an implementation of the PC-1 algorithm. This implementation takes as input a covariance matrix, and it outputs a pattern. No information about the orientation of the variables was given to the program. Run on a Decstation 3100, for each data set the program required less than fifteen seconds to return a pattern. In each trial the output pattern omitted two edges in the ALARM network; in one of the cases it also added one edge that was not present in the ALARM network.

In a related test, another ten samples were generated, each with 10,000 units. The results were scored as follows: We call the pattern the PC algorithm would generate given the population correlations the **true pattern**. We call the pattern the algorithm infers from the sample data the **output pattern**. An **edge existence error of commission (Co)** occurs when any pair of variables are adjacent in the output pattern but not in the true pattern. If an edge e between A and B occurs in both the true and output patterns, there is an **edge direction error of commission** when e has an arrowhead at A in the output pattern but not in the true pattern, (and similarly for B .) **Errors of omission (Om)** are defined analogously in each case. The results are tabulated as the average over the trial distributions of the ratio of the number of actual errors to the number of possible errors of each kind. The results at sample size 10,000 are summarized below:

#trials	%Edge Existence Errors		%Edge Direction Errors	
	Commission	Omission	Commission	Omission
10	.06	4.1	17	3.5

For similar data from a similarly connected graph with 100 variables, for ten trials the PC-1 algorithm required an average of 134 seconds and the PC-3 algorithm required an average of 16 seconds.

Herskovits and Cooper (1990) generated discrete data for the ALARM network, using variables with two, three and four values. Given their data, the TETRAD II program with the PC algorithm reconstructs almost all of the undirected graph (it omitted two edges in one trial; and in another also added one edge) and orients most edges correctly. In most orientation errors an edge was oriented in both directions. Broken down by the same measures as were used for the linear data from the same network, (with simulated data obtained from Herskovits and Cooper at sample size 10,000) the results are:

trial	%Edge Existence Errors		%Edge Direction Errors	
	Commission	Omission	Commission	Omission
1	0	4.3	27.1	10.0
2	0.2	4.3	5.0	10.4

5.8.6 Virginity

A retrospective study by Reiss, Banwart and Foreman (1975) considered the relationship among a sample of undergraduate females between a number of attitudes, including attitude toward premarital intercourse, use of a university contraceptive clinic, and virginity. Two samples were obtained, one of women who had used the clinic and one of women who had not; the samples did not differ significantly in relevant background variables such as age, education, parental education, and so on. Fienberg gives the cross-classified data for three variables: Attitude toward extramarital intercourse (E) (always wrong; not always wrong); virginity (V) and use of the contraceptive clinic (C) (used; not used). All variables are binary. The PC and SGS procedures immediately produces the following pattern:

$$E \longrightarrow V \longrightarrow C$$

Figure 17

which is consistent with any of the orientations of the edges that do not produce a collision at V . One sensible interpretation is that attitude affects sexual behavior which causes clinic use. Fienberg (1977) obtains the same result with log linear methods.

5.8.7 The Leading Crowd

Coleman (1964) describes a study in which 3398 schoolboys were interviewed twice. At each interview each subject was asked to judge whether or not he was a member of the "leading crowd" and whether his attitude toward the leading crowd was favorable or unfavorable. The data have been reanalyzed by Goodman (1973a, b) and by Fienberg (1977). Using Fienberg's notation, let A and B stand for the questions at the first interview and C and D stand for the corresponding questions at the second interview. The data are given by Fienberg as follows:

		Second Interview				
Membership Attitude		+	+	-	-	
		+	-	+	-	
Membership Attitude						
First Interview	+	+	458	140	110	49
	+	-	171	182	56	87
	-	+	184	75	531	281
	-	-	85	97	338	554

Fienberg summarizes his conclusions after a log-linear analysis in the path diagram in figure 18. He does not explain what interpretation is to be given to the double-headed arrow:

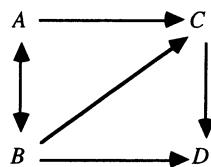


Figure 18

When the PC algorithm is told that C and D occur after A and B , with the usual .05 significance level for tests the program produces the pattern in figure 19:

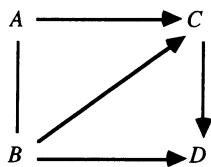


Figure 19

Orienting the undirected edge in the PC model as a directed edge from A to B produces expected values for the various cell counts that are almost identical with Fienberg's (p. 127) expected

counts.⁷ Note, however, this is a nearly complete graph, which may indicate that the sample is a mixture of different causal structures.

5.8.8 Influences on College Plans

Sewell and Shah (1968) studied five variables from a sample of 10,318 Wisconsin high school seniors. The variables and their values are:

<i>SEX</i>	[male = 0, female = 1]
<i>IQ</i> = Intelligence Quotient,	[lowest = 0, highest = 2]
<i>CP</i> = college plans	[yes = 0, no = 1]
<i>PE</i> = parental encouragement	[low = 0, high = 1]
<i>SES</i> = socioeconomic status	[lowest = 0, highest = 3]

They offer the following causal hypothesis:

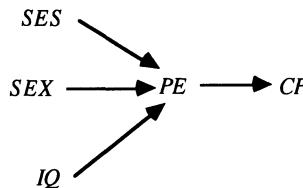


Figure 20

The data were reanalyzed by Fienberg (1977), who attempted to give a causal interpretation using log-linear models, but found a model that could not be given a graphical interpretation.

Given prior information that orders the variables by time as follows

- | | | | |
|---|------------|-----------|------------|
| 1 | <i>SEX</i> | | |
| 2 | <i>IQ</i> | <i>PE</i> | <i>SES</i> |
| 3 | <i>CP</i> | | |

⁷The small differences are presumably attributable to round-off errors.

so that later variables cannot be specified to be causes of earlier variables, the output with the PC algorithm is the structure:

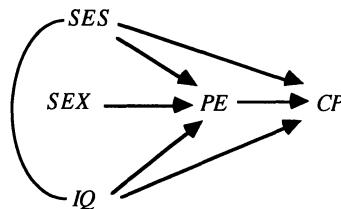
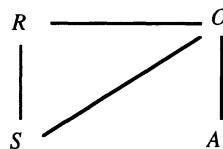


Figure 21

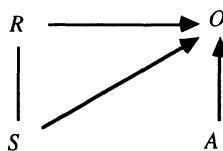
The program cannot orient the edge between *IQ* and *SES*. It seems very unlikely that the child's intelligence causes the family socioeconomic status, and the only sensible interpretation is that *SES* causes *IQ*, or they have a common unmeasured cause. Choosing the former, we have a directed graph whose joint distribution can be estimated directly from the sample. We find, for example, that the maximum likelihood estimate of the probability that males have college plans (*CP*) is .35, while the probability for females is .31. Judged by this sample the probability a child with low *IQ*, no parental encouragement (*PE*) and low socioeconomic status (*SES*) plans to go to college is .011; more distressing, the probability that a child otherwise in the same conditions but with a high *IQ* plans to go to college is only .124.

5.8.9 Abortion Opinions

Christensen (1990) illustrates log-linear model selection and search procedures with a data set whose variables are Race (*R*) [white, non-white], Sex (*S*), Age (*A*) [six categories] and Opinion (*O*) on legalized abortion (supports, opposes, undecided). Forward selection procedures require fitting 43 log-linear models. A backwards elimination method requires 22 fits; a method due to Aitkin requires 6 fits; another backwards method due to Wermuth requires 23 fits. None of these methods would work at all on large variable sets. Christensen suggests that the "best" log-linear model is an undirected conditional independence graphical model whose maximal cliques are [*RSO*] and [*OA*]. This is shown in figure 22.

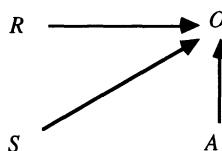
**Figure 22**

Subsequently, Christensen proposes a recursive causal model (in the terminology of Kiiveri and Speed, 1982) for the data. He suggests on substantive grounds a mixed graph

**Figure 23**

and says "The undirected edge between R and S ...represents an interaction between R and S ." Figure 23 is not a causal model in the sense we have described. It can be interpreted as a *pattern* representing the equivalence class of causal graphs whose members are the two orientations of the $R - S$ edge, but R and S in Christensen's data are very nearly independent .

This example is small enough to use the PC* algorithm, which with significance level .05 for independence tests gives exactly figure 24. Assuming faithfulness, the statistical hypothesis of figure 24 is inconsistent with the independence of $\{R, S\}$ and A conditional on O , required by the log-linear model of figure 22.

**Figure 24**

At a slightly lower significance level (.01) R and O are judged independent, and the same algorithm omits the $R \rightarrow O$ connection. On this data with significance level .05 the PC algorithm also produces the graph of figure 24 but with the $R \rightarrow O$ connection omitted. The difference in the outputs of the PC* and PC algorithms occur in the following way. Both algorithms produce at an intermediate stage the undirected graph underlying figure 24. In that undirected graph A does not lie on any undirected path between R and O . For that reason, the PC* algorithm never tests the conditional independence of R and O on A , and leaves the $R - O$ edge in. In contrast, the PC algorithms does test the conditional independence of R and O on A , with a positive result, and removes the $R - O$ edge.

5.8.10 Simulation Tests with Random Graphs

In order to test the speed and the reliability of the algorithms discussed in this chapter, we have tested the algorithms SGS, PC-1, PC-2, PC-3, and IG on a large number of simulated examples. The graphs themselves, the linear parameters, and the samples were all pseudo-randomly generated. This section describes the sample generation procedures for both linear and discrete data and gives simulation results for the linear case. Simulation results with discrete data are considered in the chapter on regression.

The average degree of the vertices in the graphs considered are 2, 3, 4, or 5; the number of variables is 10 or 50; and the sample sizes are 100, 200, 500, 1000, 2000, and 5000. For each combination of these parameters, 10 graphs were generated, and a single distribution obtained faithful to each graph, and a single sample taken from each such distribution.

Because of its computational limitations, the SGS algorithm was tested only with graphs of 10 variables.

5.8.10.1 Sample Generation

All pseudo-random numbers were generated by the UNIX "random" utility. Each sample is generated in three stages:

- (i) The graph is pseudo-randomly generated.
- (ii) The linear coefficients (in the linear case) or the conditional probabilities (in the discrete case) are pseudo-randomly generated.
- (iii) A sample for the model is pseudo-randomly generated.

We will discuss each of these steps in more detail.

(i) The input to the random graph generator is an average degree and the number of variables. The variables are ordered so that an edge can only go from a variable lower in the order to a variable higher in the order, eliminating the possibility of cycles. Since some of the procedures use a lexicographic ordering, variable names were then randomly scrambled so that no systematic lexicographic relations obtained among variable pairs connected by edges. Each variable pair is assigned a probability p equal to

$$\frac{\text{average degree}}{\text{number of variables} - 1}$$

For each variable pair a number is drawn from a uniform distribution over the interval 0 to 1. The edge is placed in the graph if and only if the number drawn is less than or equal to p .⁸

(ii) For simulated continuous distributions, an "error" variable was introduced for each endogenous variable and values for the linear coefficients between .1 and .9 were generated randomly for each edge in the graph. For the discrete case, a range of values of variables is selected by hand, and for each variable taking n values, the unit interval is divided into n sub-intervals by random choice of cut-off points. A distribution (e.g., uniform) is then imposed on the unit interval.

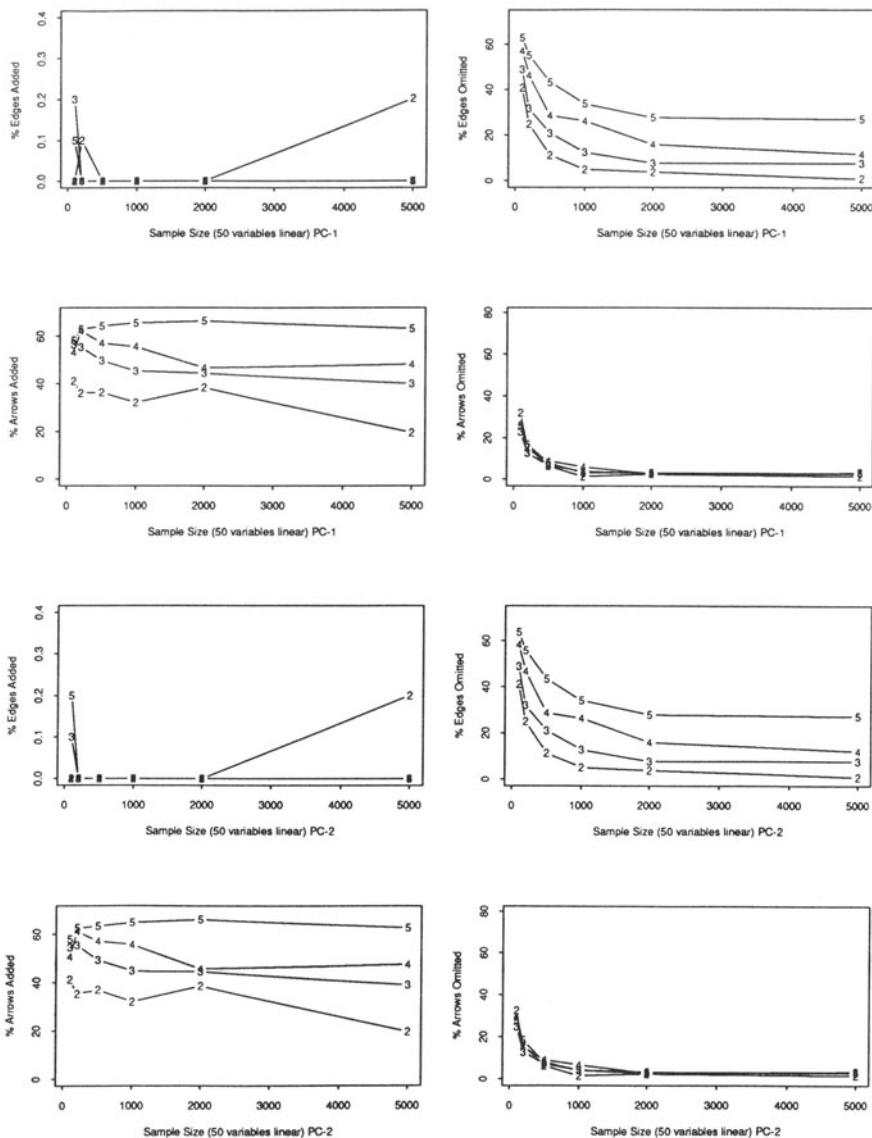
(iii) In the discrete case for each such distribution produced, each sample unit is obtained by generating, for each exogenous variable, a random number between 0 and 1.0 according to the distribution and assigning the variable value according to the category into which the number falls. Values for the endogenous variables were obtained by choosing a value randomly with probability given by the conditional probabilities on the obtained values of the parents of the variable. In the linear case, the exogenous variables--including the error terms--were generated independently from a standard normal distribution, and values of endogenous variables were computed as linear functions of their parents.

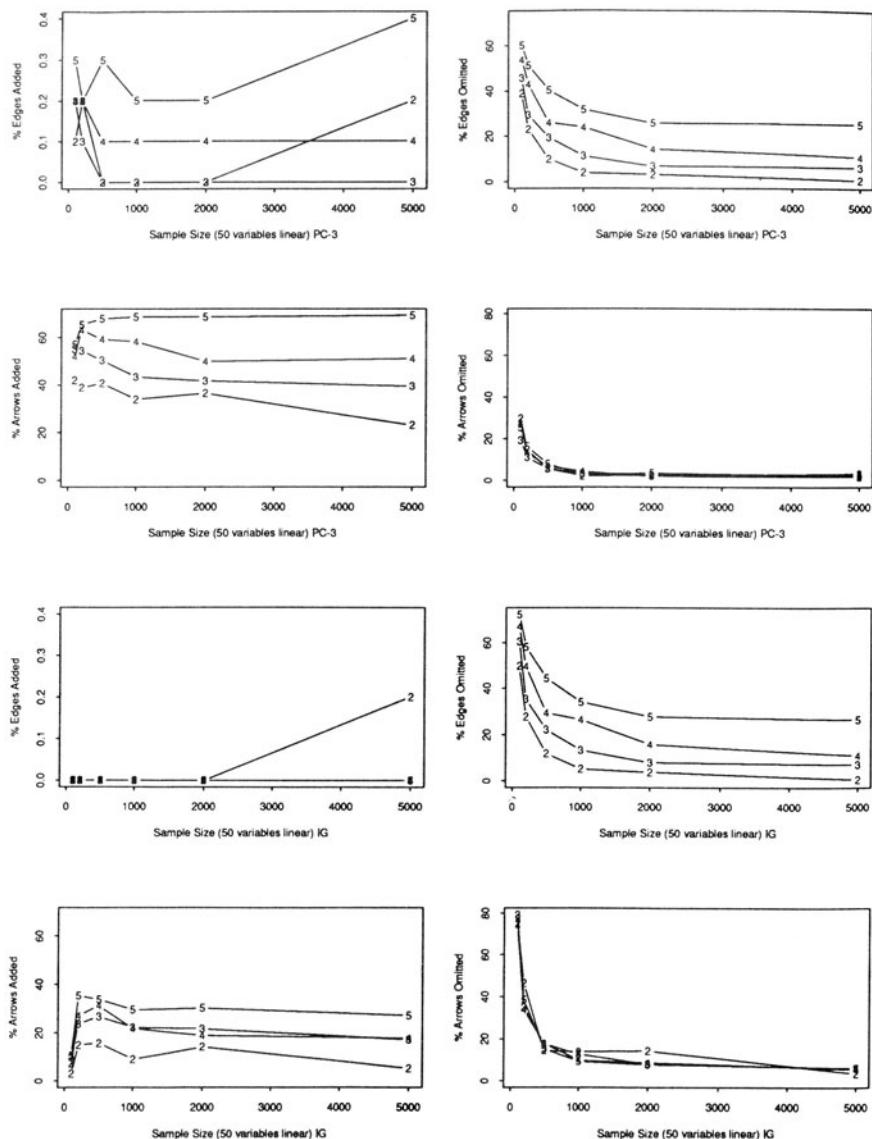
⁸We do not know whether this method of graph generation produces "realistic" graphs. One feature of some of the graphs generated in this fashion that may not be desirable is the existence of isolated variables. An informal examination showed topologies not unlike the Alarm network.

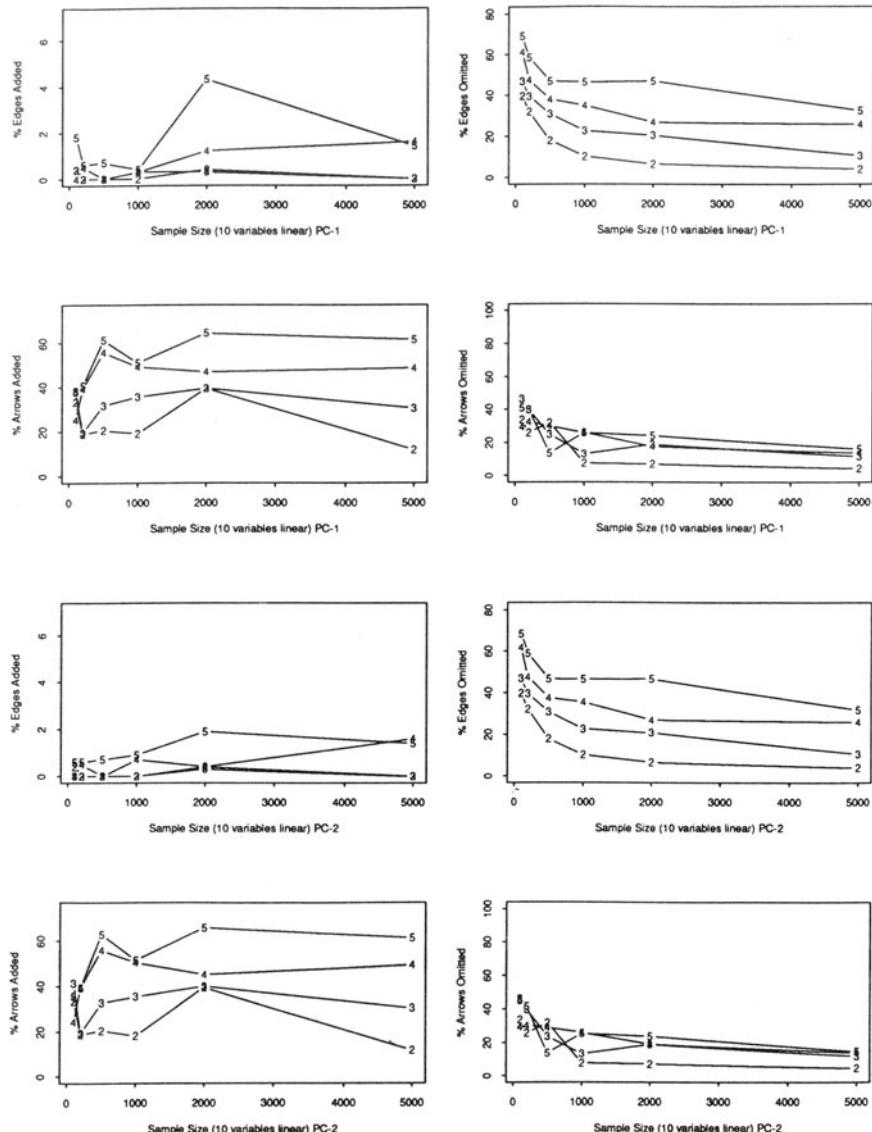
5.8.10.2 Results

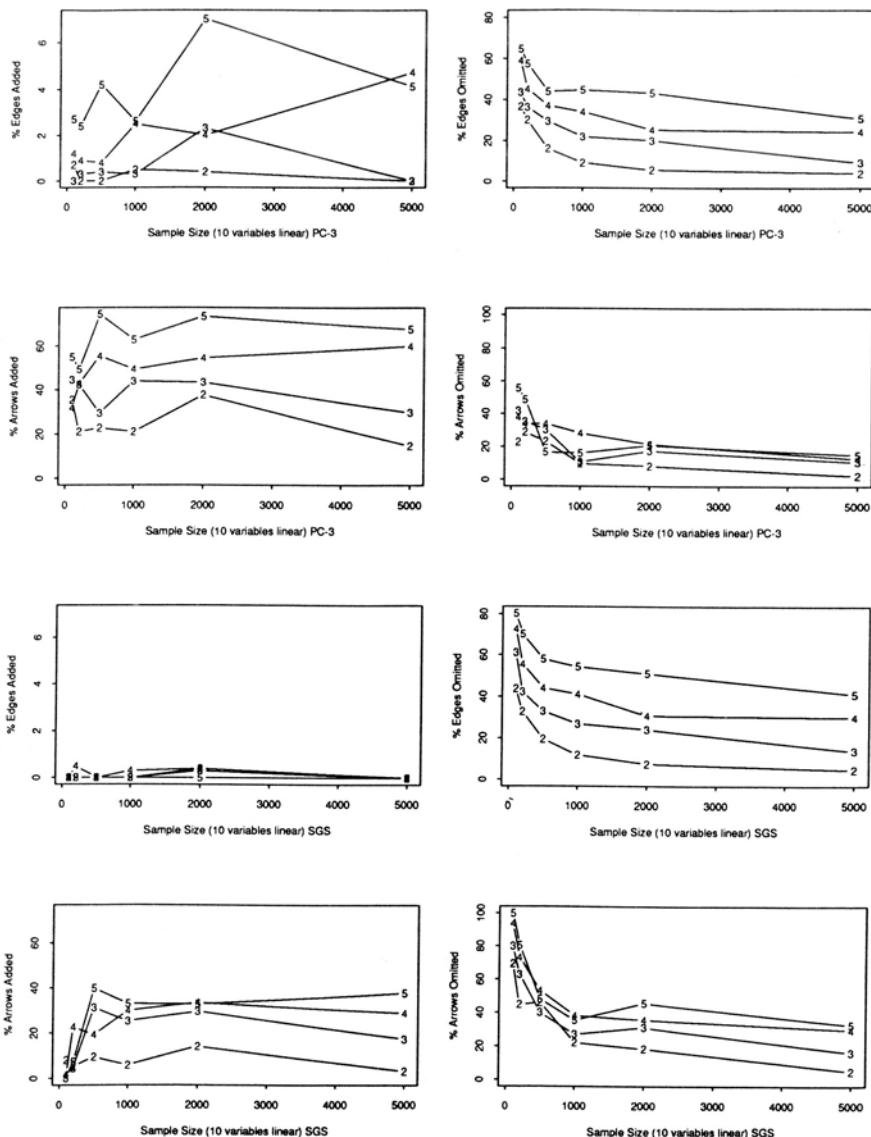
As before, reliability has several dimensions. A procedure may err by omitting undirected edges in the true graph or by including edges--directed or undirected--between vertices that are not adjacent in the true graph. For an edge that is not in the true graph, there is no fact of the matter about its orientation, but for edges that are in the true graph, a procedure may err by omitting an arrowhead in the true graph or by including an arrowhead not in the true graph. We count errors in the same way as in section 5.8.5.

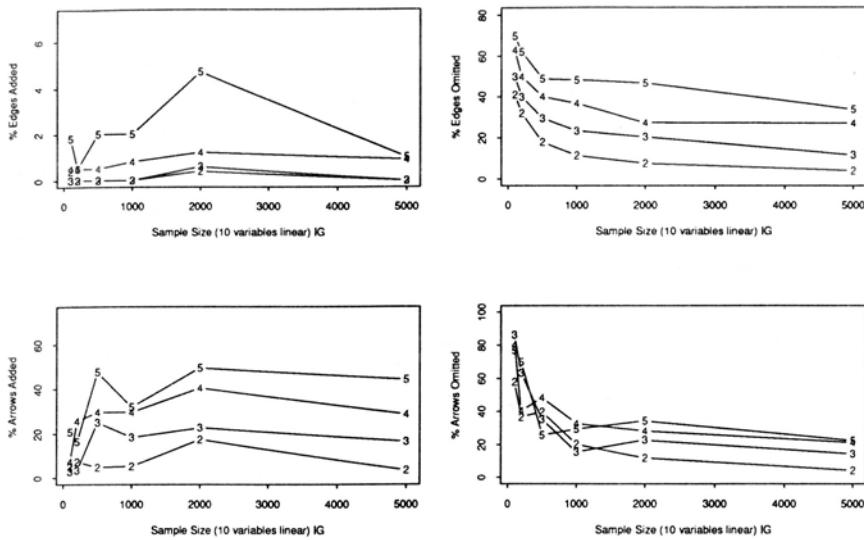
Each of the procedures was run using a significance level of .05 on all trials. The five procedures tested are not equally reliable or equally fast. The SGS algorithm is much the slowest, but in several respects it proves reliable. The graphs on the following pages show the results. Each point on the graph is a number, which represents the average degree of the vertices in the directed graphs generating the data. We plot the run times and reliabilities of the PC-1 PC-2, PC-3, IG, and SGS algorithms against sample size for data from linear models based on randomly generated graphs with 10 variables, and similarly the reliabilities of the first four of these algorithms for linear models based on randomly generated graphs with 50 variables. In each case the results are plotted separately for graphs of degree 2, 3, 4, and 5.











The following qualitative conclusions can be drawn.

The rates of arrow and edge omission decrease dramatically with sample size up to about sample size 1000; after that the decreases are much more gradual.

The rates of arrow and edge commission vary much less dramatically with sample size than do the rates of arrow and edge omission.

As the average degree of the variables increases, the average error rates increase in a very roughly linear fashion, but the PC-2 algorithm tends to be less reliable than the other algorithms with respect to edge omissions when the average degree of the graph is high.

The PC-1, PC-3, IG, and SGS algorithms have compensating virtues and disadvantages. None of the procedures are reliable on all dimensions when the graphs are not sparse. One reliable dimension is the addition of edges: If two vertices are not adjacent in the true graph, there is very little chance they will be mistakenly output by any of these four procedures, no matter what the average degree of the graph and no matter what the sample size.

In contrast, at high average degree and low sample sizes the output of each of the procedures tends to omit over 50% of the edges in the true graph. At large sample sizes and low average degree only a few percent of the true edges are omitted, but with high average degree the percentage of edges omitted even at large sample sizes is significant. For example, at sample size 5000 and average degree 5, PC-1 omits over 30% of the edges in the true graph.

Arrow commission errors are much more common than edge commission errors. If an arrow does not occur in a graph, there is a considerable probability for any of the procedures that the arrow will be output, unless the sample size is large and the true graph is of low degree. For 10 variables with average degree about 2 and sample sizes of 1,000 or more, the SGS and IG algorithms are quite reliable, with errors of commission for arrows around 6%. Under the same conditions the error rates for the PC-1 and PC-2 algorithms run about 20%. In the case of the SGS algorithm, these relations are reversed for the question of arrow omission--if an arrow occurs in the true graph, what is the chance that the procedure will fail to include the arrow in its output? The answer is about 8% for PC-1 and PC-2 and about 20% for the SGS procedure. The IG algorithm, while much less reliable for arrow omissions at low sample sizes, is only slightly more unreliable at high sample sizes.

The return time of the PC-3 algorithm is dramatically smaller than the other algorithms. Its run time also does not increase as sharply with average degree, but the procedure does produce many more edge commission errors as the average degree increases.

The results suggest that the programs can reasonably be used in various ways according to the size of the problem, the questions one wants answered, and the character of the output. Roughly the same conclusions about reliability can be expected for the discrete case, but with lower absolute reliabilities. For larger number of variables the same patterns should hold, save that the SGS algorithm cannot be run at all.

More research needs to be done on local "repairs" to the graphs generated by these procedures, especially for edge omission errors and arrow commission errors. In order for the method to converge to the correct decisions with probability 1, the significance level used in making decisions should decrease as the sample sizes increase, and the use of higher significance levels (e.g., .2 at samples sizes less than 100, and .1 at sample sizes between 100 and 300) may improve performance at small sample sizes.

5.9 Conclusion

This chapter describes several algorithms that can reliably recover sparse causal structures even for quite large numbers of variables, and illustrates their application. The algorithms have each been implemented using tests for conditional independence in the discrete case and for vanishing partial correlations in the linear case. We make no claim that these uses of tests to decide relevant probability relations is optimal, but any improvements in the statistical decision methods can be prefixed to the algorithms. With the exception of the PC* and SGS algorithms, the procedures described are feasible for large numbers of variables so long as the true causal graphs are sparse.

The algorithms we have described scarcely exhaust the possibilities, and a number of very simple alternative procedures should work reasonably well, at least for finding adjacency relations in the causal graph.

5.10 Background Notes

The idea of discovery problems is already contained in the notion of an estimation problem, and the requirement that an estimator be consistent is essentially a demand that it solve a particular kind of discovery problem. An extension of the idea to general non-statistical settings was proposed by Putnam (1965) and independently by Gold (1965) and has subsequently been extensively developed in the literature of computer science, mathematical linguistics and logic (Osherson, Stob and Weinstein, 1986).

A more or less systematic search procedure for causal/statistical hypotheses can be found in the writings of Spearman (1904) and his students early in this century. A Bayesian version of stepwise search was proposed by Harold Jeffreys (1957). Thurstone's (1935) factor analysis inaugurated a form of algorithmic search separated from any precise discovery problem: Thurstone did not view factor analysis as anything more than a device for finding simplifications of the data, and a similar view has been expressed in many subsequent proposals for statistical search. The vast statistical literature on search has focused almost exclusively on optimizing fitting functions.

The SGS algorithm was proposed by Glymour and Spirtes in 1989, and appeared in Spirtes, Glymour and Scheines (1990c). Verma and Pearl (1990b) subsequently proposed a more efficient version that examines cliques. A version of the PC algorithm was developed by Spirtes and Glymour (1990). The version presented here contains an improvement suggested by Pearl and Verma in the efficiency of step C) of the algorithm. Bayesian discovery procedures have been studied in Herskovits' thesis (1992).

The maximum likelihood estimation procedure for "recursive causal models" was developed in Kiiveri's (1982) doctoral thesis. The mathematical properties of the structures are further described in Kiiveri, Speed and Carlin (1984).

Chapter 6

Discovery Algorithms without Causal Sufficiency

6.1 Introduction

The preceding chapter complied with a common statistical fantasy, namely that in typical data sets it is known that no part of the statistical dependencies among measured variables are due to unmeasured common causes. We almost always fail to measure all of the causes of variables we do measure, and we often fail to measure variables that are causes of two or more measured variables. Any examination of collections of social science data gives the striking impression that variables in one study often seem relevant to those in other studies. Record keeping practices sometimes force econometricians to ignore variables in studies of one economy thought to have a causal role in studies of other economies (Klein, 1961). In many studies in psychometrics, social psychology and econometrics, the real variables of interest are unmeasured or measured only by proxies or "indicators." In epidemiological studies that claim to show that exposure to a risk factor causes disease, a burden of the argument is to show that the statistical association is not due to some common cause of risk factor and disease; since not everything imaginably relevant can be measured, the argument is radically incomplete unless a case can be made that unmeasured variables do not "confound" the association. If, as we believe, no reliable empirical study can proceed without considering whether relevant variables are unmeasured, then few published uncontrolled empirical studies are reliable.

In both experimental and non-experimental studies the unrecognized presence of unmeasured variables can lead to erroneous conclusions about the causal relations among the variables that are measured, and to erroneous predictions of the effects of policies that manipulate some of these variables. Until reliable, data-based methods are used to identify the presence or absence of unmeasured common causes, most causal inferences from observational data can be no more than guesswork at best and pseudo-science at worst. *Are such methods possible?* That question surely ought to be among the most important theoretical issues in statistics.

Statistical methods for detecting unmeasured common causes, or "confounding" in the terminology epidemiologists prefer, has been chiefly developed in psychometrics, where criteria for the existence and numbers of common causes have been sought since the turn of the century for special statistical models. The results include a literature on linear systems that contain criteria (e.g., Charles Spearman's (1904) vanishing tetrad differences) for latent variables that proved, however, to be neither necessary nor sufficient even assuming linearity. Criteria for two latent common causes were introduced by Kelley (1928), and related criteria are used in factor analysis, but they are not correct unless it is assumed that all statistical dependencies are due to unmeasured common causes. For problems in which the measured variables are discrete and their values a stochastic function of an unobserved continuous vector parameter θ , a number of criteria have been developed for the dimensionality of θ (Holland and Rosenbaum, 1986). Suppes and Zanotti (1981) showed that for discrete variables there always exists a formal latent variable model in which all measured variables are effects of an unmeasured common cause and all pairs of measured variables are independent conditional on the latent variable. Their argument assumes the model must satisfy only the Markov Condition; the result does not hold if it is required that the distributions be faithful.

Among epidemiologists, (Breslow and Day, 1980; Kleinbaum, Kupper and Morgenstern, 1982) the criteria introduced by the Surgeon General's report on Smoking and Health (1964) are sometimes still advocated as a means for deciding whether a statistical dependency between exposure to risk factor A and disease B is "causal," apparently meaning that A causes B and A and B have no common causes. The criteria include (i) increase in response with dosage; (ii) that the statistical dependency between a risk factor and disease be specific to particular disease subgroups and to particular conditions of risk exposure; (iii) that the statistical association be strong; (iv) that exposure to a risk factor precede the period of increased risk; (v) lack of alternative explanations.

Even in causally sufficient systems, where all common causes of measured variables are themselves measured, such criteria do not separate causes from correlated variables. They fail even to come to grips with the problem of unmeasured "confounders." Criterion (v) is an evasion; the problem in uncontrolled studies is exactly that there are too many alternative explanations of the data. Criterion (iv) is a banality which is of no use at all in deciding whether there are measured or unmeasured common causes at work. Criterion (iii) is defended on the grounds that "If an observed association is not causal, but simply the reflection of a causal association between some other factor and disease, then this latter factor must be more strongly related to disease (in terms of relative risk) than is the former factor," (Breslow and Day, 1980).

But the inference is incorrect: if there are two or more common causes, measured or not, none of them need be more strongly related to the disease than is the putative measured cause; and if A causes B and A and B *also* have a common cause, the latter need not be more strongly associated with B than is A . On behalf of Breslow and Day one might appeal to simplicity against all hypotheses of multiple common causes, but that would be an implausible claim in medical science, where multiple causal mechanisms abound. Nothing about the first two criteria separates the situation in which A and B have common causes from circumstances in which they do not.

In this chapter we present a more or less systematic account of how the presence of unmeasured common causes can mislead an investigator about causal relationships among measured variables, and of how the presence of unmeasured common causes can be detected. We deal with these questions separately for the general case and for the case in which all structures are linear. But the central aim of this chapter is to show how, assuming the Markov and Faithfulness conditions, in principle reliable causal inferences can be made from appropriate sample data without any prior knowledge as to whether the system of measured variables is causally sufficient.

6.2 The PC Algorithm and Latent Variables

A natural idea is that a slight modification of the PC algorithm will give correct information about causal structure even when unmeasured variables may be present. Suppose that P' is a distribution over V that is faithful to a causal graph, and P is the marginal of P' over O , properly included in V . We will refer to the members of O as measured or observed variables. As we have already seen, if there are unmeasured common causes, the output of the PC algorithm can include bi-directed edges of the form $A <> B$. We could interpret a bi-directed edge between A and B to mean that there is an unmeasured cause C that directly causes A and B relative to O . We modify the algorithm by using a "o" on the end of an arrow to indicate that it is not known whether an arrowhead should occur in that place. We use a "*" as a metasymbol to stand for any of the three kinds of endmarks that an arrow can have: EM (empty mark), ">", or "o".

Modified PC Algorithm:

A.) Form the complete undirected graph C on the vertex set V .

B.)

$n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in C such that $\text{Adjacencies}(C,X \setminus \{Y\})$ has cardinality greater than or equal to n , and a subset S of $\text{Adjacencies}(C,X \setminus \{Y\})$ of cardinality n , and if X and Y are d-separated given S delete edge $X - Y$ from C and record S in $\text{Sepset}(X,Y)$ and $\text{Sepset}(Y,X)$

until all ordered pairs of adjacent variables X and Y such that $\text{Adjacencies}(C,X \setminus \{Y\})$ has cardinality greater than or equal to n and all subsets of $\text{Adjacencies}(C,X \setminus \{Y\})$ of cardinality n have been tested for d-separation.

$n = n + 1$.

until for each ordered pair of adjacent vertices X, Y , $\text{Adjacencies}(C,X \setminus \{Y\})$ is of cardinality less than n .

C.) Let F be the graph resulting from step B). If X and Y are adjacent in F , orient the edge between X and Y as $X \rightarrowtail Y$.

D.) For each triple of vertices X, Y, Z such that the pair X, Y and the pair Y, Z are each adjacent in F but the pair X, Z are not adjacent in F , orient $X *-* Y *-* Z$ as $X *-> Y <-* Z$ if and only if Y is not in $\text{Sepset}(X,Z)$.

E.) repeat

If $A *-> B, B *-* C, A$ and B are not adjacent, and there is no arrowhead at B on $B *-* C$, then orient $B *-* C$ as $B \rightarrow C$.

If there is a directed path from A to B , and an edge between A and B , then orient the edge as $A *-> B$.

until no more edges can be oriented.

(When we say orient $X *-* Y$ as $X *-> Y$ we mean leave the same mark on the X end of the edge and put an arrowhead at the Y end of the edge.)

The result of this modification applied to the examples of the previous chapter is perfectly sensible. For example, in figure 1 we show both the model obtained from the Rodgers and Maranto data at significance level .1 by the PC algorithm and the model that would be obtained

by the modified PC algorithm from a distribution faithful to the the graph in the PC output. (In each case with the known time order of the variables imposed as a constraint.)

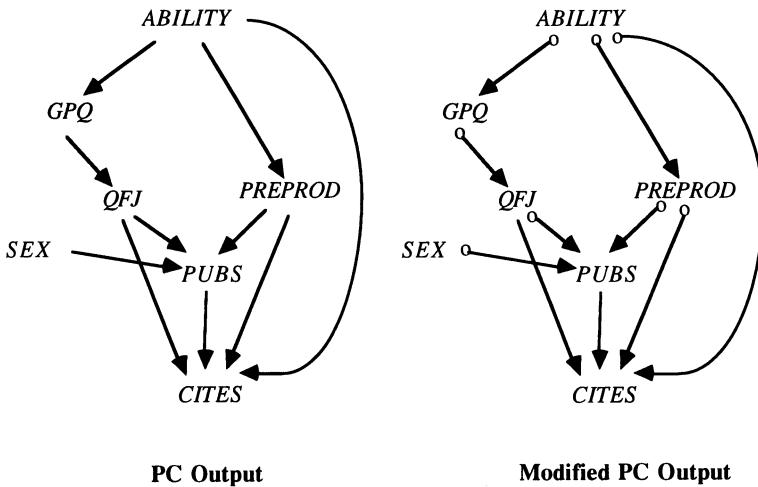


Figure 1

The output of the Modified PC Algorithm indicates that *GPQ* and *ABILITY*, for example, may be connected by an unmeasured common cause, but that *PUBS* is a direct cause of *CITES*, unconfounded by a common cause. Where a single vertex has "o" symbols for two or more edges connecting it with vertices that are not adjacent, a special restriction applies. *ABILITY*, for example has an edge to *GPQ* and to *PREPROD*, each with an "o" at the *ABILITY* end, and *GPQ* and *PREPROD* are not adjacent to one another. In that case the two "o" symbols cannot both be arrowheads. There cannot be an unmeasured cause of *ABILITY* and *GPQ* and an unmeasured cause of *ABILITY* and *PREPROD*, because if there were, *GPQ* and *PREPROD* would be dependent conditional on *ABILITY*, and the modified pattern entails instead that they are independent.

In many cases--perhaps most practical cases--in which the sampled distribution is the marginal of a distribution faithful to a graph with unmeasured variables, this simple modification of the PC algorithm gives a correct answer if the required statistical decisions are correctly made.

6.3 Mistakes

Unfortunately, this straightforward modification of the PC algorithm is not correct in general. An imaginary example will show why.

Everyone is familiar with a simple mistake occasioned by failing to recognize an unmeasured common cause of two variables, X , Y , where X is known to precede Y . The mistake is to think that X causes Y , and so to predict that a manipulation of X will change the distribution of Y . But there are more interesting cases that are seldom noticed, cases in which omitting a common cause of X and Y might lead one to think, erroneously, that some third variable Z directly causes Y . Consider an imaginary case:

A chemist has the following problem. According to received theory, which he very much doubts, chemicals A and B combine in a low yield mechanism to form chemical D through an intermediate C . Our chemist thinks there is another mechanism in which A and B combine to form D without the intermediate C . He wishes to do an experiment to establish the existence of the alternative mechanism. He can readily obtain reagents A and B , but available samples may be contaminated with varying quantities of D and other impurities. He can measure the concentration of the unstable alleged intermediate C photometrically, and he can measure the equilibrium concentration of D by standard methods. He can manipulate the concentrations of A and B , but he has no means to manipulate the concentration of C .

The chemist decides on the following experimental design. For each of ten different values of the concentration of A and B , a hundred trials will be run in which the reagents are mixed, the concentration of C is monitored, and the equilibrium concentration of D is measured. Then the chemist will calculate the partial correlation of A with D conditional on C , and likewise the partial correlation of B with D conditional on C . If there is an alternative mechanism by which A and B produce D without C , the chemist reasons, then there should be a positive correlation of A with D and of B with D in the samples in which the concentration of C is all the same; and if there is no such alternative mechanism, then when the concentration of C is controlled for, the concentrations of A , B on the one hand, and D on the other, should have zero correlation.

The chemist finds that the equilibrium concentrations of A , B on the one hand and of D on the other hand are correlated when C is controlled for--as they should be if A and B react to produce D directly--and he announces that he has established an alternative mechanism.

Alas within the year his theory is disproved. Using the same reagents, another chemist performs a similar experiment in which, however, a masking agent reacts with the intermediate C preventing it from producing D . The second chemist finds no correlation in his experiment between the concentrations of A and B and the concentration of D . What went wrong with the first chemist's procedure?

By substituting a statistical control for the manipulation of C the chemist has run afoul of the fact that the marginal probability distribution with unmeasured variables can give the appearance of a spurious direct connection between two variables. The chemist's picture of the mechanism is given in graph G_1 , and that is one way in which the observed results can be produced. Unfortunately, they can also be produced by the mechanism in graph G_2 , which is what happened in the chemist's case: impurities (F) in the reagents are causes of both C and D :

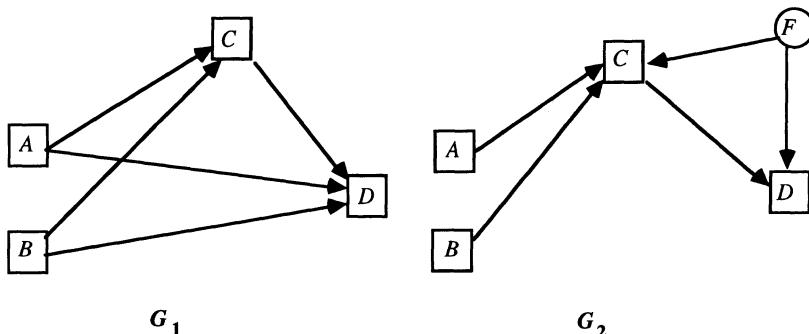


Figure 2

The general point is that a theoretical variable F acting on two measured variables C and D can produce statistical dependencies that suggest causal relations between A and D and between B and D that do not exist. For faithful distributions, if we use the SGS or PC algorithms, a structure such as G_2 will produce a directed edge from A to D in the output.

We can see the same point more analytically as follows: In a directed acyclic graph G over a set of variables \mathbf{V} , if A and D are adjacent in G , then A and D are not d-separated given any subset of $\mathbf{V} \setminus \{A,D\}$. Hence under the assumption of causal sufficiency, either A is a direct cause of D or D is a direct cause of A relative to \mathbf{V} if and only if A and D are independent conditional on no subset of $\mathbf{V} \setminus \{A,D\}$. However, if \mathbf{O} is not causally sufficient, it is not the case that if A and D are independent conditional on every subset of $\mathbf{O} \setminus \{A,D\}$ that either A is a direct cause of D relative to \mathbf{O} , or D is a direct cause of A relative to \mathbf{O} , or there is some latent variable F that is a common cause of both A and D .

This is illustrated by G_2 in figure 2, where $\mathbf{V} = \{A,B,C,D,F\}$ and $\mathbf{O} = \{A,C,D\}$. \mathbf{O} is not causally sufficient because F is a cause of both C and D which are in \mathbf{O} , but F itself is not in \mathbf{O} . A and D are not d-separated given any subset of $\mathbf{O} \setminus \{A,D\}$, so in any marginal of a distribution faithful to G , A and D are not independent conditional on any subset of $\mathbf{O} \setminus \{A,D\}$, and the modified PC algorithm would leave an edge between A and D . Yet A is not a direct cause of D relative to \mathbf{O} , D is not a direct cause of A relative to \mathbf{O} , and there is no latent common cause of A and D . The directed acyclic graph G_1 shown in figure 2, in which A is a direct cause of D , and in which there is a path from A to D that does not go through C , has the same set of d-separation relations over $\{A,C,D\}$ as does graph G_2 . Hence, given faithful distributions, they cannot be distinguished by their conditional independence relations alone.

A further fundamental problem with the simple modification of the PC algorithm described above is that if we allow bi-directed edges in the graphs constructed by the PC algorithm, it is no longer the case that if A and B are d-separated given some subset of \mathbf{O} , then they are d-separated given a subset of **Adjacencies**(A) or **Adjacencies**(B). Consider the graph in figure 3, where T_1 and T_2 are assumed to be unmeasured.

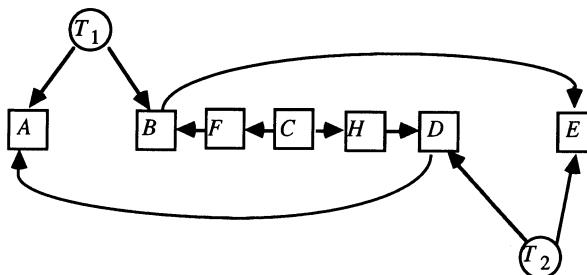


Figure 3

Among the measured variables, $\text{Parents}(A) = \{D\}$ and $\text{Parents}(E) = \{B\}$, but A and E are not d-separated given any subset of $\{B\}$ or any subset of $\{D\}$; the only sets that d-separate them are sets that contain F , C , or H . The Modified PC algorithm would correctly find that C , F , and H are not adjacent to A or E . It would then fail to test whether A and E are d-separated given any subset containing C . Hence it would fail to find that A and E are d-separated given $\{B,C,D\}$ and would erroneously leave A and E adjacent. This means that it is not possible to determine which edges to remove from the graph by examining only local features (i.e. the adjacencies) of the graph constructed at a given stage of the algorithm. Similarly, once bidirected edges are allowed in the output of the PC algorithm, it is not possible to extract all of the information about the orientation of edges by examining local features (i.e. pairs of edges sharing a common endpoint) of the graph constructed at a given stage of the algorithm.

Because of these problems, for full generality we must make major changes to the PC algorithm and in the interpretation of the output. We will show that there is a procedure, which we optimistically call the Fast Causal Inference (FCI) algorithm, that is feasible in large variable sets provided the true graph is sparse and there are not many bidirected edges chained together. The algorithm gives asymptotically correct information about causal structure when latent variables may be acting, assuming the measured distribution is the marginal of a distribution satisfying the Markov and Faithfulness conditions for the true graph. The FCI algorithm avoids the mistakes of the modified PC algorithm, and in some cases provides more information.

For example, with a marginal distribution over the boxed variables from the imaginary structure in figure 4, the modified PC algorithm gives the correct output shown in the first diagram in figure 5, whereas the FCI algorithm produces the correct and much more informative result in the second diagram in figure 5:

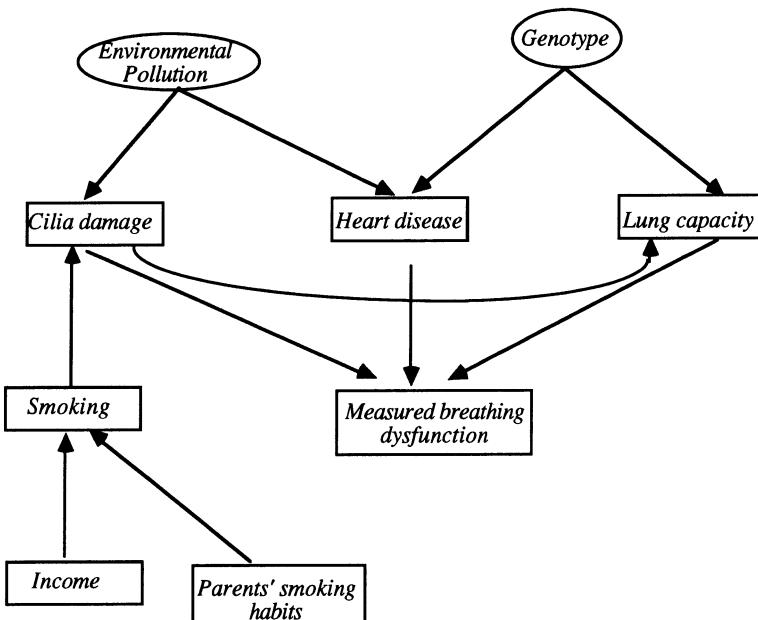


Figure 4

In figure 5, the double headed arrows indicate the presence of unmeasured common causes, and as in the modified PC algorithm the edges of the form $o \rightarrow$ indicate that the algorithm cannot determine whether the circle at one end of the edge should be an arrowhead. Notice that the adjacencies among the set of variables $\{Cilia\ damage, Heart\ disease, Lung\ capacity, Measured\ breathing\ dysfunction\}$ form a complete graph, but even so the edges can be completely oriented by the FCI algorithm.

The derivation of the FCI algorithm requires a variety of new graphical concepts and a rather intricate theory. We introduce Verma and Pearl's notions of an inducing path and an inducing path graph, and show that these objects provide information about causal structure. Then we consider algorithms that infer a class of inducing path graphs from the data.

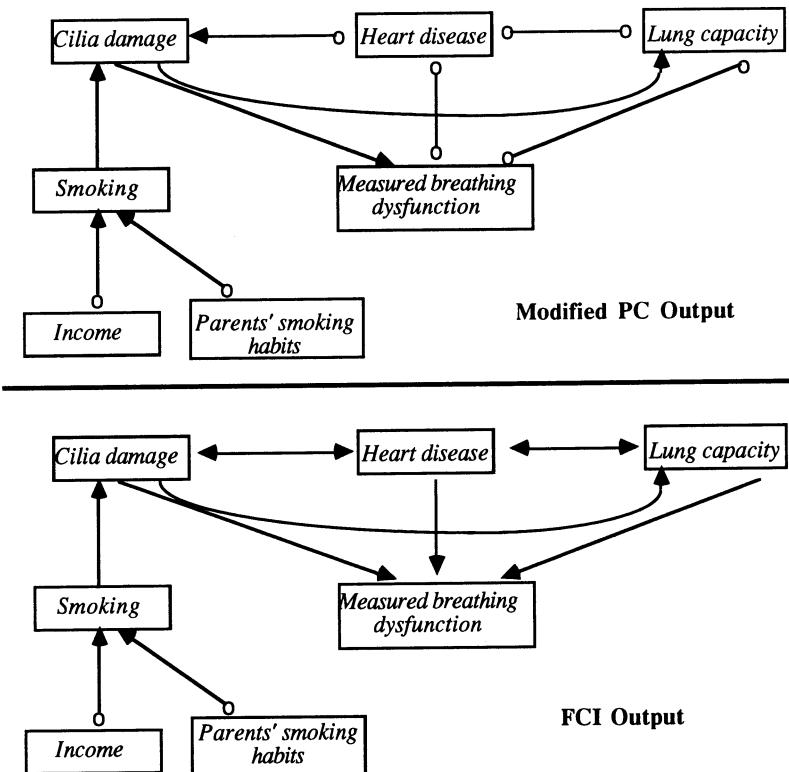
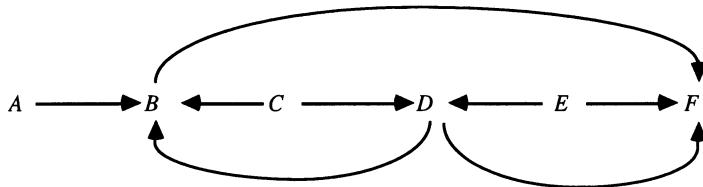


Figure 5

6.4 Inducing Paths

Given a directed acyclic graph G over a set of variables V , and O a subset of V , Verma and Pearl (1991) have characterized the conditions under which two variables in O are not d-separated given any subset of $O \setminus \{A, B\}$. If G is a directed acyclic graph over a set of variables V , O is a subset of V containing A and B , and $A \neq B$, then an undirected path U between A and B is an **inducing path relative to O** if and only if every member of O on U except for the endpoints is a collider on U , and every collider on U is an ancestor of either A or B . We will sometimes refer to members of O as **observed variables**.

Figure 6: Graph G_3

For example, in graph G_3 , the path $U = \langle A, B, C, D, E, F \rangle$ is an inducing path over $\mathbf{O} = \{A, B, D, F\}$ because each collider on U (B and D) is an ancestor of one of the endpoints, and each variable on U that is in \mathbf{O} (except for the endpoints of U) is a collider on U . Similarly, U is an inducing path over $\mathbf{O} = \{A, B, F\}$. However, U is not an inducing path over $\mathbf{O} = \{A, B, C, D, F\}$ because C is in \mathbf{O} , but C is not a collider on U .

Theorem 6.1: If G is a directed acyclic graph with vertex set V , and \mathbf{O} is a subset of V containing A and B , then A and B are not d-separated by any subset Z of $\mathbf{O} \setminus \{A, B\}$ if and only if there is an inducing path over the subset \mathbf{O} between A and B .

It follows from Theorem 6.1 and the fact that U is an inducing path over $\mathbf{O} = \{A, B, D, F\}$ that A and F are d-connected given every subset of $\{B, D\}$. Because in graph G_3 there is no inducing path between A and F over $\mathbf{O} = \{A, B, C, D, F\}$ it follows that A and F are d-separated given some subset of $\{B, C, D\}$ (in this case, $\{B, C, D\}$ itself.)

6.5 Inducing Path Graphs

The inducing paths relative to \mathbf{O} in a graph G over V can be represented in the following structure described (but not named) in Verma and Pearl (1990b). G' is an **inducing path graph over \mathbf{O} for directed acyclic graph G** if and only if \mathbf{O} is a subset of the vertices in G , there is an edge between variables A and B with an arrowhead at A if and only if A and B are in \mathbf{O} , and there is an inducing path in G between A and B relative to \mathbf{O} that is into A . (Using the notation of Chapter 2, the set of marks in an inducing path graph is $\{>, \text{EM}\}$.) In an inducing path graph, there are two kinds of edges: $A \rightarrow B$ entails that every inducing path over

\mathbf{O} between A and B is out of A and into B , and $A <-> B$ entails that there is an inducing path over \mathbf{O} that is into A and into B . This latter kind of edge can only occur when there is a latent common cause of A and B .

Figures 7 through 9 depict the inducing path graphs of G_3 over $\mathbf{O} = \{A,B,D,E,F\}$, $\mathbf{O} = \{A,B,D,F\}$ and $\mathbf{O} = \{A,B,F\}$ respectively. Note that in G_3 $\langle B,D \rangle$ is an inducing path between B and D over $\mathbf{O} = \{A,B,D,E,F\}$ that is out of D . However, in the inducing path graph the edge between B and D has an arrowhead at D because there is another inducing path $\langle B,C,D \rangle$ over $\mathbf{O} = \{A,B,D,E,F\}$ that is into D . There is no edge between A and F in the inducing path graph over $\mathbf{O} = \{A,B,D,E,F\}$, but there is an edge between A and F in the inducing path graphs over $\mathbf{O} = \{A,B,D,F\}$ and $\mathbf{O} = \{A,B,F\}$.

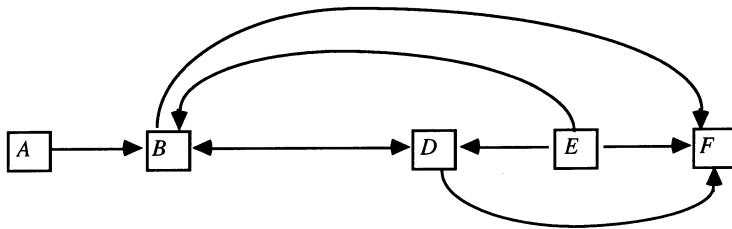


Figure 7: Inducing Path Graph of G_3 Over $\{A,B,D,E,F\}$

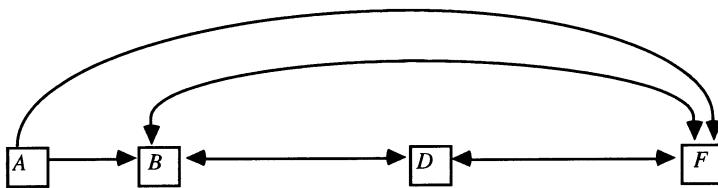


Figure 8: Inducing Path Graph of G_3 Over $\{A,B,D,F\}$

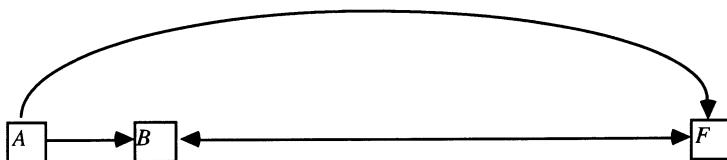


Figure 9: Inducing Path Graph of G_3 Over $\{A,B,F\}$

We can extend without modification the concept of d-separability to inducing path graphs if the only kinds of edges that can occur on a directed path are edges with one arrowhead, and undirected paths may contain edges with either single or double arrowheads. If G is a directed acyclic graph, G' is the inducing path graph for G over O , and X , Y , and S are disjoint sets of variables included in O , then X and Y are **d-separated** given S in G' if and only they are d-separated given S in G .

Double-headed arrows make for a very important difference between d-separability relations in an inducing path graph and in a directed acyclic graph. In a directed acyclic graph over O , if A and B are d-separated given any subset of $O \setminus \{A, B\}$ then A and B are d-separated given either **Parents**(A) or **Parents**(B). This is not true in inducing path graphs. For example, in inducing path graph G_4 , which is the inducing path graph of figure 3 over $O = \{A, B, C, D, E, F, H\}$, **Parents**(A) = $\{D\}$ and **Parents**(E) = $\{B\}$, but A and E are not d-separated given any subset of $\{B\}$ or any subset of $\{D\}$; all of the sets that d-separate A and E contain C , H , or F .

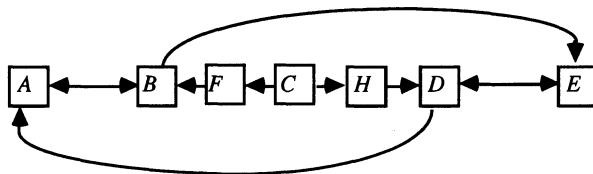


Figure 10: Inducing Path Graph G_4

There is, however, a kind of set of vertices in inducing path graphs that, so far as d-separability is concerned, behaves much like the parent sets in directed acyclic graphs.

If G' is an inducing path graph over O and $A \neq B$, let $V \in \mathbf{D-SEP}(A, B)$ if and only if $A \neq V$ and there is an undirected path U between A and V such that every vertex on U is an ancestor of A or B and (except for the endpoints) is a collider on U .

Theorem 6.2: In an inducing path graph G' over O , where A and B are in O , if A is not an ancestor of B , and A and B are not adjacent then A and B are d-separated given $\mathbf{D-SEP}(A, B)$.

In an inducing path graph either A is not an ancestor of B or B is not an ancestor of A . Thus we can determine whether A and B are adjacent in an inducing path graph without determining whether A and B are dependent conditional on *all* subsets of O .

If \mathbf{O} is not a causally sufficient set of variables, then although we can infer the existence of an inducing path between A and B if A and B are dependent conditional on every subset of $\mathbf{O} \setminus \{A, B\}$, we cannot infer that either A is a direct cause of B relative to \mathbf{O} , or that B is a direct cause of A relative to \mathbf{O} , or that there is a latent common cause of A and B . Nevertheless, the existence of an inducing path between A and B over \mathbf{O} does contain information about the causal relationships between A and B , as the following lemma shows.

Lemma 6.1.4: If G is a directed acyclic graph over \mathbf{V} , \mathbf{O} is a subset of \mathbf{V} that contains A and B , and G contains an inducing path over \mathbf{O} between A and B that is out of A , and A and B are in \mathbf{O} , then there is a directed path from A to B in G .

It follows from Lemma 6.1.4 that if \mathbf{O} is a subset of \mathbf{V} and we can determine that there is an inducing path between A and B over \mathbf{O} that is out of A , then we can infer that A is a (possibly indirect) cause of B . Hence, if we can infer properties of the inducing path graph over \mathbf{O} from the distribution over \mathbf{O} , we can draw inferences about the causal relationships among variables, regardless of what variables we have failed to measure. In the next section we describe algorithms for inferring properties of the inducing path graph over \mathbf{O} from the distribution over \mathbf{O} .

6.6 Partially Oriented Inducing Path Graphs

A **partially oriented inducing path graph** can contain several sorts of edges: $A \rightarrow B$, $A \circ \rightarrow B$, $A \circ \circ B$, or $A \leftarrow \rightarrow B$. We use "*" as a metasymbol to represent any of the three kinds of ends (EM (the empty mark), ">", or "o"); the "*" symbol itself does not appear in a partially oriented inducing path graph. (We also use "*" as a metasymbol to represent the two kinds of ends (EM or ">") that can occur in an inducing path graph.)

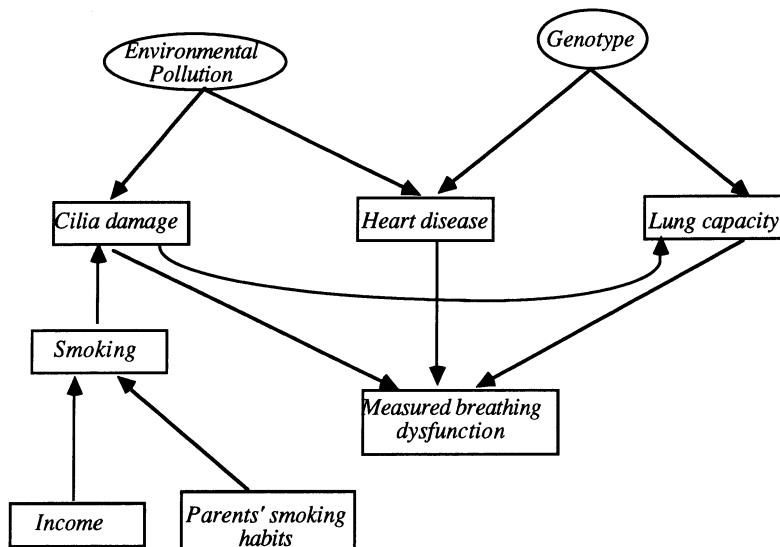
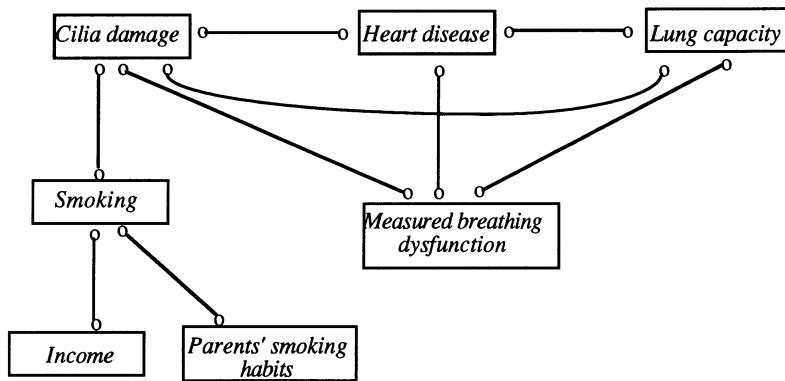
A partially oriented inducing path graph π for directed acyclic graph G with inducing path graph G' over \mathbf{O} is intended to represent the adjacencies in G' , and some of the orientations of the edges in G' that are common to all inducing path graphs with the same d-connection relations as G' . If G' is an inducing path graph over \mathbf{O} , $\text{Equiv}(G')$ is the set of inducing path graphs over the same vertices with the same d-connections as G . Every inducing path graph in $\text{Equiv}(G')$ shares the same set of adjacencies. We use the following definition:

π is a partially oriented inducing path graph of directed acyclic graph G with inducing path graph G' over O if and only if

- (i). if there is any edge between A and B in π , it is one of the following kinds:
 $A \rightarrow B, B \rightarrow A, A \text{ o-} > B, B \text{ o-} > A, A \text{ o-o } B$, or $A <-> B$;
- (ii). π and G' have the same vertices;
- (iii). π and G' have the same adjacencies;
- (iv). if $A *-> B$ is in π , then $A \rightarrow B$ or $A <-> B$ is in every inducing path graph in $\text{Equiv}(G')$;
- (v). if $A \rightarrow B$ is in π , then $A \rightarrow B$ is in every inducing path graph in $\text{Equiv}(G')$;
- (vi). if $A *-* B *-* C$ is in π , then the edges between A and B , and B and C do not collide at B in any inducing path graph in $\text{Equiv}(G')$.

(Strictly speaking a partially oriented inducing path graph is not a graph as we have defined it because of the extra structure added by the underlining.) Note that an edge $A *-\text{o } B$ does not constrain the edge between A and B either to be into or to be out of B in any subset of $\text{Equiv}(G')$. The adjacencies in a partially oriented inducing path graph π for G can be constructed by making A and B adjacent in π if and only if A and B are d-connected given every subset of $O \setminus \{A, B\}$.

Once the adjacencies have been determined, it is trivial to construct an uninformative partially oriented inducing path graph π for G . Simply orient each edge $A *-* B$ as $A \text{ o-o } B$. Of course this particular partially oriented inducing path graph π for G is very uninformative about what features of the orientation of G' are common to all inducing path graphs in $\text{Equiv}(G')$. For example, figure 11 shows again the imaginary graph of causes of measured breathing dysfunction. Figure 12 shows an uninformative partially oriented inducing path graphs of graph G_5 over $O = \{Cilia\ damage, Smoking, Heart\ disease, Lung\ capacity, Measured\ breathing\ dysfunction, Income, Parents'\ smoking\ habits\}$.

Figure 11: Graph G_5 Figure 12: Uninformative Partially Oriented Inducing Path Graph of G_5 Over O

Let us say that B is a **definite non-collider** on undirected path U if and only if either B is an endpoint of U , or there exist vertices A and C such that U contains one of the subpaths $A \prec B \ast\ast C$, $A \ast\ast B \succ C$, or $A \ast\ast \underline{B} \ast\ast C$. In a **maximally informative partially oriented inducing path graph π** for G with inducing path graph G' ,

- (i) an edge $A \ast\text{-o} B$ appears only if the edge between A and B is into B in some members of $\text{Equiv}(G')$, and out of B in other members of $\text{Equiv}(G')$, and
- (ii) for every pair of edges between A and B , and B and C , either the edges collide at B , or they are definite non-colliders at B , unless the edges collide in some members of $\text{Equiv}(G)$ and not in others.

Such a maximally informative partially oriented inducing path graph π for G could be oriented by the simple but inefficient algorithm of constructing every possible inducing path graph with the same adjacencies as G' , throwing out the ones that do not have the same d-connection relations as G' , and keeping track of which orientation features are common to all members of $\text{Equiv}(G')$. Of course, this is completely computationally infeasible. Figure 13 shows the maximally oriented partially oriented inducing path graph of graph G_5 over $O = \{\text{Cilia damage}, \text{Smoking}, \text{Heart disease}, \text{Lung capacity}, \text{Measured breathing dysfunction}, \text{Income}, \text{Parents' smoking habits}\}$.

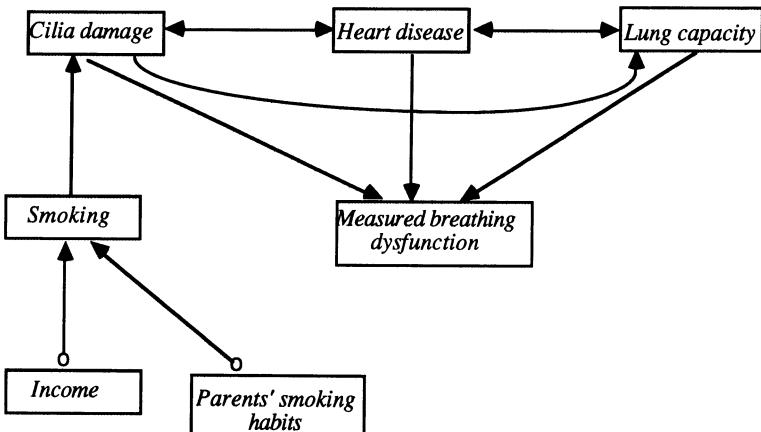


Figure 13: Maximally Informative Partially Oriented Inducing Path Graph of G_5 Over O

Our goal is to state algorithms that construct a partially oriented inducing path graph for a directed acyclic graph G containing as much orientation information as is consistent with computational feasibility. The algorithm we propose is divided into two main parts. First, the

adjacencies in the partially oriented inducing path graph are determined. Then the edges are oriented in so far as possible.

6.7 Algorithms for Causal Inference with Latent Common Causes

In order to state the algorithm, a few more definition are needed. In a partially oriented inducing path graph π :

- (i). A is a **parent** of B if and only if $A \rightarrow B$ in π .
- (ii). B is a **collider** along path $\langle A, B, C \rangle$ if and only if $A * \rightarrow B <-* C$ in π .
- (iii). An edge between B and A is **into** A if and only if $A <-* B$ in π .
- (iv). An edge between B and A is **out of** A if and only if $A \rightarrow B$ in π .
- (v). In a partially oriented inducing path graph π' , U is a **definite discriminating path** for B if and only if U is an undirected path between X and Y containing B , $B \neq X, B \neq Y$, every vertex on U except for B and the endpoints is a collider or a definite non-collider on U , and
 - (i) if V and V' are adjacent on U , and V' is between V and B on U , then $V * \rightarrow V'$ on U ,
 - (ii) if V is between X and B on U and V is a collider on U then $V \rightarrow Y$ in π , else $V <-* Y$ in π ,
 - (iii) if V is between Y and B on U and V is a collider on U then $V \rightarrow X$ in π , else $V <-* X$ in π .
 - (iv) X and Y are not adjacent in π .

Figure 14 illustrates the concept of a definite discriminating path.

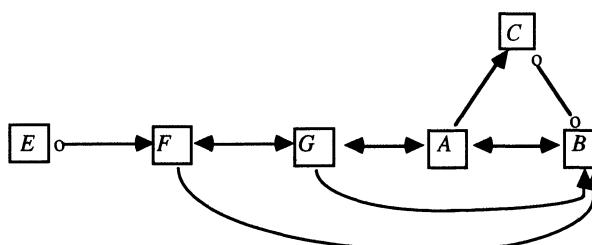


Figure 14: $\langle E, F, G, A, C, B \rangle$ is a definite discriminating path for C

In practice, the Causal Inference Algorithm and the Fast Causal Inference Algorithm (described later in this section) take as input either a covariance matrix or cell counts. Where d-separation facts are needed by the algorithms, the procedure performs tests of conditional independence (in the discrete case) or of vanishing partial correlations (in the linear, continuous case. (Recall that if P is a discrete distribution faithful to a graph G , then A and B are d-separated given a set of variables C if and only A and B are conditionally independent given C , and if P is a distribution linearly faithful to a graph G , then A and B are d-separated given C if and only if $\rho_{AB|C} = 0$.) Both algorithms construct a partially oriented inducing path graph of some directed acyclic graph G , where G contains both measured and unmeasured variables.

Causal Inference Algorithm¹

- A). Form the complete undirected graph Q on the vertex set V .
- B). If A and B are d-separated given any subset S of V , remove the edge between A and B , and record S in $\text{Sepset}(A,B)$ and $\text{Sepset}(B,A)$.
- C). Let F be the graph resulting from step B). Orient each edge as o-o. For each triple of vertices A, B, C such that the pair A, B and the pair B, C are each adjacent in F but the pair A, C are not adjacent in F , orient $A \dashv\vdash B \dashv\vdash C$ as $A \dashv\rightarrow B \dashv\leftarrow C$ if and only if B is not in $\text{Sepset}(A,C)$, and orient $A \dashv\vdash B \dashv\vdash C$ as $A \dashv\rightarrow \underline{B} \dashv\leftarrow C$ if and only if B is in $\text{Sepset}(A,C)$.

D). repeat

If there is a directed path from A to B , and an edge $A \dashv\vdash B$, orient $A \dashv\vdash B$ as $A \dashv\rightarrow B$,

else if B is a collider along $\langle A, B, C \rangle$ in π , B is adjacent to D , and A and C are not d-connected given D , then orient $B \dashv\vdash D$ as $B \dashv\leftarrow D$,

else if U is a definite discriminating path between A and B for M in π , and P and R are adjacent to M on U , and $P - M - R$ is a triangle, then

if M is in $\text{Sepset}(A,B)$ then M is marked as a non-collider on subpath $P \dashv_\underline{M} \dashv\vdash R$

else $P \dashv_\underline{M} \dashv\vdash R$ is oriented as $P \dashv\rightarrow M \dashv\leftarrow R$.

else if $P \dashv\rightarrow \underline{M} \dashv\vdash R$ then orient as $P \dashv\rightarrow M \dashv\rightarrow R$.

until no more edges can be oriented.

If the CI or FCI algorithms use as input a covariance matrix from the marginal over O of a distribution linearly faithful to G , or cell counts from the marginal over O of a distribution faithful to G , we will say the input is data over O that is faithful to G .

Theorem 6.3: If the input to the CI algorithm is data over O that is faithful to G , the output is a partially oriented inducing path graph of G over O .

If data over $O = \{Cilia\ damage, Smoking, Heart\ disease, Lung\ capacity, Measured\ breathing\ dysfunction, Income, Parents'\ smoking\ habits\}$ that is faithful to the graph in figure 11 is input

¹We thank Thomas Verma (personal communication) for pointing out an error in the original formulation of the CI algorithm.

to the CI algorithm, the output is the maximally informative partially oriented inducing path graph over \mathbf{O} shown in figure 13.

Unfortunately, the Causal Inference (CI) algorithm as stated is not practical for large numbers of variables because of the way the adjacencies are constructed. While it is theoretically correct to remove an edge between A and B from the complete graph if and only if A and B are d-separated given some subset of $\mathbf{O} \setminus \{A, B\}$, this is impractical for two reasons. First, there are too many subsets of $\mathbf{O} \setminus \{A, B\}$ on which to test the conditional independence of A and B . Second, for discrete distributions, unless the sample sizes are enormous there are no reliable tests of independence of two variables conditional on a large set of other variables.

In order to determine that a given pair of vertices, such as X and Y are not adjacent in the inducing path graph, we have to find that X and Y are d-separated given some subset of $\mathbf{O} \setminus \{X, Y\}$. Of course, if X and Y are adjacent in the inducing path graph, they are d-connected given every subset of $\mathbf{O} \setminus \{X, Y\}$. We would like to be able to determine that X and Y are d-connected given every subset of $\mathbf{O} \setminus \{X, Y\}$ without actually examining every subset of $\mathbf{O} \setminus \{X, Y\}$.

In a directed acyclic graph over a causally sufficient set V , by using the PC algorithm we are able to reduce the order and number of d-separation tests performed because of the following fact: if X and Y are d-separated by any subset of $V \setminus \{X, Y\}$, then they are d-separated either by **Parents**(X) or **Parents**(Y). While the PC algorithm is constructing the graph it does not know which variables are in **Parents**(X) or in **Parents**(Y), but as the algorithm progresses it is able to determine that some variables are definitely not in **Parents**(X) or **Parents**(Y) because they are definitely not adjacent to X or Y . This reduces the number and the order of the d-separation tests that the PC algorithm performs (as compared to the SGS algorithm).

In contrast, an inducing path graph over \mathbf{O} it is not the case that if X and Y are d-separated given some subset of $\mathbf{O} \setminus \{X, Y\}$, then X and Y are d-separated given either **Parents**(X) or given **Parents**(Y). However, if X and Y are d-separated given *some* subset of $\mathbf{O} \setminus \{X, Y\}$, then X and Y are d-separated given either **D-Sep**(X) or given **D-Sep**(Y). If we know that some variable V is not in **D-Sep**(X) and not in **D-Sep**(Y), we do not need to test whether X and Y are d-separated by any set containing V . Once again, we do not know which variables are in **D-Sep**(X) or **D-Sep**(Y) until we have constructed the graph. But there is an algorithm that can determine that some variables are *not* in **D-Sep**(X) or **D-Sep**(Y) as the algorithm progresses.

Let G be the directed acyclic graph of figure 3 (reproduced below in figure 15.) Let G' be the inducing path graph of G over $\mathbf{O} = \{A, B, C, D, E, F, H\}$. A and E are not d-separated given any

subset of the variables adjacent to A or adjacent to D (in both cases $\{B,D\}$). Because A and E are not adjacent in the inducing path graph of A and E , they are d-separated given some subset of $O\{A,E\}$. Hence they are d-separated by either $\mathbf{D}\text{-Sep}(A,E)$ (equal to $\{B,D,F\}$) or by $\mathbf{D}\text{-Sep}(E,A)$ (equal to $\{B,D,H\}$). (In this case A and E are d-separated by both $\mathbf{D}\text{-Sep}(A,E)$ and by $\mathbf{D}\text{-Sep}(E,A)$.) The problem is: how can we know to test whether A and E are d-separated given $\{B,D,H\}$ or $\{B,D,F\}$ without testing whether A and E are d-separated given every subset of $O\{A,E\}$?

A variable V is in $\mathbf{D}\text{-Sep}(A,E)$ in G' if and only if $V \neq A$ and there is an undirected path between A and V on which every vertex except the endpoints is a collider, and each vertex is an ancestor of A or E . If we could find some method of determining that a variable V does not lie on such a path, then we would not have to test whether A and E were d-separated given any set containing V (unless of course V was in $\mathbf{D}\text{-Sep}(E,A)$.) We will illustrate the strategy on G . At any given stage of the algorithm we will call the graph constructed thus far π .

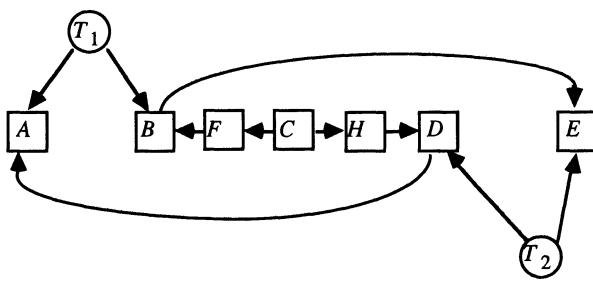
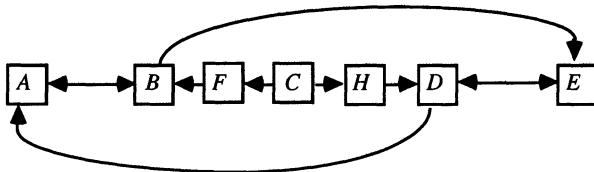
Graph G Inducing Path Graph G'

Figure 15

The FCI algorithm determines which edges to remove from the complete graph in three stages. The first stage is just like the first stage of the PC Algorithm. We initialize π to the complete undirected graph, and then we remove an edge between X and Y if they are d-separated given subsets of vertices adjacent to X or Y in π . This will eliminate many, but perhaps not all of the edges that are not in the inducing path graph. When this operation is performed on data faithful to the graph in figure 15, the result is the graph in figure 16.

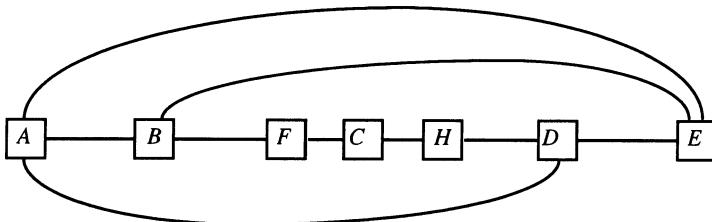


Figure 16

Note that A and E are still adjacent at this stage of the procedure because the algorithm, having correctly determined that A is not adjacent to F or H or C , and that E is not adjacent to F or H or C , never tested whether A and E are d-separated by any subset of variables containing F , H , or C .

Second, we orient edges by determining whether they collide or not, just as in the PC algorithm. The graph at this stage of the algorithm is show in figure 17.

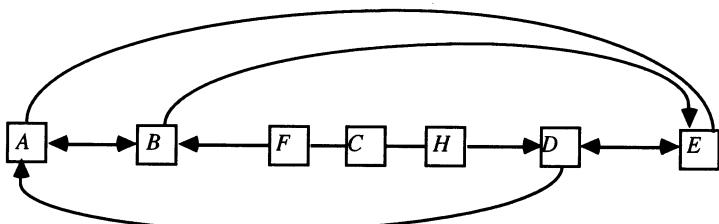


Figure 17

Figure 17 is essentially the graph constructed by the PC algorithm given data faithful to the graph in figure 15, after steps A), B), and C) have been performed.

We can now determine that some vertices are definitely not in $\mathbf{D}\text{-Sep}(A,E)$ or in $\mathbf{D}\text{-Sep}(E,A)$; it is not necessary to every test whether A and E are d-separated given any subset of $O\{A,E\}$ that contains these vertices in order to find the correct adjacencies. At this stage of the algorithm, a necessary condition for a vertex V to be in $\mathbf{D}\text{-Sep}(A,E)$ in G' is that in π there is an undirected path U between A and V in which each vertex except for the endpoints is either a collider, or has its orientation hidden because it is in a triangle. Thus C and H are definitely not in $\mathbf{D}\text{-Sep}(A,E)$ and C and F are definitely not in $\mathbf{D}\text{-Sep}(E,A)$. All of the vertices that we have not definitely determined are not in $\mathbf{D}\text{-Sep}(A,E)$ in G' we place in $\mathbf{Possible-D-Sep}(A,E)$, and similarly for $\mathbf{Possible-D-Sep}(E,A)$. In this case, $\mathbf{Possible-D-Sep}(A,E)$ is $\{B,F,D\}$ and $\mathbf{Possible-D-Sep}(E,A)$ is $\{B,D,H\}$. We now know that if A and E are d-separated given any subset of $O\{A,E\}$ then they are d-separated given some subset of $\mathbf{Possible-D-Sep}(A,E)$ or some subset of $\mathbf{Possible-D-Sep}(E,A)$. In this case we find that A and E are d-separated given a subset of $\mathbf{Possible-D-Sep}(A,E)$ (in this case the entire set) and hence remove the edge between A and E .

Once we have obtained the correct set of adjacencies, we unorient all of the edges, and then proceed to re-orient them exactly as we did in the Causal Inference Algorithm. The resulting output is shown in figure 18.

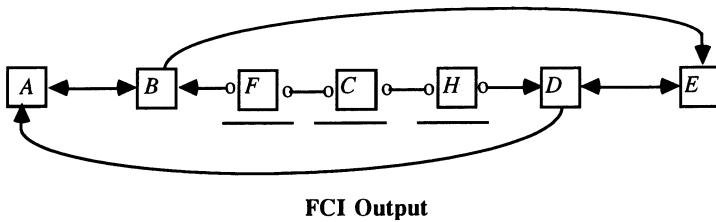


Figure 18

For a given partially constructed partially oriented inducing path graph π , $\mathbf{Possible-D-Sep}(A,B)$ is defined as follows: If $A \neq B$, V is in $\mathbf{Possible-D-Sep}(A,B)$ in π if and only if $V \neq A$, and there is an undirected path U between A and V in π such that for every subpath $\langle X, Y, Z \rangle$ of U either Y is a collider on the subpath, or Y is not a definite non-collider and on U , and X , Y , and Z form a triangle in π .

Using this definition of $\mathbf{Possible-D-Sep}(A,E)$, we can prove that every vertex not in $\mathbf{Possible-D-Sep}(A,E)$ in π is not in $\mathbf{D}\text{-Sep}(A,E)$ in G' . However, it may be possible to determine from π that some members that we are including in $\mathbf{Possible-D-Sep}(A,E)$ are not

in $\mathbf{D-Sep}(A,E)$ in G' . There is clearly a trade-off between reducing the size of **Possible-D-Sep**(A,E) (so that the number and order of tests of d-separability performed by the algorithm is reduced) and performing the extra work required to reduce the size of the set, while ensuring that it is still a superset of $\mathbf{D-Sep}(A,E)$ in G' . We do not know what the optimal balance is. If G is sparse (i.e. each vertex is not adjacent to a large number of other vertices in G), then the algorithm does not need to determine whether A and B are d-separated given C for any C containing a large number of variables.

Fast Causal Inference Algorithm

A). Form the complete undirected graph Q on the vertex set V .

B). $n = 0$.

repeat

repeat

select an ordered pair of variables X and Y that are adjacent in Q such that $\mathbf{Adjacencies}(Q,X)\setminus\{Y\}$ has cardinality greater than or equal to n , and a subset S of $\mathbf{Adjacencies}(Q,X)\setminus\{Y\}$ of cardinality n , and if X and Y are d-separated given S delete the edge between X and Y from Q , and record S in $\mathbf{Sepset}(X,Y)$ and $\mathbf{Sepset}(Y,X)$

until all ordered variable pairs of adjacent variables X and Y such that $\mathbf{Adjacencies}(Q,X)\setminus\{Y\}$ has cardinality greater than or equal to n and all subsets S of $\mathbf{Adjacencies}(Q,X)\setminus\{Y\}$ of cardinality n have been tested for d-separation;

$n = n + 1$;

until for each ordered pair of adjacent vertices X, Y , $\mathbf{Adjacencies}(Q,X)\setminus\{Y\}$ is of cardinality less than n .

C). Let F' be the undirected graph resulting from step B). Orient each edge as o-o. For each triple of vertices A, B, C such that the pair A, B and the pair B, C are each adjacent in F' but the pair A, C are not adjacent in F' , orient $A \dashv \dashv B \dashv \dashv C$ as $A \dashv \dashv B \dashv \dashv C$ if and only if B is not in $\mathbf{Sepset}(A,C)$.

D). For each pair of variables A and B adjacent in F' , if A and B are d-separated given any subset S of $\mathbf{Possible-D-SEP}(A,B)\setminus\{A,B\}$ or any subset S of $\mathbf{Possible-D-SEP}(B,A)\setminus\{A,B\}$ in F remove the edge between A and B , and record S in $\mathbf{Sepset}(A,B)$ and $\mathbf{Sepset}(B,A)$.

The algorithm then re-orients an edge between any pair of variables X and Y as X o-o Y , and proceeds to re-orient the edges in the same way as steps C) and D) of the Causal Inference algorithm.

Theorem 6.4: If the input to the FCI algorithm is data over \mathbf{O} that is faithful to G , the output is a partially oriented inducing path graph of G over \mathbf{O} .

The Fast Causal Inference Algorithm (FCI) always produces a partially oriented inducing path graph for a graph G given correct statistical decisions from the marginal over the measured variables of a distribution faithful to G . We do not know whether the algorithm is complete, i.e. whether it in every case produces a maximally informative partially oriented inducing path graph.

As with the CI algorithm, if the input to the FCI algorithm is data faithful to the graph of figure 11, the output is the maximally informative partially oriented inducing path graph of figure 13.

Two directed acyclic graphs G and G' that have the same FCI partially oriented inducing path graph over \mathbf{O} have the same d-connection relations involving just members of \mathbf{O} .

Corollary 6.4.1: If G is a directed acyclic graph over \mathbf{V} , G' is a directed acyclic graph over \mathbf{V}' , and \mathbf{O} is a subset of \mathbf{V} and of \mathbf{V}' , then G and G' have the same d-separation relations among only the variables in \mathbf{O} if and only if they have the same FCI partially oriented inducing path graph over \mathbf{O} .

Given a directed acyclic graph G , it is possible to determine what d-separation relations involving just members of \mathbf{O} are true of G from the FCI partially oriented inducing path graph of G over \mathbf{O} . In a partially oriented inducing path graph π , if $X \neq Y$, and X and Y are not in \mathbf{Z} , then an undirected path U between X and Y **definitely d-connects** X and Y given \mathbf{Z} if and only if every collider on U has a descendant in \mathbf{Z} , every definite non-collider on U is not in \mathbf{Z} , and every other vertex on U is not in \mathbf{Z} but has a descendant in \mathbf{Z} . In a partially oriented inducing path graph π , if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint sets of variables, then \mathbf{X} is definitely d-connected to \mathbf{Y} given \mathbf{Z} if and only if some member of \mathbf{X} is d-connected to some member of \mathbf{Y} given \mathbf{Z} .

Corollary 6.4.2: If G is a directed acyclic graph over \mathbf{V} , \mathbf{O} is a subset of \mathbf{V} , π is the FCI partially oriented inducing path graph of G over \mathbf{O} , and \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint subsets of \mathbf{O} , then \mathbf{X} is d-connected to \mathbf{Y} given \mathbf{Z} in G if and only if \mathbf{X} is definitely d-connected to \mathbf{Y} given \mathbf{Z} in π .

These corollaries are proved in Spirtes and Verma (1992).

6.8 Theorems on Detectable Causal Influence

In this section we show that a number of different kinds of causal inferences can be drawn from a partially oriented inducing path graph.

Theorem 6.5: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is a directed path U from A to B in π , then there is a directed path from A to B in G .

If G is a directed acyclic graph over \mathbf{V} , and \mathbf{O} is included in \mathbf{V} , if the input to the CI algorithm is data faithful to G over \mathbf{O} , then we call the output of the CI algorithm the **CI partially oriented inducing path graph** of G over \mathbf{O} . We adopt a similar terminology for the FCI algorithm. A **semi-directed path from A to B** in partially oriented inducing path graph π is an undirected path U from A to B in which no edge contains an arrowhead pointing towards A , that is, there is no arrowhead at A on U , and if X and Y are adjacent on the path, and X is between A and Y on the path, then there is no arrowhead at the X end of the edge between X and Y .

Theorem 6.6: If π is the CI partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is no semi-directed path from A to B in π , then there is no directed path from A to B in G .

Recall that a **trek** between distinct variables A and B is either a directed path from A to B , a directed path from B to A , or a pair of directed paths from a vertex C to A and B respectively that intersect only at C . The following theorem states a sufficient condition for when the edges in a partially oriented inducing path graph indicate a trek in the graph that contains no measured vertices except for the endpoints.

Theorem 6.7: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , A and B are adjacent in π , and there is no undirected path between A and B in π except for the edge between A and B , then in G there is a trek between A and B that contains no variables in \mathbf{O} other than A or B .

Theorem 6.8: If π is the CI partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and every semi-directed path from A to B contains some member of \mathbf{C} in π , then every directed path from A to B in G contains some member of \mathbf{C} .

Theorem 6.9: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and $A <-> B$ in π , then there is a latent common cause of A and B in G .

Parallel results holds for the FCI algorithm.

To illustrate the application of these theorems, consider the maximally informative partially oriented inducing path graph in figure 13 of the causal structure of G_5 . Applying Theorem 6.5 we infer that *Smoking* causes *Cilia damage*, *Lung capacity* and *Measured breathing dysfunction*. Applying Theorem 6.6, we infer that *Smoking* does not cause *Heart disease* or *Income* or *Parents' Smoking Habits*. It is impossible to determine from the conditional independence relations among the measured variables whether *Income* causes *Smoking*, or there is a common cause of *Smoking* and *Income*. The statistics among the measured variables determine that *Cilia damage* and *Heart disease* have a latent common cause, *Cilia damage* does not cause *Heart disease*, and *Heart disease* does not cause *Cilia damage*.

We note here a topic that will be more fully explored in the next chapter. In the example from figure 11, in order to infer that smoking causes breathing dysfunction, it is necessary to measure two causes of *Smoking* (whose collision at *Smoking* orients the edge from *Smoking* to *Cilia damage*.) In general, this suggests that in the design of studies intended to determine if there is a causal path from variable A to variable B , it is useful to measure not only variables that might mediate the connection between A and B , but also to measure possible causes of A .

6.9 Non-Independence Constraints

The Markov and Faithfulness conditions applied to a causally insufficient graph may entail constraints on the marginal distribution of measured variables that are not conditional independence relations, and hence are not used in the FCI algorithm. Consider, the example in figure 19, due to Thomas Verma (Verma and Pearl, 1991).

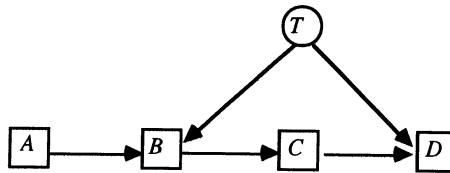


Figure 19

Assume T is unmeasured. Then a joint distribution faithful to the entire graph must satisfy the constraint that the quantity

$$\sum_B^{\rightarrow} P(B|A)P(D|B,C,A)$$

is a function only of the values of C and D .

$$\begin{aligned} \sum_B^{\rightarrow} P(B|A)P(D|B,C,A) &= \sum_T^{\rightarrow} \sum_B^{\rightarrow} P(B|A)P(D|B,C,A,T)P(T|B,C,A) = \\ &= \sum_T^{\rightarrow} P(D|C,T) \sum_B^{\rightarrow} P(B|A)P(T|B,A) \end{aligned}$$

(because A, B are independent of D given $\{C, T\}$ and C is independent of T given $\{A, B\}$). Hence

$$\begin{aligned} \sum_T^{\rightarrow} P(D|C,T) \sum_B^{\rightarrow} P(B|A)P(T|B,A) &= \sum_T^{\rightarrow} P(D|C,T)P(T|A) = \\ &= \sum_T^{\rightarrow} P(D|C,T)P(T) = g(C,D) \end{aligned}$$

(because T and A are independent).

This constraint is not entailed if a directed edge from A to D is added to the graph. The moral is that there is further marginal structure not in the form of conditional independence relations that could in principle be used to help identify latent structure. We will see a similar point when we turn to linear models in the next section.

6.10 Generalized Statistical Indistinguishability and Linearity

Suppose that for whatever reasons an investigation were to be confined to linear structures and to probability distributions that are consistent with the assumption that each random variable is a linear function of its parents and of unmeasured factors. The effect of restrictions such as linearity is to make distinguishable causal structures that would otherwise be indistinguishable. That happens because the restriction, whatever it is, together with the conditional dependence and independence relations required by the Markov, Minimality or Faithfulness Conditions, entails additional constraints on the measured variables. These additional constraints may not be in the form of conditional independence relations. In the linear case they typically are not. Consider for example the two structures shown below, where the X variables are measured and the T variables are unmeasured.

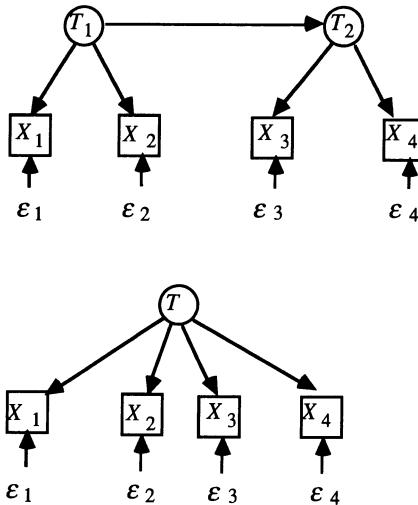


Figure 20

These structures each imply that in the marginal distribution over the measured variables every pair of variables is dependent conditional on every other set of measured variables. In each case the maximally informative partially oriented inducing path graph on the X variables is a complete undirected graph. By examining conditional independence relations among these variables, one could not tell which structure obtains. But if linearity is required, then it is easy to tell which structure obtains. For under the linearity assumption, the second structure entails all three of the following constraints on the correlations of the measured variables, while the first structure entails only the first of these constraints (where we denote the correlation between X_1 and X_2 as ρ_{12} in order to avoid subscripts with subscripts.) :

$$\begin{aligned}\rho_{13}\rho_{24} - \rho_{14}\rho_{23} &= 0 \\ \rho_{12}\rho_{34} - \rho_{14}\rho_{23} &= 0 \\ \rho_{13}\rho_{24} - \rho_{12}\rho_{34} &= 0\end{aligned}$$

Early in this century Charles Spearman (1928) called constraints of these sorts **vanishing tetrads**, and we will use his terminology.

Characterizing statistical indistinguishability under the linearity restriction thus presents an entirely new problem, and one for which we will offer no general solution. It is not true, for

example, that conditional independence relations and vanishing tetrad differences jointly determine the faithful indistinguishability classes of linear structures with unmeasured variables. For example, each of the following linear structures entails that a single tetrad difference vanishes in the marginal distribution over A, B, C and D , and has a partially oriented inducing path graph for these variables consisting of a complete undirected graph:

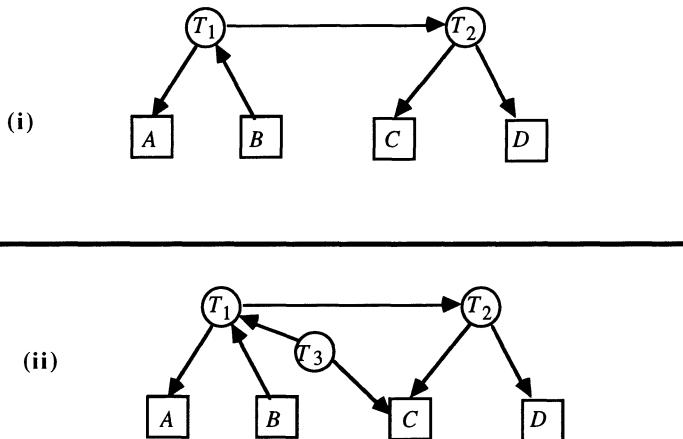


Figure 21

But the two graphs are not faithfully indistinguishable over the class of linear structures. Structure (ii) permits distributions consistent with linearity in which the correlation of A and B is positive, the correlation of B and C is positive and the correlation of A and C is negative. Structure (i) admits no distributions consistent with linearity whose marginals satisfy this condition.

Structures (i) and (ii) are not typical of the linear causal structures with unmeasured variables one finds in the social science literature. For practical purposes, the examination of vanishing tetrad constraints provides a powerful means to distinguish between alternative causal structures, even in structures that are only partially linear. Tests for hypotheses of vanishing tetrad differences were introduced by Wishart in the 1920s assuming normal variates, and asymptotically distribution free tests have been described by Bollen (1989).

Algorithms that take advantage of vanishing tetrad differences will be described and illustrated later in this book. In order to take that advantage, we need to be able to determine

algorithmically when a structure with or without unmeasured common causes entails a particular vanishing tetrad difference among the measured variables. This question leads to an important theorem.

6.11 The Tetrad Representation Theorem

We wish to characterize entirely in graph theoretic terms a necessary and sufficient condition for a distribution on the vertices of an arbitrary directed acyclic graph G to **linearly imply** a vanishing tetrad difference, that is the tetrad difference vanishes in all of the distributions linearly represented by G . We will call a distribution linearly represented by some directed acyclic graph G a **linear model**. (A slightly more formal definition is given in Chapter 13.) A linear model is uniquely determined by the directed acyclic graph G that represents it, and linear coefficients and the independent marginal distributions on the variables (including error terms) of zero indegree.

First some terminology: Given a trek $T(I,J)$ between vertices I and J , $I(T(I,J))$ denotes the directed path in $T(I,J)$ from the source of $T(I,J)$ to I and $J(T(I,J))$ denotes the directed path in $T(I,J)$ from the source of $T(I,J)$ to J . (Recall that one of the directed paths in a trek may be an empty path.) $\mathbf{T}(I,J)$ denotes the set of all treks between I and J .

In a directed acyclic graph G , if for all $T(K,L)$ in $\mathbf{T}(K,L)$ and all $T(I,J)$ in $\mathbf{T}(I,J)$, $L(T(K,L))$ and $J(T(I,J))$ intersect at a vertex Q , then Q is an $LJ(T(I,J),T(K,L))$ **choke point**. Similarly, if for all $T(K,L)$ in $\mathbf{T}(K,L)$ and all $T(I,J)$ in $\mathbf{T}(I,J)$, $L(T(K,L))$ and all $J(T(I,J))$ intersect at a vertex Q , and for all $T(I,L)$ in $\mathbf{T}(I,L)$ and all $T(J,K)$ in $\mathbf{T}(J,K)$, $L(T(I,L))$ and $J(T(J,K))$ also intersect at Q , then Q is an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ **choke point**. Also see the definition of trek.

The fundamental theorem for vanishing tetrad differences in linear models is this:

Tetrad Representation Theorem 6.10: In a directed acyclic graph G , there exists an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point if and only if G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

A consequence of Theorem 6.10 is

Theorem 6.11: An acyclic graph G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ only if either ρ_{IJ} or $\rho_{KL} = 0$, and ρ_{IL} or $\rho_{JK} = 0$, or there is a (possibly empty) set \mathbf{Q} of random variables in G such that $\rho_{IJ}.\mathbf{Q} = \rho_{KL}.\mathbf{Q} = \rho_{IL}.\mathbf{Q} = \rho_{JK}.\mathbf{Q} = 0$.

Theorem 6.10 provides a fast algorithm for calculating the vanishing tetrad differences linearly implied by any directed acyclic graph. Theorem 6.11 provides a means to determine when unmeasured common causes are acting in linear structures. In later chapters we describe some of the implications of these facts for investigating the structure of causal relations among unmeasured variables.

6.12 An Example: Math Marks and Causal Interpretation

In several places in his recent text on graphical models in statistics, Whittaker (1990) discusses a data set from Mardia, Kent and Bibby (1979) concerning the grades of 88 students on examinations in five mathematical subjects: mechanics, vectors, algebra, analysis and statistics. The example illustrates one of the uses of the Tetrad Representation Theorem, and provides occasion to comment on some important differences of interpretation between our methods and those Whittaker describes. The variance/covariance matrix for the data is as follows:

<i>Mechanics</i>	<i>Vectors</i>	<i>Algebra</i>	<i>Analysis</i>	<i>Statistics</i>
302.29				
125.78	170.88			
100.43	84.19	111.60		
105.07	93.60	110.84	217.88	
116.07	97.89	120.49	153.77	294.37

When given these data, the PC algorithm immediately determines the following pattern:

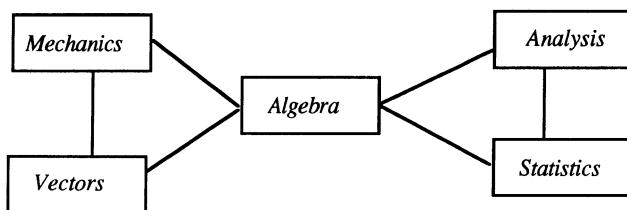


Figure 22

Whittaker obtains the same graph under a different interpretation. Recall that an undirected independence graph is any pair $\langle G, P \rangle$ where G is an undirected graph and P is a distribution such that vertices X, Y in G are not adjacent if they are independent conditional on the set of all other vertices of G ; or to state the contrapositive: if X, Y are dependent conditional on the set of all other vertices of G , then X, Y are adjacent in G . Undirected independence graphs hide much of the causal structure, and sometimes many of the independence relations. Thus if variables X and Z are causes of variable Y but X and Z are statistically independent and have no causal relations whatsoever, the undirected independence graph has an edge between X and Z . In effect, the independence graph fails to represent the conditional independence relations that hold among proper subsets of a set of variables.

Every *undirected* pattern graph obtained from a faithful distribution (or sample) is a subgraph of the undirected independence graph obtained from that distribution. In the case at hand the two graphs are the same, but they need not be in general.

Whittaker claims that identifying the undirected independence graph is important for four reasons: (i) it reduces the complex five dimensional object into two simpler three dimensional objects--the two maximal cliques in the graph; (ii) it groups the variables into two sets; (iii) it highlights *Algebra* as the one crucial examination in analyzing the interrelationship between different subjects in exam performance; (iv) it asserts that *Algebra* and *Analysis* alone will be sufficient to predict *Statistics* and that *Algebra* and *Vectors* will be sufficient to predict *Mechanics*; but that all four marks are needed to predict *Algebra* (p. 6)

The second reason seems simply a consequence of the first, and the first seems of little consequence: the cognitive burden of noting that there are five variables is not very great. There is a long tradition in statistics of introducing representations on grounds that they simplify the data and in practice treating the objects of such reductions as causes. That is, for example, the history of factor analysis after Thurstone. But as with factor analysis, causal conclusions drawn

from independence graphs would be unreliable. The third reason seems too vague to be worth much trouble. The assertion given in the fourth reason is sound, but only if "predict" is understood in all cases to have nothing to do with predicting the values of variables when they are deliberately altered, as by coaching. We suspect statistical analyses of such educational data are apt to be given a causal significance, and for such purposes directed graphical models better represent the hypotheses.

Applying Theorem 6.11, the vanishing tetrad test for latent variables, we find that there are four vanishing tetrad differences that cannot be explained by vanishing partial correlations among the measured variables. This suggests that it is entailed by vanishing partial correlations involving latent variables, and thus suggests the introduction of latent variables. A natural idea in view of the mathematical structure of the subjects tested is that *Algebra* is an indicator of *Algebraic knowledge*, which is a factor in the *Knowledge of vector algebra* measured by *Vector* and *Mechanics* and is also a factor in *Knowledge of real analysis* that affects *Analysis* and *Statistics*. The explanation of the data then looks like this:

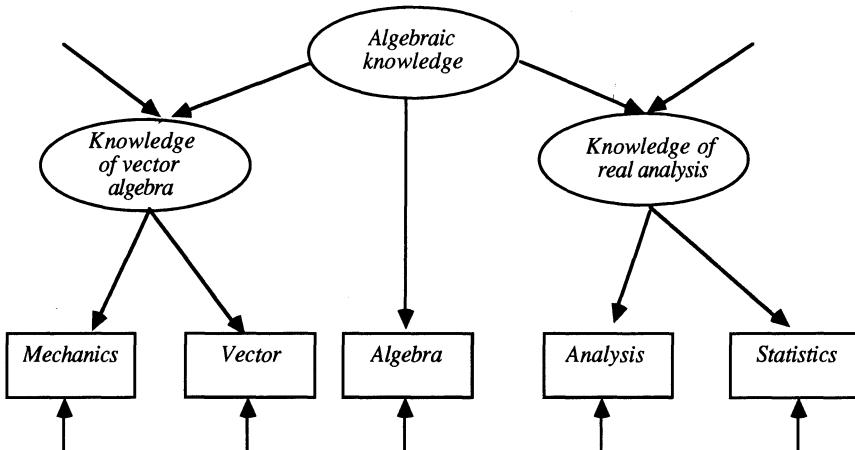


Figure 23

The arrows without notation attached to them indicate other sources of variation. Assuming a faithful distribution and linearity, this graph does not entail the vanishing first order partial correlations among the measured variables that the data suggest. But if the variance in *Algebra* due to factors other than algebraic knowledge is sufficiently small, a linear distribution faithful to this graph will to good approximation give exactly those vanishing partial correlations.

This structure (assuming linearity) entails eight vanishing tetrad differences, all of which the TETRAD II program identifies and tests and cannot reject ($p > .7$). The model itself, when treated as the null hypotheses in a likelihood ratio test, yields a p value of about .9, roughly the value Whittaker reports for the undirected graphical independence model.

6.13 Background Notes

In a series of papers (Pearl and Verma 1990, 1991, Verma and Pearl 1990a, 1990b, 1991) Verma and Pearl describe an "Inductive Causation" algorithm that outputs a structure that they call a pattern (or sometimes a "completed hybrid graph") of a directed acyclic graph G over a set of variables \mathbf{O} . The most complete description of their theory appears in Verma and Pearl (1990b). The key ideas of an inducing path, an inducing path graph, and the proof of (what we call) Theorem 6.1 all appear in this paper. Unfortunately, the two main claims about the output of the Inductive Causation Algorithm made in the paper, given in their lemma A2 and their Theorem 2, are false (see Spirtes, 1992).

Early versions of the Inductive Causation Algorithm did not distinguish between $A \rightarrow B$ and $A \circ \rightarrow B$, and hence could not be used to infer that A causes B as in Theorem 6.5. This distinction was introduced (in a different notation) in order to prove a version of Theorem 6.5 and Theorem 6.6 in Spirtes and Glymour (1990); Verma and Pearl incorporated it in a subsequent version of the Inductive Causation Algorithm. The Inductive Causation Algorithm does not use definite discriminating paths to orient edges, and hence in some cases gives less orientation information than the FCI procedure. The output of the Inductive Causation Algorithm has no notation distinguishing between edges in triangles that definitely do not collide and merely unoriented edges. Like the CI algorithm, the Inductive Causation Algorithm cannot be applied to large numbers of variables because testing the independence of some pairs of variables conditional on every subset of $\mathbf{O} \setminus \{A, B\}$ is required.

The vanishing tetrad difference was used as the principle technique in model specification by Spearman and his followers. A brief account of their methods is given in Glymour, Scheines, Spirtes and Kelly (1987). Spearman's inference to common causes from vanishing tetrad differences was challenged by Godfrey Thomson in a series of papers between 1916 and 1935. In our terms, Thomson's models all violated linear faithfulness.

Chapter 7

Prediction

7.1 Introduction

The fundamental aim of many empirical studies is to predict the effects of changes, whether the changes come about naturally or are imposed by deliberate policy: Will the reduction of sources of environmental lead increase the intelligence of children in exposed regions? Will increased taxation of cigarettes decrease lung cancer? How large will these effects be? What will be the differential yield if a field is planted with one species of wheat rather than another; or the difference in number of polio cases per capita if all children are vaccinated against polio as against if none are; or the difference in recidivism rates if parolees are given \$600 per month for six months as against if they are given nothing; or the reduction of lung cancer deaths in middle aged smokers if they are given help in quitting cigarette smoking; or the decline in gasoline consumption if an additional dollar tax per gallon is imposed?

One point of experimental designs of the sort found in randomized trials is to attempt to *create* samples that, from a statistical point of view, are from the very distributions that would result if the corresponding treatments were made general policy and applied everywhere. For such experiments under such assumptions, the problems of statistical inference are conventional, which is not to say they are easy, and the prediction of policy outcomes is not problematic in principle. But in empirical studies in the social sciences, in epidemiology, in economics, and in many other areas, we do not know or cannot reasonably assume that the observed sample is from the very distribution that would result if a policy were adopted. Implementing a policy may change relevant variables in ways not represented in the observed sample. The inference task is to move from a sample obtained from a distribution corresponding to passive observation or quasi-experimental manipulation, to conclusions about the distribution that would result if a policy were imposed. In our view one of the most fundamental questions of statistical inference is when, if ever, such inferences are possible, and, if ever they are possible,

by what means. The answer, according to Mosteller and Tukey, is "never." We will see whether that answer withstands analysis.

7.2 Prediction Problems

The possibilities of prediction may be analyzed in a number of different sorts of circumstances, including at least the following:

Case 1: We know the causal graph, which variables will be directly manipulated, and what the direct manipulation will do to those variables. We want to predict the distribution of variables that will not be directly manipulated. More formally, we know the set X of variables being directly manipulated, $P(X|Parents(X))$ in the manipulated distribution, and that $Parents(X)$ in the manipulated population is a subset of $Parents(X)$ in the unmanipulated population. That is essentially the circumstance that Rubin, Holland and Pratt and Schlaifer address, and in that case the causal graph and the Manipulation Theorem specify a relevant formula for calculating the manipulated distribution in terms of marginal conditional probabilities from the unmanipulated distribution. The latter can be estimated from samples; we can find the distribution of Y (or of Y conditional on Z) under direct manipulation of X by taking the appropriate marginal of the calculated manipulated distribution.

Case 2: We know the set X of variables being directly manipulated, $P(X|Parents(X))$ in the manipulated distribution, that $Parents(X)$ in the manipulated population is a subset of $Parents(X)$ in the unmanipulated population, and that the measured variables are causally sufficient; unlike case 1, we do not know the causal graph. The causal graph must be conjectured from sample data. In this case the sample and the PC (or some other) algorithm determine a pattern representing a class of directed graphs, and properties of that class determine whether the distribution of Y following a direct manipulation of X can be predicted.

Case 3. The difficult, interesting and realistic case arises when we know the set X of variables being directly manipulated, we know $P(X|Parents(X))$ in the manipulated population, and that $Parents(X)$ in the manipulated population is a subset of $Parents(X)$ in the unmanipulated population, but prior knowledge and the sample leave open the possibility that there *may be* unmeasured common causes of the measured variables. If observational studies were treated without unsupported pre-conceptions, surely that would be the typical circumstance. It is

chiefly because of this case that Mosteller and Tukey concluded that prediction from uncontrolled observations is not possible. One way of viewing the fundamental problem of predicting the distribution of Y or conditional distribution of Y on Z upon a direct manipulation of X can be formulated this way: *find conditions sufficient for prediction, and conditions necessary for prediction, given only a partially oriented inducing path graph and conditional independence facts true in the marginal (over the observed variables) of the unmanipulated distribution. Show how to calculate features of the predicted distribution from the observed distribution.* The ultimate aim of this chapter is to provide a partial solution to this problem.

We will take up these cases in turn. Case 1 is easy but we take time with it because of the connection with Rubin's theory. Case 2 is dealt with very briefly. In our view Case 3 describes the more typical and theoretically most interesting inference problems. The reader is warned that even when the proofs are postponed the issue is intricate and difficult.

7.3 Rubin-Holland-Pratt-Schlaifer Theory¹

Rubin's framework has a simple and appealing intuition. In experimental or observational studies we sample from a population. Each unit in the population, whether a child or a national economy or a sample of a chemical, has a collection of properties. Among the properties of the units in the population, some are *dispositional*--they are propensities of a system to give a response to a treatment. A glass vase, for example, may be fragile, meaning that it has a disposition to break if struck sharply. A dispositional property isn't exhibited unless the appropriate treatment is applied--fragile vases don't break unless they are struck. Similarly, in a population of children, for each reading program each child has a disposition to produce a certain post-test score (or range of test scores) if exposed to that reading program. In experimental studies when we give different treatments to different units, we are attempting to estimate dispositional properties of units (or their averages, or the differences of their averages) from data in which only some of the units have been exposed to the circumstances in which that disposition is manifested. Rubin associates with each such dispositional quantity, Q , and each value x of relevant treatment variable X , a random variable, $Q_{xf=x}$, whose value for each unit in the population is the value Q would have if that unit were to be given treatment x , or in other words if the system were forced to have X value equal to x . If unit i is actually given treatment

¹This section is based on Spirtes, Glymour, Scheines, Meek, Fienberg and Slate, 1992.

x_1 and a value of Q is measured for that unit, the measured value of Q equals the value of $Q_{Xf=x_1}$.

Experimentation may give a set of paired values $\langle x, y_{Xf=x} \rangle$, where $y_{Xf=x}$ is the value of the random variable $Y_{Xf=x}$. But for a unit i that is given treatment x_1 , we also want to know the value of $Y_{Xf=x_2}$, $Y_{Xf=x_3}$, and so on for each possible value of X , representing respectively the values for Y unit i is disposed to exhibit if unit i were exposed to treatment x_2 or x_3 , that is, if the X value for these units were forced to be x_2 or x_3 rather than x_1 . These unobserved values depend on the causal structure of the system. For example, the value of Y that unit i is disposed to exhibit on treatment x_2 might depend on the treatments given to other units. We will suppose that there is no dependence of this kind, but we will investigate in detail other sorts of connections between causal structure and Rubin's counterfactual random variables.

A typical inference problem in Rubin's framework is to estimate the distribution of $Y_{Xf=x}$ for some value x of X , over all units in the population, from a sample in which only some members have received the treatment x . A number of variations arise. Rather than forcing a unique value on X , we may contemplate forcing some specified distribution of values on X , or we may contemplate forcing different specified distributions on X depending on the (unforced) values of some other variables Z ; our "experiment" may be purely observational so that an observed value q of variable Q for unit i when X is observed to have value x is not necessarily the same as $Q_{Xf=x}$. Answers to various problems such as these can be found in the papers cited. For example, in our paraphrasing, Pratt and Schlaifer claim the following:

When all units are systems in which Y is an effect of X and possibly of other variables, and no causes of Y other than X are measured, in order for the conditional distribution of Y on $X = x$ to equal $Y_{Xf=x}$ for all values x of X , it is sufficient and "almost necessary" that X and each of the random variables $Y_{Xf=x}$ (where x ranges over all possible values of X) be statistically independent.

In our terminology, when the conditional distribution of Y on $X = x$ equals $Y_{Xf=x}$ for all values x of X we say that the conditional distribution of Y on X is "invariant"; in their terminology it is "observable." Pratt and Schlaifer's claim may be clarified with several examples, which will also serve to illustrate some tacit assumptions in the application of the framework. Suppose X and U , which is unobserved, are the only causes of Y , and they have no causal connection of any kind with one another, a circumstance that we will represent by the graph in figure 1.

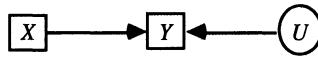


Figure 1

For simplicity let's suppose the dependencies are all linear, and that for all possible values of X , Y and U , and all units, $Y = X + U$. Let X_f represent values of X that could possibly be forced on all units in the population. X is an observed variable; X_f is not. X is a random variable; X_f is not. Consider values in Table 1.

Table 1

X	Y	U	X_f	$U_{X_f=1}$	$Y_{X_f=1}$
1	1	0	1	0	1
1	2	1	1	1	2
1	3	2	1	2	3
2	2	0	1	0	1
2	3	1	1	1	2
2	4	2	1	2	3

Suppose for simplicity that each row (ignoring X_f , which is not a random variable) is equally probable. Here the X and Y columns give possible values of the measured variables. The U column gives possible values of the unmeasured variable U . X_f is a variable whose column indicates values of X that might be forced on a unit; we have not continued the table beyond $X_f = 1$. The $U_{X_f=1}$ column represents the range of values of U when X is forced to have the value 1; the $Y_{X_f=1}$ gives the range of values of Y when X is forced to have the value 1. Notice that in the table $Y_{X_f=1}$ is uniquely determined by the value of X_f and the value of $U_{X_f=1}$ and is independent of the value of X .

The table illustrates Pratt and Schlaifer's claim: $Y_{X_f=1}$ is independent of X and the distribution of Y conditional on $X = 1$ equals the distribution of $Y_{X_f=1}$.

We constructed the table by letting $U = U_{X_f=1}$, and $Y_{X_f=1} = 1 + U_{X_f=1}$. In other words, we obtained the table by assuming that save for the distribution of X , the causal structure and probabilistic structure are completely unaltered if a value of X is forced on all units. By

applying the same procedure with $Y_{Xf=2} = 2 + U_{Xf=2}$, the table can be extended to obtain values when $Xf = 2$ that satisfy Pratt and Schlaifer's claim.

Consider a different example in which, according to Pratt and Schlaifer's rule, the conditional probability of Y on X is *not* invariant under direct manipulation. In this case X causes Y and U causes Y , and there is no causal connection of any kind between X and U , as before, but in addition an unmeasured variable V is a common cause of both X and Y , a situation represented in figure 2.

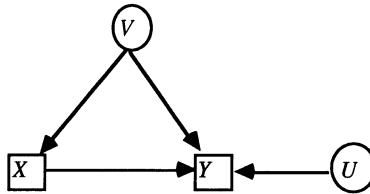


Figure 2

Consider the following distribution, with the same conventions as in Table 1:

Table 2

X	V	U	Y	Xf	$V_{Xf=1}$	$U_{Xf=1}$	$Y_{Xf=1}$
0	0	0	0	1	0	0	1
0	0	1	1	1	0	1	2
0	0	2	2	1	0	2	3
0	0	3	3	1	0	3	4
1	1	0	2	1	1	0	2
1	1	1	3	1	1	1	3
1	1	2	4	1	1	2	4
1	1	3	5	1	1	3	5

Again, assume all rows are equally probable, ignoring the value of Xf which is not a random variable. Notice that $Y_{Xf=1}$ is now *dependent* on the value of X . And, just as Pratt and Schlaifer require, the conditional distribution of Y on $X = 1$ is *not* equal to the distribution of $Y_{Xf=1}$.

The table was constructed so that when $X = 1$ is forced, and hence $Xf = 1$, the distributions of $U_{Xf=1}$, and $V_{Xf=1}$ are independent of Xf . In other words, while the system of equations

$$\begin{aligned} Y &= X + V + U \\ X &= V \end{aligned}$$

was used to obtain the values of X , Y , and U , the assumptions $U_{Xf=1} = U$, $V_{Xf=1} = V$ and the equation

$$Y_{Xf=1} = Xf + V_{Xf=1} + U_{Xf=1}$$

were used to determine the values of $U_{Xf=1}$, $V_{Xf=1}$ and $Y_{Xf=1}$. The forced system was treated as if it were described by the diagram depicted in figure 3.

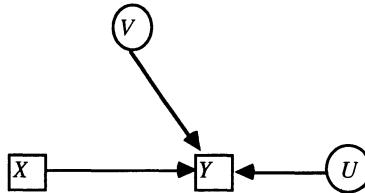


Figure 3

For another example, suppose $Y = X + U$, but there is also a variable V that is dependent on both Y and X , so that the system can be depicted as in figure 4.

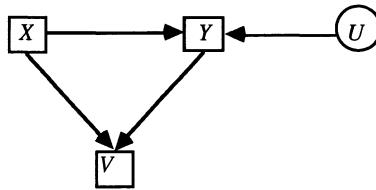


Figure 4

Here is a table of values, obtained by assuming $Y = X + U$ and $V = Y + X$, and these relations are unaltered by a direct manipulation of X :

Table 3

X	Y	V	U	Xf	$V_{Xf=1}$	$U_{Xf=1}$	$Y_{Xf=1}$
0	0	0	0	1	2	0	1
0	1	1	1	1	3	1	2
0	2	2	2	1	4	2	3
1	1	2	0	1	2	0	1
1	2	3	1	1	3	1	2
1	3	4	2	1	4	2	3

Again assume all rows are equally probable. Note that $Y_{Xf=1}$ is independent of X , and $Y_{Xf=1}$ has the same distribution as Y conditional on $X = 1$. So Pratt and Schlaifer's principle is again satisfied, and in addition the conditional probability of Y on X is invariant. The table was constructed by supposing the manipulated system satisfies the very same system of equations as the unmanipulated system, and in effect that the graph of dependencies in figure 4 is unaltered by forcing values on X .

Pratt and Schlaifer's rules, as we have reconstructed them, are consequences of the Markov Condition. So are other examples described by Rubin. To make the connection explicit we require some results. We will assume the technical definitions introduced in Chapter 3, and we will need some further definitions.

If G is a directed acyclic graph over a set of variables $V \cup W$, W is exogenous with respect to V in G , Y and Z are disjoint subsets of V , $P(V \cup W)$ is a distribution that satisfies the Markov condition for G , and **Manipulated(W) = X**, then $P(Y|Z)$ is **invariant** under direct manipulation of X in G by changing W from w_1 to w_2 if and only if $P(Y|Z, W = w_1) = P(Y|Z, W = w_2)$ wherever they are both defined. Note that a sufficient condition for $P(Y|Z)$ to be invariant under direct manipulation of X in G by changing W is that W be d-separated from Y given Z in G . In a directed acyclic graph G containing Y and Z , $ND(Y)$ is the set of all vertices that do not have a descendant in Y . If $Y \cap Z = \emptyset$, then V is in $IV(Y, Z)$ (informative variables for Y given Z) if and only if V is d-connected to Y given Z , and V is not in $ND(YZ)$. (Note that this entails that V is not in $Y \cup Z$.) If $Y \cap Z = \emptyset$, W is in $IP(Y, Z)$ (W has a parent who is an informative variable for Y given Z) if and only if W is a member of Z , and W has a parent in $IV(Y, Z) \cup Y$. We will use the following result.

Theorem 7.1: If G_{Comb} is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G_{Comb} , Y and Z are disjoint subsets of V , $P(V \cup W)$ is a distribution that satisfies the Markov condition for G_{Comb} , no member of $X \cap Z$ is a member of $\text{IP}(Y, Z)$ in G_{Unman} , and no member of $X \cap Z$ is a member of $\text{IV}(Y, Z)$ in G_{Unman} , then $P(Y|Z)$ is invariant under a direct manipulation of X in G_{Comb} by changing W from w_1 to w_2 .

The importance of Theorem 7.1 is that whether $P(Y|Z)$ is invariant under a direct manipulation of X in G_{Comb} by changing W from w_1 to w_2 is determined by properties of G_{Unman} alone. Therefore, we will sometimes speak of the invariance of $P(Y|Z)$ under a direct manipulation of X in G_{Unman} without specifying W or G_{Comb} .

Each of the preceding examples, and Pratt and Schlaifer's general rule, are consequences of a corollary to Theorem 7.1:

Corollary 7.1: If G_{Comb} is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G_{Comb} , X and Y are in V , and $P(V \cup W)$ is a distribution that satisfies the Markov condition for G_{Comb} , then $P(Y|X)$ is invariant under direct manipulation of X in G_{Comb} by changing W from w_1 to w_2 if in G_{Unman} no undirected path *into* X d-connects X and Y given the empty set of vertices. Equivalently, if (1) Y is not a (direct or indirect) cause of X , and (2) there is no common cause of X and Y in G_{Unman} .

In graphical terms, Pratt and Schlaifer's claim amounts to requiring that for "observability" (invariance) G and G' --the graph of a manipulated system obtained by removing from G all edges into X --and their associated probabilities must give the same conditional distribution of Y on X . Corollary 7.1 characterizes the sufficiency side of this claim. Pratt and Schlaifer say their condition is "almost necessary." What they mean, we take it, is that there are cases in which the antecedent of their condition fails to hold and the consequent does hold, and, furthermore, that when the antecedent fails to hold the consequent will not hold unless a special constraint is satisfied by the conditional probabilities. Parallel remarks apply to the graphical condition we have given. There exist cases in which there are d-connecting paths between X and Y given the empty set that are into X and the probability of Y when X is directly manipulated is equal to the original conditional probability of Y on X . Again the antecedent will fail and the consequent will hold only if a constraint is satisfied by the conditional probabilities, so the condition is "almost necessary."

It may happen that the distribution of Y when a value is forced on X cannot be predicted from the unforced conditional distribution of Y on X but, nonetheless, the conditional distribution of Y on Z when a value is forced on X can be predicted from the unforced conditional distribution of Y on X and Z . Pratt and Schlaifer consider the general case in which, besides X and Y , some further variables Z are measured. Pratt and Schlaifer say that the law relating Y to X is "observable with concomitant Z " when the unforced conditional distribution of Y on X and Z equals the conditional distribution of Y on Z in the population in which X is forced to have a particular value.

Pratt and Schlaifer claim sufficient and "almost necessary" conditions for observability with concomitants, namely that *for any value x of X the distribution of X be independent of the conditional distribution of $Y_{X=x}$ on the value of z of $Z_{X=x}$ when X is forced to have the value x .* This rule, too, is a special case of Theorem 7.1.

Consider an example due to Rubin. (Rubin's X is Pratt and Schlaifer's Z ; Rubin's T is Pratt and Schlaifer's X). In an educational experiment in which reading program assignments T are assigned on the basis of a randomly sampled value of some pre-test variable X which shares one or more unmeasured common causes, V , with Y , the score on a post-test, we wish to predict the average difference τ in Y values if all students in the population were given treatment $T = 1$ as against if all students were given treatment $T = 2$. The situation in the experiment is represented in figure 5.

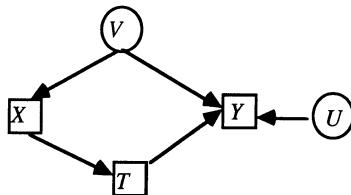


Figure 5

Provided the experimental sample is sufficiently representative, Rubin says that an unbiased estimate of τ can be obtained as follows: Let k range over values of X , from 1 to K , let \bar{Y}_{1k} be the average value of Y conditional on $T = 1$ and $X = k$, and analogously for \bar{Y}_{2k} . Let $n1k$ be the number of units in the sample with $T = 1$ and $X = k$, and analogously for $n2k$. The numbers $n1$ and $n2$ represent the total number of units in the sample with $T = 1$ and $T = 2$ respectively.

Let $\bar{Y}_{Tf=1}$ = expected value of Y if treatment 1 is forced on all units. According to Rubin, estimate $\bar{Y}_{Tf=1}$ by:

$$\sum_{k=1}^K \frac{n1k + n2k}{n1 + n2} \bar{Y1k}$$

and estimate τ by:

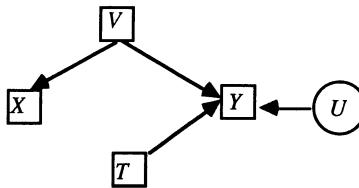
$$\sum_{k=1}^K \frac{n1k + n2k}{n1 + n2} (\bar{Y1k} - \bar{Y2k})$$

The basis for this choice may not be apparent. If we look at the hypothetical population in which every unit is forced to have $T = 1$, then it is clear from Rubin's tacit independence assumptions that he treats the manipulated population as if it had the causal structure shown in figure 6, as the following derivation shows.

$$\begin{aligned} \bar{Y}_{Tf=1} &= \overrightarrow{\sum_Y} Y \times P(Y_{Tf=1}) = \\ &\sum_Y Y \times \sum_{k=1}^K P(Y_{Tf=1} | X_{Tf=1} = k, T_{Tf=1} = 1) P(X_{Tf=1} = k | T_{Tf=1} = 1) P(T_{Tf=1} = 1) = \\ &\overrightarrow{\sum_Y} Y \times \sum_{k=1}^K P(Y_{Tf=1} | X_{Tf=1} = k, T_{Tf=1} = 1) P(X_{Tf=1} = k) \end{aligned}$$

The second equality in the above equations hold because $P(T_{Tf=1} = 1) = 1$, and $X_{Tf=1}$ and $T_{Tf=1}$ are independent according to the causal graph shown in figure 6. By Theorem 7.1, both $P(Y_{Tf=1} | X_{Tf=1}, T_{Tf=1})$ and $P(X_{Tf=1})$ are invariant under direct manipulation of T in the graph of figure 5. This entails the following equation.

$$\begin{aligned} \bar{Y}_{Tf=1} &= \overrightarrow{\sum_Y} Y \times \sum_{k=1}^K P(Y_{Tf=1} | X_{Tf=1} = k, T_{Tf=1} = 1) P(X_{Tf=1} = k) = \\ &\sum_{k=1}^K P(X = k) \times \overrightarrow{\sum_Y} Y \times P(Y | X = k, T = 1) = \frac{n1k + n2k}{n1 + n2} \times \bar{Y1k} \end{aligned}$$

**Figure 6**

Note that X and T , unlike $X_{Tf=1}$ and $T_{Tf=1}$ are *not* independent. Rubin's tacit assumption that $X_{Tf=1}$ and $T_{Tf=1}$ are independent indicates that he is implicitly assuming that the causal graph of the manipulated population is the graph of figure 6, not the graph of figure 5, which is the causal structure of the unmanipulated population. $\bar{Y}_{Tf=2}$ can be derived in an analogous fashion.

The reconstruction we have given to Rubin's theory assumes that all units in the population have the same causal structure for the relevant variables, but not, of course, that the units are otherwise homogenous. It is conceivable that someone might know the counterfactuals required for prediction according to the Pratt and Schlaifer rules even though the relevant causal structure in the population (and in the sample from which inferences are to be made) differs from unit to unit. For example, it might somehow be known that A and B have no unmeasured common cause and that B does not cause A , and the population might in fact be a mixture of systems in which A causes B and systems in which A and B are independent. In that case the distribution of B if A is forced to have the value $A = a$ can be predicted from the conditional probability of B given $A = a$, indeed the probabilities are the same. For this, and analogously for other cases of prediction for populations with a mixture of causal structures, the predictions obtained by applying Pratt and Schlaifer's rule can be derived from the Markov Condition by considering whether the relevant conditional probabilities are invariant in each of the causally homogenous subpopulations. Thus if A and B have no causal connection, $P(B|A = a)$ equals the probability of B when A is forced to have value a , and if A causes B , $P(B|A = a)$ equals the probability of B when A is forced to have value a , and so the probability is also the same in any mixture of systems with these two causal structures.

7.4 Prediction with Causal Sufficiency

The Rubin framework is specialized in two dimensions. It assumes known various counterfactual (or causal) properties, and it addresses *invariance* of conditional probability. But we very often don't know the causal structure or the counterfactuals before considering the data, and we are interested not in invariance *per se* but only as an instrument in prediction. We need to be clearer about the goal. We suppose that the investigator knows (or estimates) a distribution $P_{Unman}(\mathbf{O})$ which is the marginal over \mathbf{O} of a distribution faithful to an unknown causal graph G_{Unman} , with unknown vertex set \mathbf{V} that includes \mathbf{O} . She also knows the variable, X , that is the member of \mathbf{O} that will be directly manipulated, and the variables $Parents(G_{Man}, X)$ that will be direct causes of X in G_{Man} . She knows that X is the only variable directly manipulated. Finally she knows what the manipulation will do to X , that is, she knows $P_{Man}(X|Parents(G_{Man}, X))$. The distribution of \mathbf{Y} conditional on \mathbf{Z} is **predictable** if in these circumstances $P_{Man}(\mathbf{Y}|\mathbf{Z})$ is uniquely determined no matter what the unknown causal graph, no matter what the manipulated and unmanipulated distributions over unobserved variables, and no matter how the manipulation is brought about consistent with the assumptions just specified. The goal is to discover when the distribution of \mathbf{Y} conditional on \mathbf{Z} is predictable, and how to obtain a prediction.

The assumption that $P_{Unman}(\mathbf{O})$ is the marginal over \mathbf{O} of a distribution faithful to the unmanipulated graph G_{Unman} may fail for several reasons. First, it may fail because of the particular parameters values of the distribution. If \mathbf{W} is a set of policy variables, it also may fail because the w_2 (manipulated) subpopulation contains dependencies that are not in the w_1 (unmanipulated) subpopulation. For example, suppose that a battery is connected to a light bulb by a circuit that contains a switch. Let W be the state of the switch, w_1 be the unmanipulated subpopulation where the switch is off and w_2 be the manipulated subpopulation where the switch is on. In the w_1 subpopulation the state of the light bulb (on or off) is independent of the state of the battery (charged or not) because the bulb is always off. On the other hand in the w_2 subpopulation the state of the light bulb is dependent on the state of the battery. Hence in G_{Comb} there is an edge from the state of the battery to the state of the light bulb; it follows that there is also an edge from the state of the battery to the state of the light bulb in G_{Unman} (which is the subgraph of G_{Comb} that leaves out W .) This implies that the joint distribution over the state of the battery and the state of the light bulb in the w_1 subpopulation is not faithful to G_{Unman} . The results of the Prediction Algorithm are reliable only in circumstances where a manipulation does not introduce additional dependencies (which may or may not be part of one's background knowledge.)

Suppose we wish to make a prediction of the effect of an intervention or policy from observations of variables correctly believed to be causally sufficient for systems with a common but unknown causal structure. In that case the sample and the PC (or some other) algorithm determine a pattern representing a class of directed graphs, and properties of that class determine whether the distribution of Y following a direct manipulation of X can be predicted. Suppose for example that the pattern is $X - Y - Z$ which represents the set of graphs in figure 7.

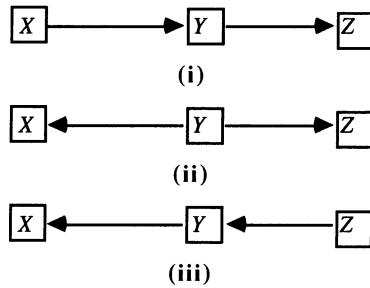


Figure 7

For each of these causal graphs, the distribution of Y after a direct manipulation of X can be calculated, but the result is different for the first graph than for the two others. $P_{Man}(Y)$ for each of the graphs can be calculated from the Manipulation Theorem and taking the appropriate marginal; the results for each graph are given below:

$$(i) \quad P_{Man}(Y) = \sum_X^{\rightarrow} P_{Unman}(Y|X)P_{Man}(X) \neq P_{Unman}(Y)$$

$$(ii) \quad P_{Man}(Y) = P_{Unman}(Y)$$

$$(iii) \quad P_{Man}(Y) = \sum_Z^{\rightarrow} P_{Unman}(Y|Z)P_{Unman}(Z) = P_{Unman}(Y)$$

If every unit in the population is forced to have the same value of X , then for (i) the manipulated distribution of Y does not equal the unmanipulated distribution of Y . For (ii) and (iii) the manipulated distribution of Y equals the unmanipulated distribution. Since the pattern does not

tell us which of these structures is correct, the distribution of Y on a manipulation of X cannot be predicted.

If a different pattern had been obtained a prediction would have been possible; for example the pattern $U - X \rightarrow Y <- Z$ can represent either of the graphs in figure 8.

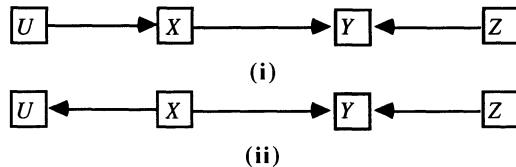


Figure 8

$P_{Man}(Y)$ for each of the graphs can be calculated from the Manipulation Theorem and taking the appropriate marginal; the results for each graph are given below:

$$(i) \quad P_{Man}(Y) = \sum_X^{\rightarrow} P_{Unman}(Y|X)P_{Man}(X)$$

$$(ii) \quad P_{Man}(Y) = \sum_X^{\rightarrow} P_{Unman}(Y|X)P_{Man}(X)$$

(Note however, that while $P_{Man}(Y)$ is the same for (i) and (ii), $P_{Man}(U,X,Y,Z)$ is not the same for (i) and (ii), so $P_{Man}(U,X,Y,Z)$ is not predictable.)

When it is known that the structure is causally sufficient, we can decide the predictability of the distribution of a variable (or conditional distribution of one set of variables on another set) by finding the pattern and applying the Manipulation Theorem and taking the appropriate marginal for every graph represented by the pattern. If all graphs give the same result, that is the prediction. Various computational shortcuts are possible, some of which are described in the Prediction Algorithm stated in the next section.

7.5 Prediction without Causal Sufficiency

We come finally to the most serious case, in which for all we know the causal structure of the manipulated systems will be different from the causal structure of the observed systems, the causal structure of the observed systems is unknown, and for all we know the observed statistical dependencies may be due to unobserved common causes. This is the situation that Mosteller and Tukey seem to think typical in non-experimental studies, and we agree. The question is whether, nonetheless, prediction is sometimes possible, and if so when and how.

Consider the following trivial example. If we have measured only smoking and lung cancer, we will find that they are correlated. The correlation could be produced by any of the three causal graphs depicted in figure 9.

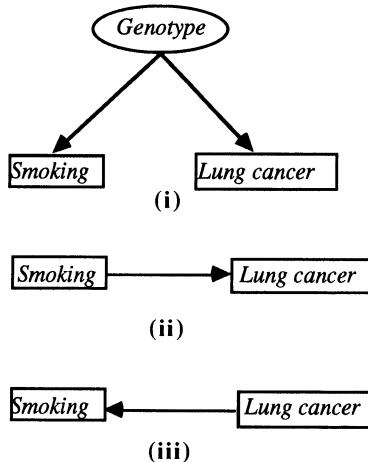
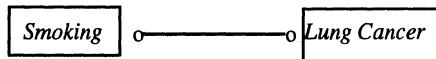


Figure 9

All three graphs yield the same maximally informative partially oriented inducing path graph, shown in figure 10.

**Figure 10**

If *Smoking* is directly manipulated in graphs (i) or (iii), then $P(\text{Lung cancer})$ will not change; but if *Smoking* is directly manipulated in graph (ii) then $P(\text{Lung cancer})$ will change. So it is not possible to predict the effects of the direct manipulation of *Smoking* from the marginal distribution of the measured variables.

In the causally sufficient case each complete orientation of the pattern yields a directed acyclic graph G . According to the Manipulation Theorem, for each directed acyclic graph G_{Unman} when we factor the distribution into a product of terms of the form $P_{\text{Unman}}(w)(V|\text{Parents}(G_{\text{Unman}}, V))$ we can calculate the effect of manipulating a variable X simply by replacing $P_{\text{Unman}}(w)(X|\text{Parents}(G_{\text{Unman}}, X))$ with $P_{\text{Man}}(w)(X|\text{Parents}(G_{\text{Man}}, X))$ (where G_{Man} is the manipulated graph). This simple substitution works because each of the terms in the factorization other than $P_{\text{Unman}}(w)(X|\text{Parents}(G_{\text{Unman}}, X))$ is guaranteed to be invariant under any direct manipulation of X in G_{Unman} , and hence can be estimated from frequencies in the unmanipulated population.

Let us now try and generalize this strategy to the causally non-sufficient case, where $P(O)$ is the marginal of a distribution $P(V)$ that is faithful to a directed acyclic graph G_{Unman} , and π is the partially oriented inducing path graph of G_{Unman} . We could search for a factorization of the distribution of $P(O)$ that is a product of terms of the form $P_{\text{Unman}}(V|M(V))$ (where membership in the set $M(V)$ is a function of V) in which each of the terms except $P_{\text{Unman}}(X|M(X))$ is invariant under all direct manipulations of X in all directed acyclic graphs for which π is a partially oriented inducing path graph over O . If we find such a factorization, then we can predict the effect of the manipulation by substituting the term $P_{\text{Man}}(X|\text{Parents}(G_{\text{Man}}, X))$ for $P_{\text{Unman}}(V|M(X))$ (where G_{Man} is the manipulated graph), just as we did in the causally sufficient case. We will not know which of the many directed acyclic graphs for which π is a partially oriented inducing path graph over O actually generated the distribution; however, it will not matter, because $P_{\text{Man}}(Y|Z)$ will be the same for each of them. This is essentially the strategy that we adopt. However, the task of finding such a factorization is considerably more difficult in the causally non-sufficient case: unlike the causally sufficient case where we can simply construct a factorization in which each term except $P(X|\text{Parents}(G_{\text{Unman}}, X))$ is invariant under direct manipulation of X in G_{Unman} , in the causally non-sufficient case we have to *search* among different factorizations in order to find a factorization in which each term

except $P_{Unman}(X|M(X))$ is invariant under all direct manipulations of X for all directed acyclic graphs G that have partially oriented inducing path graph over \mathbf{O} equal to π . Fortunately, as we will see, we do not have to search though every possible factorization of $P(\mathbf{O})$.

We will flesh out the details of this strategy and provide examples. We will use the FCI algorithm to construct a partially oriented inducing path graph π over \mathbf{O} of G_{Unman} . Note that in view of Verma and Pearl's example described in Chapter 6, it may be that some graphs for which π is a partially oriented inducing path graph over \mathbf{O} may not represent any distribution with marginal $P_{Unman}(\mathbf{O})$ because of non-independence constraints. From the theory developed in this book, we cannot hope to provide a computational procedure that decides predictability and obtains predictions whenever they are possible in principle, because we have no understanding of all constraints that graphs may entail for marginal distributions. But by considering only conditional independence constraints we can provide a sufficient condition for predictability.

Here is an example that provides a more detailed illustration of the strategy: Suppose we measure *Genotype* (G), *Smoking* (S), *Income* (I), *Parents' smoking habits* (PSH) and *Lung cancer* (L). Suppose the unmanipulated distribution is faithful to the unmanipulated graph that has the partially oriented inducing path graph shown in figure 11.

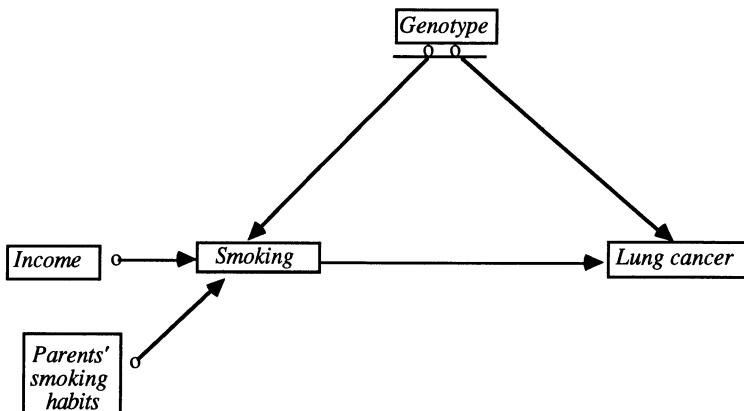


Figure 11

The partially oriented inducing path graph does not tell us whether *Income* and *Smoking* have a common unmeasured cause, or *Parents' smoking habits* and *Smoking* have a common

unmeasured cause, and so on. The measured distribution might be produced by any of several structures, including, for example those in figure 12, where T_1 and T_2 are unmeasured.

If we directly manipulate *Smoking* so that *Income* and *Parents' smoking habits* are not parents of *Smoking* in the manipulated graph, then no matter which graph produced the marginal distribution, the partially oriented inducing path graph and the Manipulation Theorem tell us that if *Smoking* is directly manipulated then in the manipulated population the resulting causal graph will look like the graph shown in figure 13.

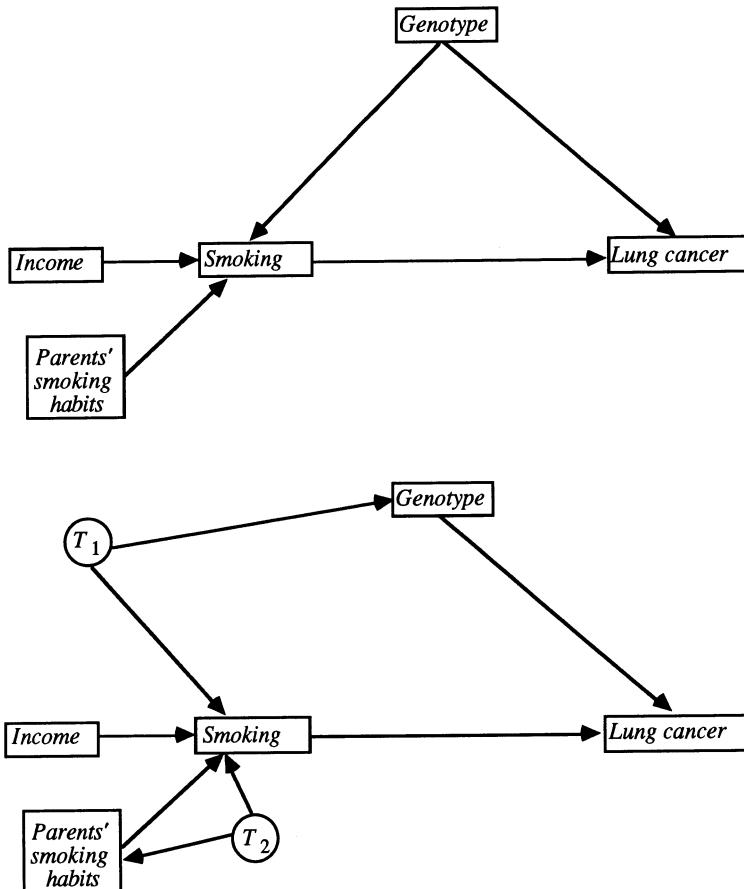


Figure 12

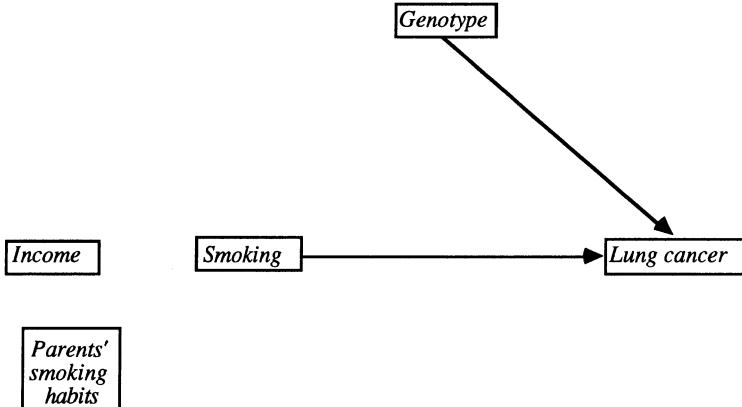


Figure 13

In this case, we can determine the distribution of *Lung cancer* given a direct manipulation of *Smoking*. Three steps are involved. Here, we simply give the results of carrying out each step. How each step is carried out is explained in more detail in the next section.

First, from the partially oriented inducing path graph we find a way to factor the joint distribution in the manipulated graph. Let P_{Unman} be the distribution on the measured variables and let P_{Man} be the distribution that results from a direct manipulation of *Smoking*. It can be determined from the partially oriented inducing path graph that

$$P_{Man}(I, PSH, S, G, L) = P_{Man}(I) \times P_{Man}(PSH) \times P_{Man}(S) \times P_{Man}(G) \times P_{Man}(L | G, S)$$

where $I = Income$, $PSH = Parents' smoking habits$, $S = Smoking$, $G = Genotype$, and $L = Lung cancer$. This is the factorization of P_{Man} corresponding to the immediately preceding graph that represents the result of a direct manipulation of *Smoking*.

Second, we can determine from the partially oriented inducing path graph which factors in the expression just given for the joint distribution are needed to calculate $P_{Man}(L)$. In this case $P_{Man}(I)$ and $P_{Man}(PSH)$ prove irrelevant and we have:

$$P_{Man}(L) = \overrightarrow{\sum}_{G,S} P_{Man}(S) \times P_{Man}(G) \times P_{Man}(L|G,S)$$

Third, we can determine from the partially oriented inducing path graph that $P_{Man}(G)$ and $P_{Man}(L|G,S)$ are equal respectively to the corresponding unmanipulated probabilities, $P_{Unman}(G)$ and $P_{Unman}(L|G,S)$. Furthermore, $P_{Man}(S)$ is assumed to be known, since it is the quantity being manipulated. Hence, all three factors in the expression for $P_{Man}(L)$ are known, and $P_{Man}(L)$ can be calculated.

Note that $P_{Man}(L)$ can be predicted even though $P(L)$ is most definitely not invariant under a direct manipulation of S . The example should be enough to show that while Mosteller and Tukey's pessimism about prediction from observation may have been justified when they wrote, it was not well-founded.

The algorithm sketched in the example is described more formally below, where we have labeled each step by a letter for easy reference. Suppose $P_{Unman}(\mathbf{V})$ is the distribution before the manipulation, $P_{Man}(\mathbf{V})$ the manipulation after the distribution, and a single variable X in \mathbf{X} is manipulated to have distribution $P_{Man}(X|\text{Parents}(G_{Man},X))$, where G_{Man} is the manipulated graph. We assume that $P_{Unman}(\mathbf{V})$ is faithful to the unmanipulated graph G_{Unman} , that $\text{Parents}(G_{Man},X)$ is known, that $P_{Man}(X|\text{Parents}(G_{Man},X))$ is known, and that we are interested in predicting $P_{Man}(\mathbf{Y}|\mathbf{Z})$. The Prediction Algorithm is simplified by the fact that if $P_{Unman}(\mathbf{O})$ satisfies the Markov Condition for a graph G_{Unman} , then so does $P_{Man}(\mathbf{O})$, and hence any factorized expression for $P_{Unman}(\mathbf{Y}|\mathbf{Z})$ is also an expression for $P_{Man}(\mathbf{Y}|\mathbf{Z})$. Recall that a total order Ord of variables in a graph G' is **acceptable** for G' if and only if whenever $A \neq B$ and there is a directed path from A to B in G' , A precedes B in Ord . If π is the FCI partially oriented inducing path graph of G over \mathbf{O} , then X is in **Definite-Non-Descendants**(\mathbf{Y}) if and only if there is no semi-directed path from any Y in \mathbf{Y} to X in π . Recall that a directed acyclic graph G is a minimal I-map of distribution P if and only if P satisfies the Markov and Minimality Conditions for G .

Prediction Algorithm

- A). $P_{Man}(\mathbf{Y}|\mathbf{Z})$ = unknown.
- B). Generate partially oriented inducing path graph π from $P_{Unman}(\mathbf{O})$.
- C). For each ordering of variables acceptable for π in which the predecessors of X in Ord equals $\text{Parents}(G_{Man},X) \cup \text{Definite-Non-Descendants}(X)$
 - C1). Form the minimal I-map F of $P_{Unman}(\mathbf{O})$ for that ordering;

- C2). Extract an expression for $P_{Unman}(Y|Z)$ from F ; call it E ;
- C3. If for each $V \neq X$, the term $P_{Unman}(V|\text{Parents}(F,V))$ in E is invariant in G_{Man} when X is directly manipulated then
- C3a). return $P_{Man}(Y|Z) = E'$, where E' is equal to E except that $P_{Unman}(X|\text{Parents}(F,X))$ is replaced by $P_{Man}(X|\text{Parents}(G_{Man},X))$
 - C3b). exit

(The algorithm can also be applied to the case where a set X of variables is manipulated, as long as it is possible to find an ordering of variables such that for each X in X all of the predecessors of X are in **Definite-Non-Descendants**(X) or **Parents**(G_{Man},X), there are no causal connections among the variables in X , and if some X in X is a parent of some variable V not in X , then every member of X is a predecessor of V .) The description leaves out important details. How can we find the partially oriented inducing path graph (step B), the graph for which $P_{Unman}(V)$ satisfies the Minimality and Markov conditions for a given ordering of variables (step C1), the expression E for $P_{Man}(Y|Z)$ (step C2); how do we determine if a given conditional probability term that appears in the expression for $P_{Unman}(Y|Z)$ is invariant under a direct manipulation of X in G_{Unman} when we do not know what G_{Unman} is (step C3)? The details are described below.

Step B: We carry out step B) with the FCI Algorithm.

Step C: Say steps C1) and C2) are *successful* if they produce an expression for $P_{Unman}(Y|Z)$ in which for every V in $O \setminus \{X\}$, $P_{Unman}(V|\text{Parents}(F,V))$ is invariant under direct manipulation of X in G_{Unman} . We conjecture that if there is an ordering of variables for which some directed acyclic graph makes C1) and C2) successful, then there is such an ordering that is acceptable for π . (Notice that the correctness of the algorithm does not depend upon the correctness of this conjecture, although if it is wrong the algorithm will be less informative than some other algorithm that searches a larger set of variable orderings.)

Step C1: For a given ordering Ord , let **Predecessors**(Ord, V) be the predecessors of V in Ord . For each V in F over O , let **Parents**(V) be the smallest subset of **Predecessors**(V) such that V is independent of **Predecessors**(Ord, V)\Parents(V) given Parents(V). Then F is a minimal I-map of $P(O)$. See Pearl (1988). Under the assumption that $P(O)$ is the marginal of a faithful distribution $P(V)$ we can test whether V is independent of **Predecessors**(Ord, V)\Parents(V) given Parents(V) by testing whether each member of **Predecessors**(Ord, V)\Parents(V) is independent of V given Parents(V). This clearly suggests testing whether small sets of variables are equal to Parents(V) first.

For inducing path graph G' and acceptable total ordering Ord , W is in $\mathbf{SP}(Ord, G', V)$ (separating predecessors of V in G' for ordering Ord) if and only if W precedes V in Ord and there is an undirected path U between W and V such that each vertex on U except for the endpoints precedes V in Ord and is a collider on U . If G is a directed acyclic graph over V , G_{IP} is the inducing path graph of G over O , Ord is an ordering acceptable for G_{IP} , and $P(V)$ is faithful to G , then the directed acyclic graph G_{Min} in which for each X in O $\mathbf{Parents}(X) = \mathbf{SP}(Ord, X)$ is a minimal I-map of $P(O)$. Of course we are not generally given G_{IP} . However, we can construct a partially oriented inducing path graph and identify sets of variables that narrow down the search for $\mathbf{SP}(Ord, X)$. For a partially oriented inducing path graph π and ordering Ord acceptable for π , let V be in $\mathbf{Possible-SP}(Ord, X)$ if and only if $V \neq X$ and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and no vertex on U except for the endpoints is a definite-non-collider on U . For a partially oriented inducing path graph π over O and ordering Ord acceptable for π , V is in $\mathbf{Definite-SP}(Ord, X)$ if and only if $V \neq X$ and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and every vertex on U except for the endpoints is a collider on U .

Theorem 7.2: If $P(O)$ is the marginal of a distribution faithful to G over V , π is a partially oriented inducing path graph of G over O , and Ord is an ordering of variables in O acceptable for some inducing path graph over O with partially oriented inducing path graph π , then there is a minimal I-map G_{Min} of $P(O)$ in which $\mathbf{Definite-SP}(Ord, X)$ in π is included in $\mathbf{Parents}(G_{Min}, X)$ which is included in $\mathbf{Possible-SP}(Ord, X)$ in π .

We can use Theorem 7.2 as a heuristic for searching for a minimal I-map of $P(O)$. The procedure is only a heuristic for the following reason. While from π we can identify orderings that are not acceptable for any inducing path graph over O with partially oriented inducing path graph π , we cannot always definitely tell that some ordering acceptable for π is acceptable for some inducing path graph over O with partially oriented inducing path graph π . For orderings not acceptable for any such inducing path graph over O , it is possible that making $\mathbf{SP}(Ord, X)$ the parents of X in G_{Min} does not make G_{Min} a minimal I-map, in which case it may be that no set M including $\mathbf{Definite-SP}(Ord, X)$ and included in $\mathbf{Possible-SP}(Ord, X)$ makes $\mathbf{Predecessors}(Ord, V) \setminus M$ independent of X given M . If that is the case, we must conduct a wider search.

Step C2: If P satisfies the Markov condition for directed acyclic graph G , the following lemma shows how to determine an expression E for $P(\mathbf{Y}|\mathbf{Z})$. (For a related result see Geiger, Verma, and Pearl 1990)

Lemma 3.3.5: If P satisfies the Markov condition for directed acyclic graph G over \mathbf{V} , then

$$P(\mathbf{Y}|\mathbf{Z}) = \frac{\sum_{\substack{\mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \\ W \in \mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}}} \prod_{W \in \mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}} P(W|\text{Parents}(W))}{\sum_{\substack{\mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y} \\ W \in \mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}}} \prod_{W \in \mathbf{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}} P(W|\text{Parents}(W))}$$

for all values of \mathbf{V} for which the conditional distributions in the factorization are defined, and for which $P(\mathbf{z}) \neq 0$.

Step C3: We use Theorems 7.3 and 7.4 below to determine from π whether a given conditional distribution is invariant under a direct manipulation of X in G_{Unman} . If π is a partially oriented inducing path graph over \mathbf{O} , then a vertex B on an undirected path U in a partially oriented inducing path graph π over \mathbf{O} is a **definite non-collider** on U if and only if B is an endpoint of U or there are edges $A \dashv \vdash B \dashv \vdash C$, $A \dashv \vdash B \rightarrow C$, or $A \leftarrow B \dashv \vdash C$ on U . If $A \neq B$, and A and B are not in \mathbf{Z} , then an undirected path U between A and B in a partially oriented inducing path graph π over \mathbf{O} is a **possibly d-connecting** path between A and B given \mathbf{Z} if and only if every collider on U is the source of a semi-directed path to a member of \mathbf{Z} , and every definite non-collider is not in \mathbf{Z} . If \mathbf{Y} and \mathbf{Z} are disjoint, then X is in **Possibly-IP**(\mathbf{Y}, \mathbf{Z}) if and only if X is in \mathbf{Z} , and there is a possibly d-connecting path between X and some Y in \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ that is not out of X . If \mathbf{Y} and \mathbf{Z} are disjoint, X is in **Possibly-IV**(\mathbf{Y}, \mathbf{Z}) if and only if X is not in \mathbf{Z} , there is a possibly d-connecting path between X and some Y in \mathbf{Y} given \mathbf{Z} , and there is a semi-directed path from X to a member of $\mathbf{Y} \cup \mathbf{Z}$. Note that Theorems 7.3 and 7.4 also entail that if there is a directed acyclic graph G for which an ordering of variables is acceptable that makes steps C1 and C2 successful, then so does the minimal I-map for which that ordering is acceptable.

Theorem 7.3: If G is a directed acyclic graph over $\mathbf{V} \cup \mathbf{W}$, \mathbf{W} is exogenous with respect to \mathbf{V} in G , \mathbf{O} is included in \mathbf{V} , G_{Unman} is the subgraph of G over \mathbf{V} , π is the FCI partially oriented inducing path graph over \mathbf{O} of G_{Unman} , \mathbf{Y} and \mathbf{Z} are included in \mathbf{O} , X is included in \mathbf{Z} , \mathbf{Y} and \mathbf{Z} are disjoint, and no X in \mathbf{X} is in **Possibly-IP**(\mathbf{Y}, \mathbf{Z}) in π , then

$P(Y|Z)$ is invariant under direct manipulation of X in G by changing the value of W from w_1 to w_2 .

Theorem 7.4: If G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , O is included in V , G_{Unman} is the subgraph of G over V , π is the FCI partially oriented inducing path graph over O of G_{Unman} , X , Y and Z are included in O , X , Y and Z are pairwise disjoint, and no X in X is in **Possibly-IV**(Y, Z) in π , then $P(Y|Z)$ is invariant under direct manipulation of X in G by changing the value of W from w_1 to w_2 .

The Prediction Algorithm is based upon the construction of a partially oriented inducing path graph from $P_{Unman}(W)(V)$. Consider the model in figure 14, where the relationships among X , Z , and T are linear in graph G_1 , and W is a policy variable.

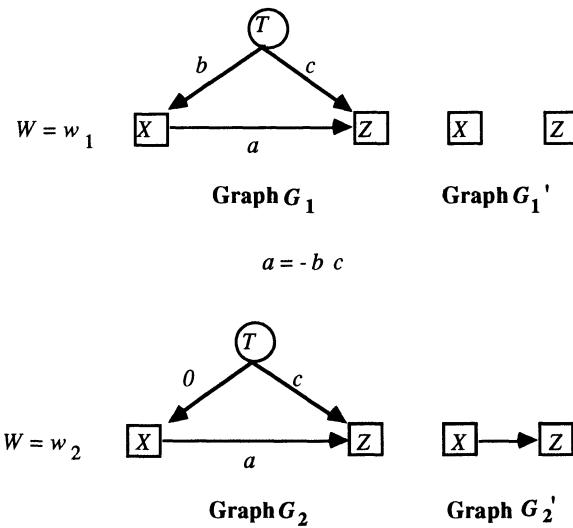


Figure 14

Although the distribution over X , Z , and T is not faithful to G_1 when $W = w_1$ if $a = -bc$, the distribution over X and Z is faithful to G_1' . In effect, although the distribution over X and Z when $W = w_1$ is faithful to a directed acyclic graph, it is not faithful to the graph of the causal process that generated the distribution. Graph G_2 depicts the model when X is directly manipulated by changing the value of W from w_1 to w_2 ; this makes the coefficient of T in the

equation for X equal to 0, and imposes some new distribution upon X . The manipulated distribution over X and Z does not satisfy the Markov condition for G_1' ; rather it satisfies the Markov condition for graph G_2' , which contains an edge between X and Z that G_1' does not contain. If we were to construct a partially oriented inducing path graph from the unmanipulated distribution over X and Z it would contain no edges, and make the prediction that the distribution of Z would be the same in the manipulated and unmanipulated distributions, we would be wrong. Hence the Prediction Algorithm is only guaranteed to be correct when the unmanipulated distribution is faithful to the unmanipulated graph (which includes the $X \rightarrow Z$ edge because the combined graph contains the $X \rightarrow Z$ edge.)

This assumption is not as restrictive as it might first appear. Suppose that we perform an experiment of the effects of *Smoking* upon *Cancer*. We decide to assign each subject a number of cigarettes smoked per day in the following way. For each subject in the experiment, we roll a die: if the die comes up 1, they are assigned to smoke no cigarettes, if the die comes up 2, they are assigned to smoke 10 cigarettes per day, etc. Let $\mathbf{W} = \{\text{Experiment}\}$ and $\mathbf{V} = \{\text{Die}, \text{Smoking}, \text{Drinking}, \text{Cancer}\}$. Figure 15 shows the causal graph for the combined population of experimental and non-experimental subjects, and G_{Unman} . The policy variable is *Experiment*: it has the same value (0) for everyone in the non-experimental population, and the same value (1) for everyone in the experimental population. *Die* is not a policy variable because it takes on different values for members of the experimental population.

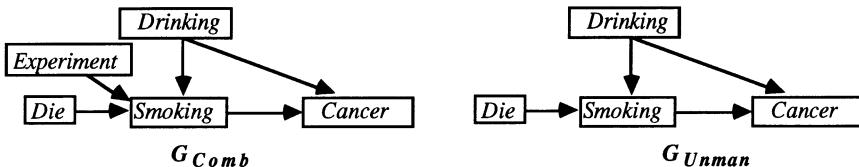


Figure 15

In this case, the assumption that $P_{\text{Unman}}(\mathbf{V})$ is faithful to G_{Unman} is clearly false because the outcome of the roll of a die and the number of cigarettes smoked by a subject are independent in the non-experimental population, but there is an edge between them in G_{Unman} . Suppose, however that we consider the subset of variables $\mathbf{V}' = \{\text{Smoking}, \text{Drinking}, \text{Cancer}\}$. The causal graphs that result from marginalizing over \mathbf{V}' are shown in figure 16. In this case, $P_{\text{Unman}}(\mathbf{V}')$ is faithful to G_{Unman} . Since variables that are causes of *Smoking* in the manipulated population but not in the unmanipulated population complicate the analysis, we will in general simply not consider them. There is no problem in leaving them out of the causal

graphs, as long as relative to the set of measured variables they are direct causes only of the manipulated variable. This guarantees that the set of variables that remain after they are removed is causally sufficient.

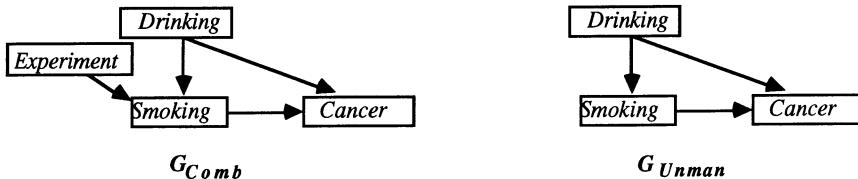


Figure 16

Theorem 7.5: If G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , G_{Unman} is the subgraph of G over V , $P_{Unman}(W)(V) = P(V|W = w_1)$ is faithful to G_{Unman} , and changing the value of W from w_1 to w_2 is a direct manipulation of X in G , then the Prediction Algorithm is correct.

The Prediction Algorithm is not complete; it may say that $P_{Man}(Y|Z)$ is unknown when it is calculable in principle.

7.6 Examples

First we consider our hypothetical example from the previous chapter, with the directed acyclic graph depicted in figure 17, and the partially oriented inducing path graph π over $O = \{Income, Parents' smoking habits, Smoking, Cilia damage, Heart disease, Lung capacity, Measured breathing dysfunction\}$ depicted in figure 18. We assume that P_{Unman} is faithful to G_{Unman} , and that in the manipulated graph that *Income* and *Parents' smoking habits* are not parents of *Smoking*. We will use the Prediction Algorithm to draw our conclusions.

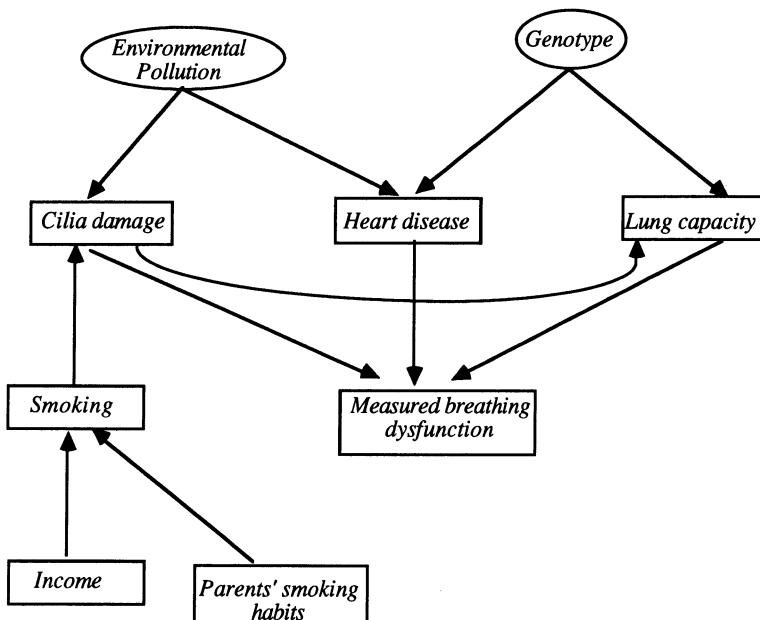


Figure 17

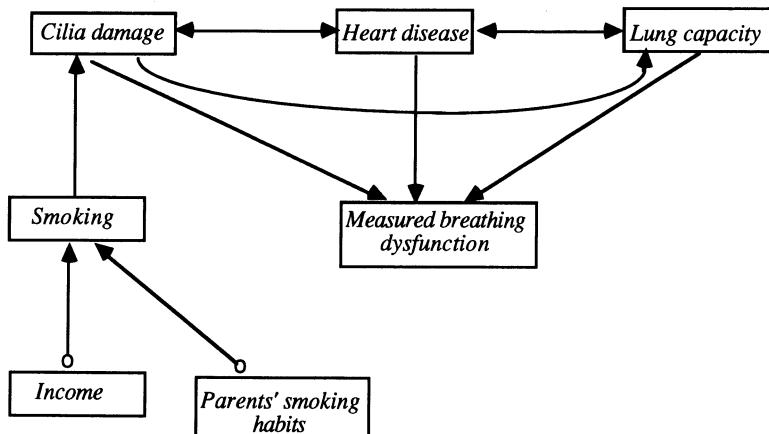


Figure 18

We will show in some detail the process of determining that the entire joint distribution of $\{Income, Parents' smoking habits, Heart disease, Lung capacity and Measured breathing dysfunction\}$ is predictable given a direct manipulation of *Smoking*. Let us abbreviate the names of the variables in the following way:

<i>Income</i>	<i>I</i>
<i>Parents' Smoking Habits</i>	<i>PSH</i>
<i>Smoking</i>	<i>S</i>
<i>Cilia damage</i>	<i>C</i>
<i>Heart disease</i>	<i>H</i>
<i>Measured breathing dysfunction</i>	<i>M</i>
<i>Lung capacity</i>	<i>L</i>

We begin by choosing an ordering for the variables. There are two constraints we impose upon the orderings. First, the only variables that precede *S* are those variables that are in **Definite-Non-Descendant**(*S*), and second, the ordering is acceptable for the partially oriented inducing path graph. That means that *I*, *PSH*, and *H* precede *S*. Second, in order to be acceptable for the partially oriented inducing path graph, *S*, *C*, *L*, and *M* have to occur in that order. We arbitrarily choose one ordering *Ord* compatible with these constraints: *I, PSH, H, S, C, L M*. (Note that the ordering among the variables that are predecessors of the directly manipulated variable never matters because each term containing only variables that are predecessors of the directly manipulated variable is always invariant.)

We generate a directed graph for which $P_{Unman}(I, PSH, S, C, H, M, LC)$ satisfies the Minimality and Markov conditions. In this case we can determine that any ordering acceptable for the partially oriented inducing path graph in figure 18 is also an ordering acceptable for the inducing path graph. Hence, we can apply Theorem 7.2. The resulting factorization is $P_{Unman}(I) \times P_{Unman}(PSH) \times P_{Unman}(H) \times P_{Unman}(S|I, PSH) \times P_{Unman}(C|S, H) \times P_{Unman}(L|C, H, S) \times P_{Unman}(M|C, H, L)$.

We now determine which terms in the factorized distribution are needed in order to predict the conditional distribution under consideration. Because we are predicting the entire joint distribution, it is trivial that we need every term in the factorized distribution.

Finally, we use the partially oriented inducing path graph to test whether each of the terms except $P_{Unman}(S|I, PSH)$ in the factorized distribution is invariant under direct manipulation of *S* in G_{Unman} . $P_{Unman}(I)$, $P_{Unman}(PSH)$, and $P_{Unman}(H)$ are invariant by Theorem 7.4 because

there are no semi-directed paths from S to I , H , or PSH . $P_{Unman}(C|S,H)$ is invariant by Theorem 7.3 because every path possibly d-connecting path S to C given H is out of S . $P_{Unman}(L|C,S,H)$, is invariant by Theorem 7.3 because every path possibly d-connecting path between S and L given C and H is out of S . Finally $P_{Unman}(M|C,H,L)$ is invariant by Theorem 7.4 because there is no possibly d-connecting path between S and M given C , H , and L .

Hence, $P_{Man}(I,PSH,H,S,C,L,M) = P_{Unman}(I) \times P_{Unman}(PSH) \times P_{Unman}(H) \times P_{Man}(S) \times P_{Unman}(C|S,H) \times P_{Unman}(L|C,H,S) \times P_{Unman}(M|C,H,L)$.

In this case, the search was simple because for the given ordering of variables, every term in the expression for $P_{Unman}(I,PSH,H,S,C,L,M)$ except for $P_{Man}(S)$ is invariant under direct manipulation of *Smoking* in G_{Unman} . If the expression had failed this test we would have repeated the process by generating different orderings of variables, until we had found a factorized expression of $P(I,PSH,H,S,C,L,M)$ in which each term except $P_{Man}(S)$ was invariant or we ran out of orderings.

For the next example, consider three alternative models of the relationship between *Smoking* and *Lung cancer* depicted in figure 19. In G_1 , *Smoking* causes *Lung cancer*, and there is a common cause of *Smoking* and *Lung cancer*; in G_2 , *Smoking* does not cause *Lung cancer*, but there is a common cause of *Lung cancer* and *Smoking*; and in G_3 , *Smoking* causes *Lung cancer*, but there is no common cause of *Smoking* and *Lung cancer*.

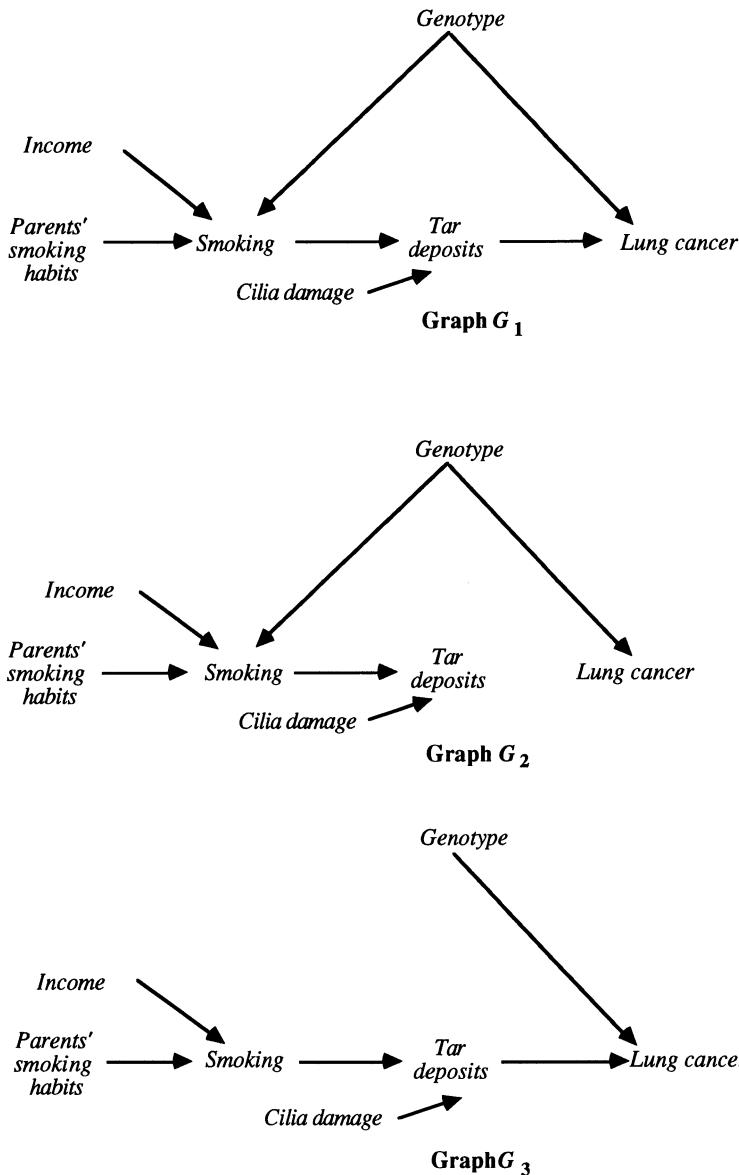


Figure 19

The maximally informative partially oriented inducing path graph of G_1 , G_2 , and G_3 over $\mathbf{O} = \{\text{Smoking}, \text{Lung cancer}\}$ is shown in figure 20.



Figure 20

From this partially oriented inducing path graph it is impossible to determine whether *Smoking* causes *Lung cancer* (as in G_3) or *Smoking* does not cause *Lung cancer* but there is a common cause of *Smoking* and *Lung cancer* (as in G_2), or *Smoking* causes *Lung cancer* and there is also a common cause (as in G_1). In addition, we cannot predict the distribution of *Lung cancer* when *Smoking* is directly manipulated. If we try the ordering of variables $\langle \text{Smoking}, \text{Lung cancer} \rangle$ then in order to apply the Prediction Algorithm, we need to show that $P(\text{Lung cancer}|\text{Smoking})$ is invariant under direct manipulation of *Smoking* in G_{Unman} . But we cannot use Theorem 7.3 to show that $P(\text{Lung cancer}|\text{Smoking})$ is invariant because the *Smoking* o-o *Lung cancer* edge guarantees that there is a possibly d-connecting path between *Smoking* and *Lung cancer* given the empty set that is not out of *Smoking*. This is a quite general feature of the method; it cannot be used to predict a conditional distribution of Y whenever there is an edge between the variable X being directly manipulated and Y that has a "o" at the X end. Of course, this feature does not of itself show that $P(\text{Lung cancer})$ is not predictable by some other method (although in this example it clearly is not.)

Suppose, however, that $\mathbf{O} = \{\text{Smoking}, \text{Lung cancer}, \text{Income}\}$. If the true graph is G_2 , the partially oriented inducing path graph is shown in figure 21.

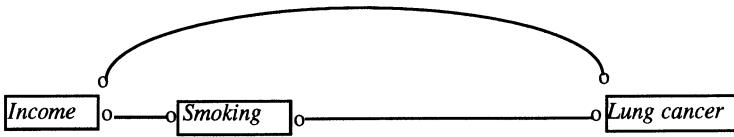


Figure 21

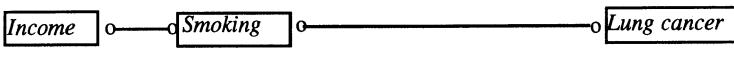
By the results of the previous chapter, we can conclude that *Smoking* does not cause *Lung cancer*, because there is no semi-directed path from *Smoking* to *Lung cancer*. In this case $P(\text{Lung cancer})$ is invariant under direct manipulation of *Smoking* in G_{Unman} , so $P_{\text{Man}}(\text{Lung cancer})$ is predictable.

The partially oriented inducing path graphs for G_1 and G_3 over $\mathbf{O} = \{\text{Lung cancer}, \text{Smoking}, \text{Income}\}$ (shown in figure 22) do not contain enough information in order to determine whether

Smoking causes Lung cancer. Because in each case there is a *Smoking o-o Lung cancer* edge it follows that we cannot use the Prediction Algorithm to predict $P_{Man}(Lung\ cancer)$.



Partially Oriented Inducing Path Graph of G_1
Over $O = \{Lung\ Cancer, Smoking, Income\}$



Partially Oriented Inducing Path Graph of G_3
Over $O = \{Lung\ Cancer, Smoking, Income\}$

Figure 22

If the true graph is G_3 it is possible to determine that *Smoking causes Lung cancer* by also measuring two causes of *Smoking* that are not directly connected in the partially oriented inducing path graph, as in figure 23. Because there is a directed path from *Smoking* to *Lung cancer* in the partially oriented inducing path graph, by the results of the preceding chapter there is a directed path from *Smoking* to *Lung cancer* in the causal graph of the process that generated the data, and *Smoking causes Lung cancer*. The output of the Prediction Algorithm is:

$$P_{Man}(Lung\ Cancer) = \sum_{Smoking}^{\rightarrow} P_{Man}(Smoking) P_{Unman}(Lung\ Cancer|Smoking)$$

Note that it is not necessary that *Parents' Smoking Habits* and *Income* be uncorrelated, or direct parents of *Smoking*. The *Smoking* to *Lung cancer* edge is oriented by any pair of variables that have edges that collide at a third variable V , that are not adjacent in the partially oriented inducing path graph, and such that there is a directed path U from V to *Smoking* and for every subpath $\langle X, Y, Z \rangle$ of U , X , Y , and Z do not form a triangle.

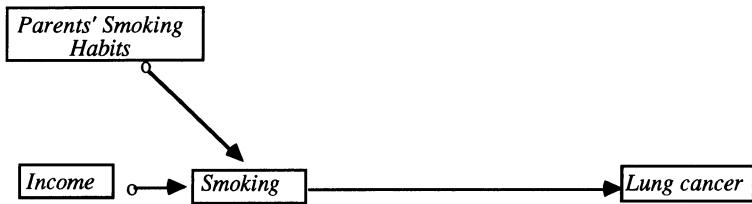


Figure 23

Unfortunately, it is more difficult to determine whether *Smoking* is a cause of *Lung cancer* if G_1 is the true causal graph. If $\mathbf{O} = \{\textit{Smoking}, \textit{Lung cancer}, \textit{Income}, \textit{Parents' Smoking Habits}\}$ and G_1 is the true causal graph, without further background knowledge we cannot determine whether *Smoking* causes *Lung cancer*. Figure 24 shows that in the partially oriented inducing path graph the *Smoking* to *Lung cancer* edge is in triangles with *Income* and *Parents' smoking habits* and hence is oriented with an 'o' at each end. It follows from the existence of the *Smoking* o-o *Lung cancer* edge that we cannot use the Prediction Algorithm to predict $P(\textit{Lung cancer})$ when *Smoking* is directly manipulated.

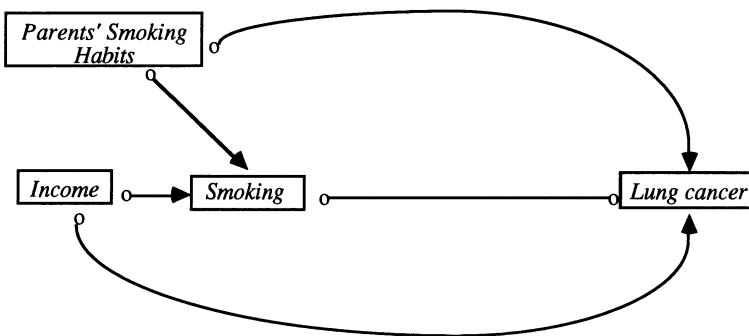


Figure 24

It is plausible that *Income* does not cause *Lung cancer* directly. If we know from background knowledge that if there is a causal connection between *Income* and *Lung cancer* it contains a causal path from *Smoking* to *Lung cancer*, then we can conclude from the partially oriented inducing path graph that *Smoking* does cause *Lung cancer*.

Alternatively, if G_1 is the correct model, we might try to determine that *Smoking* is a cause of cancer by measuring a variable such as *Tar deposits*, that is causally between *Smoking* and *Lung cancer*. While there is still an induced edge between *Income* and *Lung Cancer* in the partially oriented inducing path graph, *Income*, *Smoking*, and *Tar deposits* are not in a triangle, and the edge from *Smoking* to *Tar deposits* can be oriented. Unfortunately, as figure 25 illustrates, this now leaves one end of the edge between *Tar deposits* and *Lung cancer* oriented with a "o" at one end, so the partially oriented inducing path graph still does not entail that *Smoking* causes *Lung cancer*. And because there is a *Smoking* o-o *Lung cancer* edge, $P_{Man}(Lung\ cancer)$ is not predictable using the Prediction Algorithm.

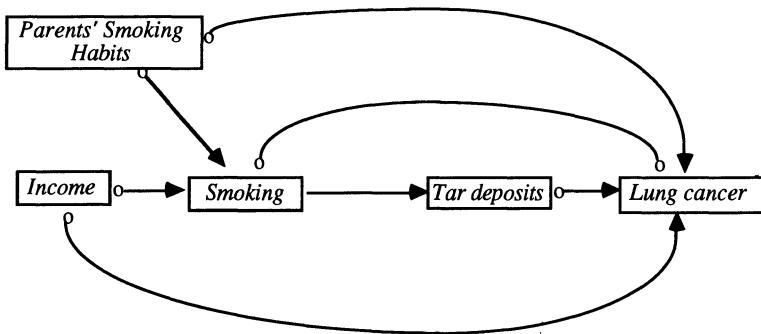


Figure 25

However, if G_1 is the correct model, and we measure a variable between *Smoking* and *Lung cancer*, such as *Tar deposits*, and another cause of *Tar deposits*, such as *Cilia damage*, we can determine that *Smoking* causes *Lung cancer*. See figure 26. However, we cannot predict $P_{Man}(Lung\ cancer)$ using the Prediction Algorithm because of the *Smoking* o-> *Lung cancer* edge.

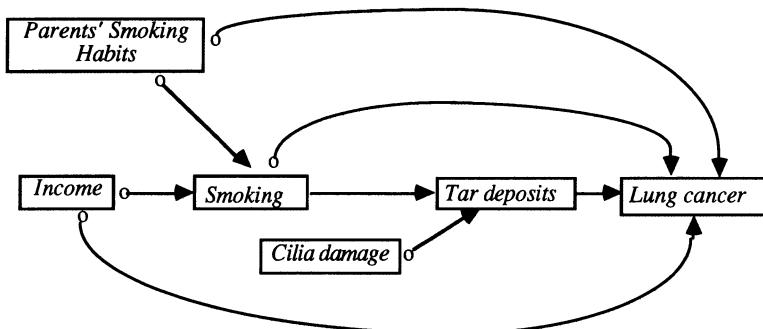


Figure 26

We can also determine that *Smoking* is a cause of *Lung cancer* by breaking the *Income-Smoking-Lung cancer* triangle by measuring all of the common causes of *Smoking* and *Lung cancer* (in this case, *Genotype*). By measuring all of the common causes of *Smoking* and *Lung cancer*, the edge between *Income* and *Lung cancer* is removed from the partially oriented inducing path graph. This breaks triangles involving *Income*, *Smoking*, and *Lung cancer*, so that the *Smoking* to *Lung cancer* edge can be oriented by the edge between *Income* and *Smoking*, as in figure 27. In addition, $P_{Man}(Lung\ cancer)$ is predictable. The output of the Prediction Algorithm is:

$$P_{Man}(Lung\ Cancer) = \sum_{Smoking, Genotype} P_{Man}(Smoking) P_{Unman}(Genotype) P_{Unman}(Lung\ Cancer | Smoking, Genotype)$$

Of course, measuring *all* of the common causes of *Smoking* and *Lung cancer* may be difficult both because of the number of such common causes, and because of measurement difficulties (as in the case of *Genotype*). So long as even one common cause remains unmeasured, the inducing path graph has an *Income - Smoking - Lung cancer* triangle, and the edge between *Smoking* and *Lung cancer* cannot be oriented.

Although we cannot determine from the partially oriented inducing path graph in figure 27 whether *Genotype* is a common cause of *Smoking* and *Lung cancer*, we can determine that there is *some* common cause of *Smoking* and *Lung cancer*.

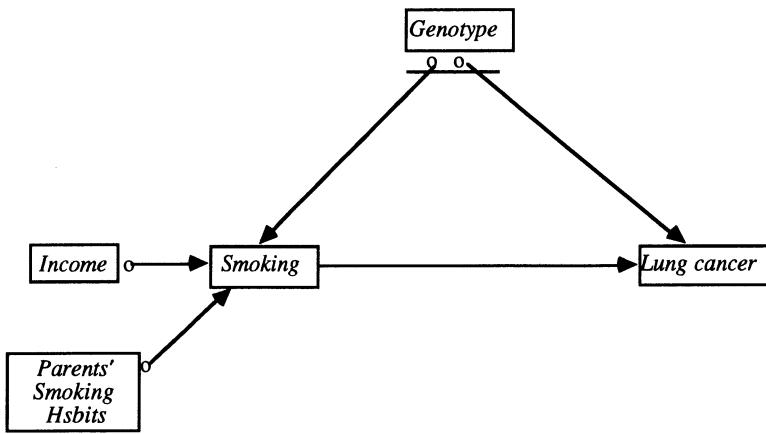


Figure 27

7.7 Conclusion

The results developed here show that there exist possible cases in which predictions of the effects of manipulations can be obtained from observations of unmanipulated systems, and predictions of experimental outcomes can be made from uncontrolled observations. Some examples from real data analysis problems will be considered in the next chapter. We do not know whether our sufficient conditions for prediction are close to maximally informative, and a good deal of theoretical work remains to be done on the question.

7.8 Background Notes

While there are of course many applications that conform to it, we have not been able to find anticipations of the theory developed in this chapter except in the tradition of work inaugurated by Rubin. The special case of the Manipulation Theorem that applies when an intervention makes a single directly manipulated variable X independent of its parents was independently conjectured by Fienberg in a seminar in 1991.

Chapter 8

Regression, Causation and Prediction

Regression is a special case, not a special subject. The problems of causal inference in regression studies are instances of the problems we have considered in the previous chapters, and the solutions are to be found there as well. What is singular about regression is only that a technique so ill suited to causal inference should have found such wide employment to that purpose.

8.1 When Regression Fails to Measure Influence

Regression models are commonly used to estimate the "influence" that regressors \mathbf{X} have on an outcome variable, Y .¹ If the relations among the variables are linear then for each X_i the expected change in Y that would be produced by a unit change in X_i if all other \mathbf{X} variables are forced to be constant can be represented by a parameter, say α_i . It is obvious and widely noted (see, for example, Fox, 1984) that the regression estimate of α_i will be incorrect if X_i and Y have one or more unmeasured common causes, or in more conventional statistical terminology, the estimate will be biased and inconsistent if the error variable for Y is correlated with X_i . To avoid such errors, it is often recommended (Pratt and Schlaifer, 1988) that investigators enlarge the set of potential regressors and determine if the regression coefficients for the original

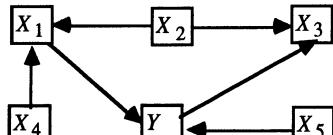
¹In linear regression, we understand the "direct influence" of X_i on Y to mean (i) the change in value of a variable Y that would be produced in each member of a population by a unit change in X_i , with all other \mathbf{X} variables forced to be unchanged. Other meanings might be given, for example: (ii) the population *average* change in Y for unit change in X_i , with all other \mathbf{X} variables forced to be unchanged; (iii) the change in Y in each member of the population for unit change in X_i ; (iv) the population average change in Y for unit change in X_i ; etc. Under interpretations (iii) and (iv) the regression coefficient is an unreliable estimate whenever X_i also influences other regressors that influence Y . Interpretation (ii) is equivalent to (i) if the units are homogeneous and the stochastic properties are due to sampling; otherwise, regression will be unreliable under interpretation (i) except in special cases, e.g., when the linear coefficients, as random variables, are independently distributed (in which case the analysis given here still applies (Glymour, Spirtes and Scheines, 1991a)).

regressors remain stable, in the hope that confounding common causes, if any, will thereby be measured and revealed. Regression estimates are known often to be unstable when the number of regressors is enlarged, because, for example, additional regressors may be common causes of previous regressors and the outcome variable (Mosteller and Tukey, 1977). The stability of a regression coefficient for X when other regressors are added is taken to be evidence that X and the outcome variable have no common cause.

It does not seem to be recognized, however, that when regressors are statistically dependent, the existence of an unmeasured common cause of regressor X_i and outcome variable Y may bias estimates of the influence of *other* regressors, X_k ; variables having no influence on Y whatsoever, nor even a common cause with Y , may thereby be given significant regression coefficients. The error may be quite large. The strategy of regressing on a larger set of variables and checking stability may compound rather than remedy this problem. A similar difficulty may arise if one of the measured candidate regressors is an *effect*, rather than a cause, of Y , a circumstance that we think may sometimes occur in uncontrolled studies.

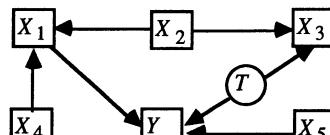
To illustrate the problem, consider the linear structures in figure 1, where for concreteness we specify that exogenous and error variables are all uncorrelated and jointly normally distributed, the error variables have zero means, and linear coefficients are not zero. Only the X variables and Y are assumed to be measured. Each set of linear equations is accompanied by a directed graph illustrating the assumed causal and functional dependencies among the non-error variables. In large samples, for data from each of these structures linear multiple regression will give all variables in the set $\{X_1, X_2, X_3, X_5\}$ non-zero regression coefficients, even though X_2 has no direct influence on Y in any of these structures, and X_3 has no influence direct or indirect on Y in structures (i), and (ii), and the effect of X_3 in structures (iii) and (iv) is confounded by an unmeasured common cause. The regression estimates of the influences of X_2 and X_3 will in all four cases be incorrect. If a specification search for regressors had selected X_1 alone or X_1 and X_5 in (i) or (ii), or X_5 alone in (i), (ii), (iii), or (iv), a regression on these variables would give consistent, unbiased estimates of their influence on Y . But the textbook procedures in commercial statistical packages will in all of these cases fail to identify $\{X_1\}$ or $\{X_5\}$ or $\{X_1, X_5\}$ as the appropriate subset of regressors.

$$\begin{aligned}Y &= a_1 X_1 + a_2 X_5 + \varepsilon_Y \\X_1 &= a_3 X_2 + a_4 X_4 + \varepsilon_1 \\X_3 &= a_5 X_2 + a_6 Y + \varepsilon_3\end{aligned}$$



(i)

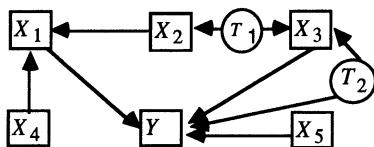
$$\begin{aligned}Y &= a_1 X_1 + a_2 X_5 + a_3 T + \varepsilon_Y \\X_1 &= a_4 X_2 + a_5 X_4 + \varepsilon_1 \\X_3 &= a_6 X_2 + a_7 T + \varepsilon_3\end{aligned}$$



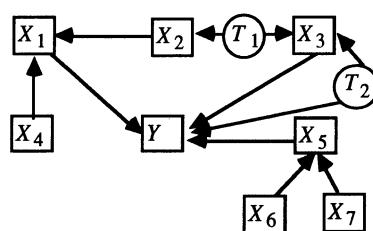
(ii)

$$\begin{aligned}Y &= a_1 X_1 + a_2 X_5 + a_3 T_2 + a_4 X_3 + \varepsilon_Y \\X_1 &= a_5 X_2 + a_6 X_4 + \varepsilon_1 \\X_2 &= a_7 T_1 + \varepsilon_2 \\X_3 &= a_8 T_1 + a_9 T_2 + \varepsilon_3\end{aligned}$$

$$\begin{aligned}Y &= a_1 X_1 + a_2 X_5 + a_3 T_2 + a_4 X_3 + \varepsilon_Y \\X_1 &= a_5 X_2 + a_6 X_4 + \varepsilon_1 \\X_2 &= a_7 T_1 + \varepsilon_2 \\X_3 &= a_8 T_1 + a_9 T_2 + \varepsilon_3 \\X_5 &= a_{10} X_6 + a_{11} X_7 + \varepsilon_5\end{aligned}$$



(iii)



(iv)

Figure 1

It is easy to produce examples of the difficulty by simulation. Using structure (i), twenty sets of values for the linear coefficients were generated, half positive and half negative, each with absolute value greater than .5. For each system of coefficient values a random sample of 5,000 units was generated by substituting those values for the coefficients of structure (i) and using uncorrelated standard normal distributions for the exogenous variables. Each sample was given to MINITAB, and in all cases MINITAB found that $\{X_1, X_2, X_3, X_5\}$ is the set of regressors with significant regression coefficients. The STEPWISE procedure in MINITAB selected the same set in approximately half the cases and in the others added X_4 to boot; selection by the lowest value of Mallow's C_p or adjusted R^2 gave results similar to the STEPWISE procedure.

The difficulty can be remedied if one measures all common causes of the outcome variable and the candidate regressors, but unfortunately nothing in regression methods informs one as to when that condition has been reached. And the addition of extra candidate regressors may create the problem rather than remedy it; in structures (i) and (ii), if X_3 were not measured the regression estimate of X_2 would be consistent and unbiased.

The problem we have illustrated is quite general; it will lead to error in the estimate of the influence of any regressor X_k that directly causes or has a common direct unmeasured common cause with any regressor X_i such that X_i and Y have an unmeasured common cause (or X_i is an effect of Y). Depending on the true structure and coefficient values the error may be quite large. It is easy to construct cases in which a variable with no influence on the outcome variable has a standardized regression coefficient larger than any other single regressor. Completely parallel problems arise for categorical data. Recall Theorem 3.4:

Theorem 3.4: If P is faithful to some graph, then P is faithful to G if and only if

- (i) for all vertices, X, Y of G , X and Y are adjacent if and only if X and Y are dependent conditional on every set of vertices of G that does not include X or Y ; and
- (ii) for all vertices X, Y, Z such that X is adjacent to Y and Y is adjacent to Z and X and Z are not adjacent, $X \rightarrow Y \leftarrow Z$ is a subgraph of G if and only if X, Z are dependent conditional on every set containing Y but not X or Z .

Consideration of the first part of this theorem explains why in structure (i) in figure 1 regression procedures incorrectly select X_2 as a variable directly influencing Y : The structure and distribution satisfy the Markov and Faithfulness conditions, but linear regression takes a variable X_i to influence Y provided the partial correlation of X_i and Y controlling for *all* of the other X variables does not vanish. Part (i) of Theorem 3.4 shows that the regression criterion is insufficient. It follows immediately from Theorem 3.4 that, assuming the Markov and Faithfulness Conditions, regression of Y on a set \mathbf{X} of variables will only yield an unbiased or consistent estimate of the influences of the \mathbf{X} variables provided in the true structure no \mathbf{X} variable is the effect of Y or has a common unmeasured cause with Y .

Since typical empirical data sets to which multiple regression methods are applied have some correlated regressors, and in uncontrolled studies it is rare to know that unmeasured common causes are not acting on both the outcome variable and the regressors, the problem is endemic. One of the most common uses of statistical methods thus appears to be little more than elaborate guessing.

8.2 A Solution and Its Application

Assuming the right variables have been measured, there is a straightforward solution to these problems: apply the PC, FCI, or other reliable algorithm, and appropriate theorems from the preceding chapters, to determine which X variables influence the outcome Y , which do not, and for which the question cannot be answered from the measurements; then estimate the dependencies by whatever methods seem appropriate and apply the results of the previous chapter to obtain predictions of the effect of manipulating the X variables. No extra theory is required. We will give a number of illustrations, both empirical and simulated.

We begin by noting that for the twenty samples from structure (i), in every case our implementation of the PC algorithm--which of course assumes there are no latent variables--selects $\{X_1, X_5\}$ as the variables that directly influence Y . Our implementation of the FCI algorithm, which makes no such assumption, in every case says that X_1 directly influences Y , that X_5 may, and that the other variables do not.

In each of the other three structures in figure 1 with sufficiently large samples multiple regression methods will make comparable errors, always including X_2 and X_3 among the "significant" or "best" or "important" variables. In contrast the FCI algorithm together with Theorems 6.5 through 6.8 give the following results:

Structure	Direct Influence	No Direct Influence	Undetermined
(ii)	X_1	X_2, X_3, X_4	X_5
(iii)	X_1	X_4	X_2, X_3, X_5
(iv)	X_1, X_5	X_4, X_6, X_7	X_2, X_3

In all of these cases the FCI procedure either determines definitely that X_2 and X_3 have no direct influence on Y , or determines that it cannot be decided whether they have any unconfounded direct influence.

8.2.1 Components of the Armed Forces Qualification Test

The *AFQT* is a test battery used by the United States armed forces. It has a number of component tests, including those listed below:

Arithmetical Reasoning (*AR*)

Numerical Operations (*NO*)

Word Knowledge (*WK*)

In addition a number of other tests, including those listed below, are not part of the *AFQT* but are correlated with it and with its components:

Mathematical Knowledge (*MK*)

Electronics Information (*EI*)

General Science (*GS*)

Mechanical Comprehension (*MC*)

Given scores for these 8 measures on 6224 armed forces personnel, a linear multiple regression of *AFQT* on the other seven variables gives significant regression coefficients to all seven and thus fails to distinguish the tests that are in fact linear components of *AFQT*. The covariance matrix is shown below.

n = 6224

<i>AFQT</i>	<i>NO</i>	<i>WK</i>	<i>AR</i>	<i>MK</i>	<i>EI</i>	<i>MC</i>	<i>GS</i>
253.9850							
29.6490	51.7649						
60.3604	6.2931	41.967					
57.6566	14.5143	16.0226	40.9329				
29.3763	18.2701	13.2055	20.6052	40.7386			
36.2318	2.10733	22.6958	16.3664	12.1773	63.1039		
35.8244	4.45539	17.4155	20.3952	16.459	35.1981	62.9647	
38.2510	5.61516	27.1492	14.7402	14.8442	29.9095	26.6842	48.9300

Given the prior information that *AFQT* is not a cause of any of the other variables, the PC algorithm in TETRAD II correctly picks out {*AR*, *NO*, *WK*} as the only variables adjacent to

AFQT, and hence the only variables that can be components of *AFQT*. (Spirtes, Glymour, Scheines and Sorensen, 1990).²

8.2.2 The Causes of Spartina Biomass

A recent textbook on regression (Rawlings 1988) skillfully illustrates regression principles and techniques for a biological study in which it is reasonable to think there is a causal process at work relating the variables. The question at issue is plainly causal: among a set of 14 variables, which have the most influence on an outcome variable, the weight of Spartina grass? Since the example is the principal application given for an entire textbook on regression, the reader who reaches the 13th chapter may be surprised to find that the methods yield almost no useful information about that question.

According to Rawlings, Linthurst (1979) obtained five samples of Spartina grass and soil from each of nine sites on the Cape Fear Estuary of North Carolina. Besides the mass of Spartina (*BIO*), fourteen variables were measured for each sample:

- Free Sulfide (H_2S)
- Salinity (*SAL*)
- Redox potentials at pH 7 (*EH₇*)
- Soil pH in water (*PH*)
- Buffer acidity at pH 6.6 (*BUF*)
- Phosphorus concentration (*P*)
- Potassium concentration (*K*)
- Calcium concentration (*CA*)
- Magnesium concentration (*MG*)
- Sodium concentration (*NA*)
- Manganese concentration (*MN*)
- Zinc concentration (*ZN*)
- Copper concentration (*CU*)
- Ammonium concentration (*NH₄*)

²In fact, we were inadvertently misinformed that all seven tests are components of *AFQT* and we first discovered otherwise with the SGS algorithm.

The correlation matrix is as follows³:

<i>BIO</i>	<i>H₂S</i>	<i>SAL</i>	<i>EH₇</i>	<i>PH</i>	<i>BUF</i>	<i>P</i>	<i>K</i>	<i>CA</i>	<i>MG</i>	<i>NA</i>	<i>MN</i>	<i>ZN</i>	<i>CU</i>	<i>NH₄</i>
1.0														
.33	1.0													
-.10	.10	1.0												
.05	.40	.31	1.0											
.77	.27	-.05	.09	1.0										
-.73	-.37	-.01	-.15	-.95	1.0									
-.35	-.12	-.19	-.31	-.40	.38	1.0								
-.20	.07	-.02	.42	.02	-.07	-.23	1.0							
.64	.09	.09	-.04	.88	-.79	-.31	-.26	1.0						
-.38	-.11	-.01	.30	-.18	.13	-.06	.86	-.42	1.0					
-.27	.00	.16	.34	-.04	-.06	-.16	.79	-.25	.90	1.0				
-.35	.14	-.25	-.11	-.48	.42	.50	-.35	-.31	-.22	-.31	1.00			
-.62	-.27	-.42	-.23	-.72	.71	.56	.07	-.70	.35	.12	.60	1.0		
.09	.01	-.27	.09	.18	-.14	-.05	.69	-.11	.71	.56	-.23	.21	1.0	
-.63	-.43	-.16	-.24	-.75	.85	.49	-.12	-.58	.11	-.11	.53	.72	.01	1.0

The aim of the data analysis was to determine for a later experimental study which of these variables most influenced the biomass of Spartina in the wild. Greenhouse experiments would then try to estimate causal dependencies out in the wild. In the best case one might hope that the statistical analyses of the observational study would correctly select variables that influence the growth of Spartina in the greenhouse. In the worst case, one supposes, the observational study would find the wrong causal structure, or would find variables that influence growth in the wild (e.g., by inhibiting or promoting growth of a competing species) but have no influence in the greenhouse.

Using the SAS statistical package, Rawlings analyzed the variable set with a multiple regression and then with two stepwise regression procedures. A search through all possible subsets of regressors was not carried out, presumably because the candidate set of regressors is too large. The results were as follows:

³The correlation matrix given in Rawlings (1988) incorrectly gives the correlation between *CU* and *NH₄* as 0.93.

- (i) a multiple regression of *BIO* on all other variables gives only *K* and *CU* significant regression coefficients;
- (ii) two stepwise regression procedures⁴ both yield a model with *PH*, *MG*, *CA* and *CU* as the only regressors, and multiple regression on these variables alone gives them all significant coefficients;
- (iii) simple regressions one variable at a time give significant coefficients to *PH*, *BUF*, *CA*, *ZN* and *NH₄*.

What is one to think? Rawling's reports that "None of the results was satisfying to the biologist; the inconsistencies of the results were confusing and variables expected to be biologically important were not showing significant effects." (p. 361). This analysis is supplemented by a ridge regression, which increases the stability of the estimates of coefficients, but the results for the point at issue--identifying the important variables--are much the same as with least squares. Rawlings also provides a principal components factor analysis and various geometrical plots of the components. These calculations provide no information about which of the measured variables influence Spartina growth.

Noting that *PH*, for example, is highly correlated with *BUF*, and using *BUF* instead of *PH* along with *MG*, *CA* and *CU* would also result in significant coefficients, Rawlings effectively gives up on this use of the procedures his book is about:

Ordinary least squares regression tends either to indicate that none of the variables in a correlated complex is important when all variables are in the model, or to arbitrarily choose one of the variables to represent the complex when an automated variable selection technique is used. A truly important variable may appear unimportant because its contribution is being usurped by variables with which it is correlated. Conversely, unimportant variables may appear important because of their associations with the real causal factors. It is particularly dangerous in the presence of collinearity to use the regression results to impart a "relative importance," whether in a causal sense or not, to the independent variables. (p. 362)

Rawling's conclusion is correct about multiple regression and about conventional methods for choosing regressors, but it is not true of more reliable inference procedures. If we apply the PC

⁴The "maximum R-square" and "stepwise" options in PROC REG in the SAS program.

algorithm to the Linthurst data then there is one robust conclusion: the only variable that may directly influence biomass in this population⁵ is *PH*; *PH* is distinguished from all other variables by the fact that the correlation of every other variable (except *MG*) with *BIO* vanishes or vanishes when *PH* is controlled for.⁶ The relation is not symmetric; the correlation of *PH* and *BIO*, for example, does not vanish when *BUF* is controlled. The algorithm finds *PH* to be the only variable adjacent to *BIO* no matter whether we use a significance level of .05 to test for vanishing partial correlations, or a level of 0.1, or a level of 0.2. In all of these cases, the PC algorithm or the FCI algorithm yield the result that *PH* and only *PH* can be directly connected with *BIO*. If the system is linear normal and the Causal Markov Condition obtains, then in this population any influence of the other regressors on *BIO* would be blocked if *PH* were held constant. Of course, over a larger range of values of the variables there is little reason to think that *BIO* depends linearly on the regressors, or that factors that have no influence in producing variation within this sample would continue to have no influence. Nor can the analysis determine whether the relationship between *PH* and *BIO* is confounded by one or more unmeasured common causes, but the principles of the theory in this case suggest otherwise. If *PH* and *BIO* have a common unmeasured cause *T*, say, and any other variable, *Z*, among the 13 others either causes *PH* or has a common unmeasured cause with *PH*, then *Z* and *BIO* should be correlated conditional on *PH*, which appears not to be the case.

The program and theory lead us to expect that if *PH* is forced to have values like those in the sample—which are almost all either below *PH* 5 or above *PH* 7—then manipulations of other variables within the ranges evidenced in the sample will have no effect on the growth of Spartina. The inference is a little risky, since growing plants in a greenhouse under controlled conditions may not be a direct manipulation of the variables relevant to growth in the wild. If for example, in the wild variations in *PH* affect Spartina growth chiefly through their influence on the growth of competing species not present in the greenhouse, a greenhouse experiment will not be a direct manipulation of *PH* for the system.

The fourth chapter of Linthurst's thesis partly confirms the PC algorithm's analysis. In the experiment Linthurst describes, samples of Spartina were collected from a salt marsh creekbank (presumably at a different site than those used in the observational study). Using a 3 x 4 x 2 (*PH* x *SAL* x *AERATION*) randomized complete block design with four blocks, after transplantation to a greenhouse the plants were given a common nutrient solution with varying

⁵Although the definition of the population in this case is unclear, and must in any case be drawn quite narrowly.

⁶More exactly, at .05, with the exception of *MG* the partial correlation of every regressor with *BIO* vanishes when some set containing *PH* is controlled for; the correlation of *MG* with *BIO* vanishes when *CA* is controlled for.

values *PH* and *SAL* and *AERATION*. The *AERATION* variable turned out not to matter in this experiment. Acidity values were *PH* 4, 6 and 8. *SAL* for the nutrient solutions was adjusted to 15, 25, 35 and 45 %.

Linthurst found that growth varied with *SAL* at *PH* 6 but not at the other *PH* values, 4 and 8, while growth varied with *PH* at all values of *SAL* (p. 104). Each variable was correlated with plant mineral levels. Linthurst considered a variety of mechanisms by which extreme *PH* values might control plant growth:

At *pH* 4 and 8, salinity had little effect on the performance of the species. The *pH* appeared to be more dominant in determining the growth response. However, there appears to be no evidence for any causal effects of high or low tissue concentrations on plant performance unless the effects of *pH* and salinity are also accounted for. (p.108)

The overall effect of *pH* at the two extremes is suggestive of damage to the root directly, thereby modifying its membrane permeability and subsequently its capacity for selective uptake. (p. 109).

A comparison of the observational and experimental data suggests that the PC Algorithm result was essentially correct and can be extrapolated through the variation in the populations sampled in the two procedures, but cannot be extrapolated through *PH* values that approach neutrality. The result of the PC search was that in the non-experimental sample, observed variations in aerial biomass were perhaps caused by variations in *PH*, but were not caused by variations in other variables. In the observational data Rawlings reports (p. 358) almost all *SAL* measurements are around 30--the extremes are 24 and 38. Compared to the experimental study rather restricted variation was observed in the wild sample. The observed values of *PH* in the wild, however, are clustered at the two extremes; only four observations are within half a *PH* unit of 6, and no observations at all occurred at *PH* values between 5.6 and 7.1. For the observed values of *PH* and *SAL*, the experimental results appear to be in very good agreement with our results from the observational study: small variations in *SAL* have no effect on *Spartina* growth if the *PH* value is extreme.

8.2.3 The Effects of Foreign Investment on Political Repression

Timberlake and Williams (1984) used regression to claim foreign investment in third-world countries promotes dictatorship. They measured political exclusion (*PO*) (i.e., dictatorship),

foreign investment penetration in 1973 (*FI*), energy development in 1975 (*EN*), and civil liberties (*CV*). Civil liberties was measured on an ordered scale from 1 to 7, with lower values indicating greater civil liberties. Their correlations for 72 "non-core" countries are:

<i>PO</i>	<i>FI</i>	<i>EN</i>	<i>CV</i>
1.0			
-.175	1.0		
-.480	.330	1.0	
.868	-.391	-.430	1.0

Their inference is unwarranted. Their model and the model obtained from the SGS algorithm using a .12 significance level to test for vanishing partial correlations) are shown in figure 2.⁷

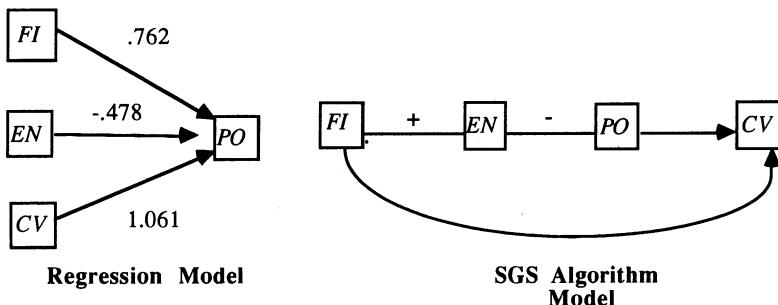
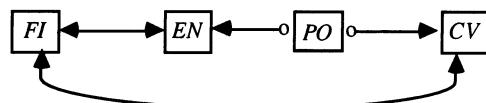


Figure 2

The SGS Algorithm will not orient the *FI-EN* and *EN-PO*, edges, or determine whether they are due to at least one unmeasured common cause. Maximum likelihood estimates of any of the SGS Algorithm models require that the influence of *FI* on *PO* (if any) be negative, and the models easily pass a likelihood ratio test with the EQS program. If one of the SGS Algorithm models is correct, Timberlake and William's regression model appears to be a case in which an effect of the outcome variable is taken as a regressor, as in structure (i) of figure 1.

This analysis of the data assumes there are no unmeasured common causes. If we run the correlations through the FCI algorithm using the same significance level, we obtain the following partially oriented inducing path graph:

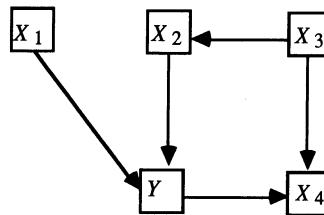
⁷Searches at lower significance levels remove the adjacency between *FI* and *EN*.

**Figure 3**

The graph together with the required signs of the dependencies, says that foreign investment and energy consumption have a common cause, as do foreign investment and civil liberties, that energy development has no influence on political exclusion, but political exclusion may have a negative effect on energy development, and that foreign investment has no influence, direct or indirect, on political exclusion.

8.2.4 More Simulation Studies

In the following simulation study we used data generated from the graph of figure 4, which illustrates some of the confusions that seem to be present in the regression produced by Timberlake and Williams.

**Figure 4**

For both the linear and the discrete cases with binary variables, one hundred trials were run at each of sample sizes 2,000 and 10,000 using the SGS algorithm. A similar set was run using the PC algorithm for linear and ternary variables. (Each of these algorithms assumes causal sufficiency.) Results were scored separately for errors concerning the existence and the directions of edges, and for correct choice of regressors. Let us call the pattern of the graph in figure 4 the true pattern. Recall that an edge existence error of commission (Co) occurs when any pair of variables are adjacent in the output pattern but not in the true pattern. An edge direction error of commission occurs when in an edge occurring in both the true pattern and the output pattern there is an arrowhead in the output pattern but not the true pattern. Errors of

omission (Om) are defined analogously in each case. The results are tabulated as the average over the trial distributions of the ratio of the number of actual errors to the number of possible errors of each kind. The proportion of trials in which both (Both) actual causes of Y were correctly identified (with no incorrect causes), and in which one (One) but not both causes of Y were correctly identified (again with no incorrect causes) were recorded for each sample size:

Variable Type	#trials	n	%Edge Existence		%Edge Direction		%Both Correct	%One Correct
			Co	Om	Co	Om		
SGS								
Linear	100	2000	1.4	3.6	3.0	5.4	85.7	3.6
Linear	100	10000	1.6	1.0	2.7	2.2	90.0	7.0
Binary	100	2000	0.6	16.6	29.5	21.8	38.0	34.0
Binary	100	10000	1.2	7.4	30.0	9.1	60.0	25.0
PC								
Linear	100	2000	6.0	2.0	1.0	6.2	80.0	15.0
Linear	100	10000	0.0	1.0	2.5	2.9	95.0	0.0
Ternary	100	2000	3.0	1.0	29.1	8.3	65.0	35.0
Ternary	100	10000	3.0	2.0	10.8	1.2	85.0	15.0

The differences in the results with the SGS and PC algorithms for discrete data are due to the choice of binary variables in the former case and ternary variables in the latter case. The tests for statistical independence with discrete variables appear to have more power when variables can have more than two values.

For purposes of prediction and policy, the numbers in the last two columns suggest that the procedure quite reliably finds real causes of the outcome variable when the statistical assumptions of the simulations are met, the sample is large and a causal structure like that in figure 4 obtains. Regression will in these cases find that all of the regressors influence the outcome variable.

8.3 Error Probabilities for Specification Searches

We have shown that various algorithms for specifying causal structure from the data are correct if the requisite statistical decisions are correctly made, but we have given no results about the probability of various sorts of errors in small and medium size samples. The Neyman-Pearson account of testing has made popular two measures of error: the probability of rejecting the null hypothesis when it is true (type I), and the probability of not rejecting the null hypothesis when an alternative is true (type II). Correspondingly, when a search procedure yields a model M from a sample, we can ask for the probability that, were the model M true, the procedure would not find it on samples of that size, and given an alternative M' , we can ask for the probability that were M' true the search procedure would find M on samples of that size. We shall also refer to the error probabilities for the outcomes of search procedures as probabilities of type I and type II errors respectively. Especially in small samples, the significance levels and powers of the tests used in deciding conditional independence may not be reliable indicators of the probabilities of corresponding errors in the search procedure.

Error probabilities for search procedures are nearly impossible to obtain analytically, and we have recommended that Monte Carlo methods be used instead. When a procedure yields M from a sample of size n , estimate M and use the estimated model to generate a number of samples of size n , run the search procedure on each and count the frequency with which something other than M is found. For plausible or interesting alternative models M' , estimate M' , use the estimated model to generate a number of samples of size n , run the search procedure on each and count the frequency with which M is found. We will illustrate the determination of error probabilities for specification searches with a case in which probability of type II error is quite high.

Weisberg, (1985) illustrates a procedure for detecting outliers with an experimental study for which regression produces anomalous results:

An experiment was conducted to investigate the amount of a particular drug present in the liver of a rat. Nineteen rats were randomly selected, weighed, placed under light ether anesthesia and given an oral dose of the drug. Because it was felt that large livers would absorb more of a given dose than smaller liver, the actual dose an animal received was approximately determined as 40 mg of the drug per kilogram of body weight....

The experimental hypothesis was that, for the method of determining the dose, there is no relationship between the percentage of the dose in the liver (Y) and the body weight (X_1), liver weight (X_2), and relative dose (X_3). (p. 121-124)

Regressing Y on (X_1, X_2, X_3) gives a result not in agreement with the hypothesis; the coefficients of Y on body weight (X_1) and dose (X_3) are both significant, even though one is determined by the other. We find the following regression values for Weisberg's data (standard errors are parenthesized and below the coefficients, and t statistics are shown just below the standard errors):

$$\begin{array}{cccc} Y & = -3.902*X_1 + .197*X_2 + 3.995*X_3 + \varepsilon \\ & (1.345) & (.218) & (1.336) \\ & -2.901 & .903 & 2.989 \end{array}$$

A multiple regression not including X_2 also yields significant regressors for X_1 and X_3 at the .05 level. Yet, Weisberg observes that no individual regression of Y on any one of the X variables is significant at that level. The results of the several statistical decisions are therefore inconsistent; we have, for example, that $\rho_{X_1Y} = 0$, $\rho_{X_2Y} = 0$, $\rho_{X_3Y} = 0$ but $\rho_{X_1Y, X_3} \neq 0$. One might take any of several views about such inconsistencies. One is that it is largely an artifact of the particular significance level used. If the .01 level were used to reject hypotheses of vanishing correlations and partial correlations, the correlations with Y would vanish and so would the partial correlations controlling for one other variable. But the partial correlation of X_1 with Y controlling for both of the other regressors could not be rejected, and an inconsistency would remain. Another view is that inconsistencies in the outcomes of sequential statistical decisions are to be expected, especially in small samples, and where possible, inferences should be based on the statistical decisions that are most reliable. In the case at hand the power of any of the statistical tests is low because of the sample size, but the lower the order of the partial correlation the greater the power. The rule of thumb is that to control for an extra variable is to throw away a data point. Thus in this case the PC algorithm never considers the partial correlations and concludes solely from the vanishing correlations that none of the X variables cause the Y variable. Weisberg instead recommends excluding one of the 19 data points, after which a multiple regression using the remaining data gives no (.05) significant regression coefficients.

From the experimental setup we can assume that body weight and liver weight are causally prior to dose, which itself is prior to the outcome, i.e., the amount of the drug found in the rat's

liver. Applying the PC algorithm to the original data set with this background knowledge, and using the .05 significance level in the program's tests, we get the pattern in figure 5.

The PC algorithm gives the supposed correct result in this case because no correlation of an X variable with Y is significant, and that is all the program needs to decide absence of influence. The regression of Y against each individual X variable alone is an essentially equivalent test. To estimate the type I error of the PC search, we obtained a maximum likelihood estimate (assuming normal distributions) of the model shown in figure 5 and used it to generate 100 simulated data sets each of size 19. The PC algorithm was then applied to each data set. In four of the 100 samples the procedure erroneously introduced an edge between Y and one or another of the X variables.

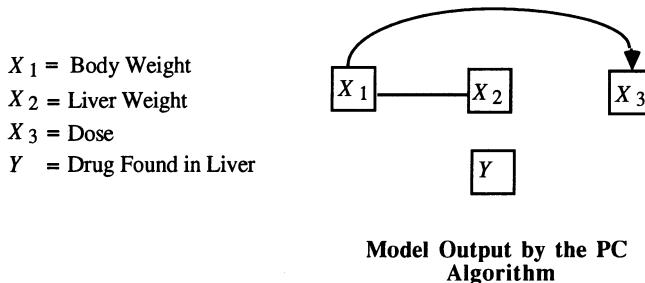


Figure 5

To investigate the power of the procedure against alternatives, we consider three models in which Y is connected to at least one X variable. The first is simply the regression model with correlations among the regressors estimated from the sample. With a correlation of about .99 between X_1 and X_3 , the regression model with correlated errors is nearly unfaithful, and we should expect the search to be liable not to find the structure. We generated 100 samples at each of the sizes 19, 50, 100 and 1000. We then ran the PC algorithm on each sample, counting the output as a type 2 error if it included no edge between Y and some X variable. Figure 6 gives the results for the first three sample sizes.

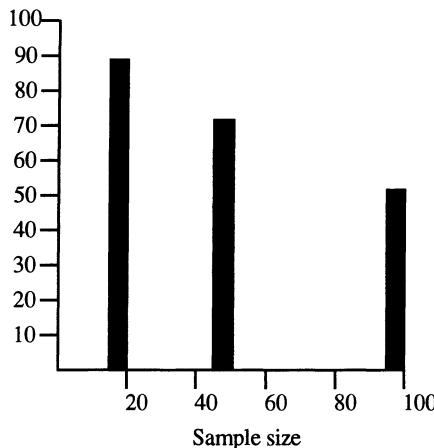


Figure 6: Percentage of Samples from the Regression Model for which PC Omits at Least One Edge

In 100 trials at sample size 1,000 PC never makes a type 2 error against this alternative.

The second alternative is an elaboration of the original PC output. We add an edge from body weight to the outcome, giving the graph in figure 7.

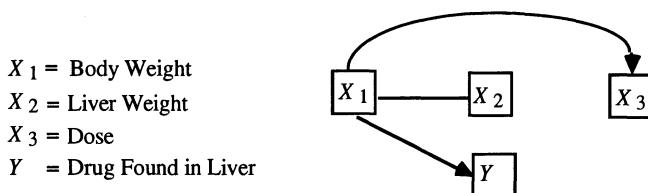


Figure 7

We estimated this model with the EQS program, which found a value of .228 for the linear coefficient associated with the $X_1 \rightarrow Y$ edge, and then used the estimated model to again generate 100 samples at each of the four sample sizes. The results for the first three sample sizes are shown in figure 8.

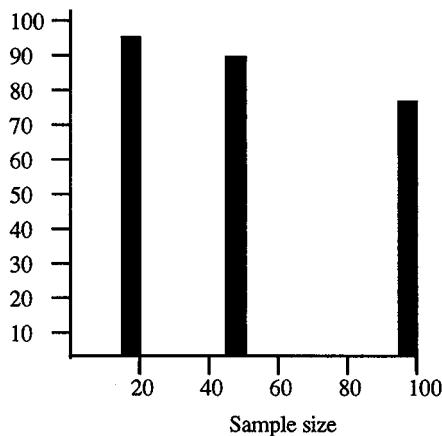


Figure 8: Percentage of samples from the model of figure 7 for which PC omits $X_1 \rightarrow Y$

Even at sample size 1,000 the search makes an error of type 2 against this alternative in 55% of the cases. "Small" influences of body weight on Y cannot be detected. We would expect the same to be true of dose.

In the third case we increased the linear coefficient connecting X_1 and Y in the model in figure 7 to 1.0.

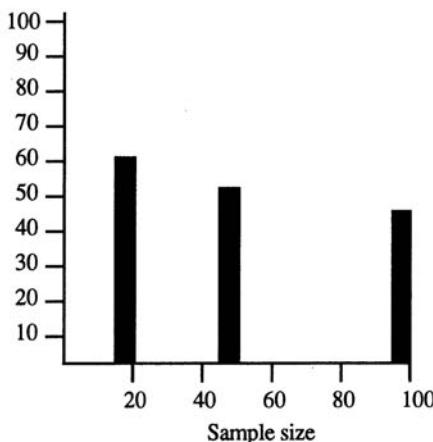


Figure 9: Proportion of samples from the model of figure 7 for which PC omits $X_1 \rightarrow Y$

At sample size 1,000 the search makes an error against this alternative in 2% of the cases.

For this problem at small sample sizes the search has little power against some alternatives, and little power even at large sample sizes against alternatives that may not be implausible.

8.4 Conclusion

In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled studies. So far as we can tell, the popular automatic regression search procedures should not be used at all in contexts where causal inferences are at stake. Such contexts require improved versions of algorithms like those described here to select those variables whose influence on an outcome can be reliably estimated by regression. In applications, the power of the specification searches against reasonable alternative explanations of the data is easy to determine by simulation and ought to be investigated.

It should be noted that the present state of the algorithm is scarcely the last word on selecting direct causes. There are cases in which a partially oriented inducing path graph of a directed acyclic graph G over \mathbf{O} contains a directed edge from X to Y even though X is not a *direct* cause of Y relative to \mathbf{O} (although of course there is a directed path from X to Y in G .) However, Theorem 6.8 states a sufficient condition for a directed edge in a partially oriented inducing path graph to entail that X is a direct cause of Y . In some cases tests based on constraints such as Verma and Pearl's, noted in section 6.9, would help with the problem, but they have not been developed or implemented.

Chapter 9

The Design of Empirical Studies

Simple extensions of the results of the preceding chapters are relevant to the design of empirical studies. In this chapter we consider only a few fundamental issues. They include a comparison of the powers of observational and experimental designs, some implications for sampling and variable selection, and some considerations regarding ethical experimental design. We conclude with a reconsideration from the present perspective of the famous dispute over the causal conclusions that could legitimately be drawn from epidemiological studies of smoking and health.

9.1 Observational or Experimental Study?

There are any number of practical issues about both experimental and non-experimental studies that will not concern us here. Questions of the practical difficulty of obtaining an adequate random sample aside, when can alternative possible causal structures be distinguished without experiment and when only by experiment?

Suppose that one is interested in whether a treatment T causes an outcome O . According to Fisher, R.Fisher (1959) one important advantage of a randomized experiment is that it eliminates from consideration several alternatives to the causal hypothesis to be tested. If the value of T is assigned randomly, then the hypothesis that O causes T or that there is an unmeasured common cause of O and T can be eliminated. Fisher argues that the elimination of this alternative hypothesis greatly simplifies causal inference; the question of whether T causes O is reduced to the question of whether T is statistically dependent on O . (This assumes, of course, instances of the Markov and Faithfulness Conditions.)

Critics of randomized experiments, e.g. Howson and Urbach (1989), have correctly questioned whether randomization in all cases does eliminate this alternative hypothesis. The treatments given to people are typically very complex and change the values of many random variables. For example, suppose one is interested in the question of whether inhaling tobacco smoke from cigarettes causes lung cancer. Imagine a randomized experiment in which one group of people is randomly assigned to a control group (not allowed to smoke) and another group is randomly assigned to a treatment group (forced to smoke 20 cigarettes a day.) Further imagine that the experimenter does not know that an unrecorded feature of the cigarettes, such as the presence of a chemical in some of the paper wrappings of the cigarettes, is the actual cause of lung cancer, and inhaling tobacco smoke does not cause lung cancer. In that case lung cancer and inhaling tobacco smoke from cigarettes are statistically dependent even though inhaling tobacco smoke from cigarettes does not cause lung cancer. They are dependent because assignment to the treatment group is a common cause of inhaling tobacco smoke from cigarettes and of lung cancer.

Fisher (1951, p. 20) suggests that "the random choice of the objects to be treated in different ways would be a complete guarantee of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction." But this does not explain how an experimenter who does not even suspect that cigarette paper might be treated with some cancer causing chemical could know that he had not eliminated all common causes of lung cancer and inhaling tobacco smoke from cigarettes, even though he had randomized assignment to the treatment group. This is an important and difficult question about randomization, made more difficult by the fact that randomization often produces deterministic relationships between such variables as drug dosage and treatment group, producing violations of the Faithfulness Condition.

In this section we will put aside this question, and simply assume that an experimenter has some method that correctly eliminates the possibility that O causes T or that there are common causes of O and T . In general, causal inferences from experiments are based on the principles described in chapters 6 and 7. The theory applies uniformly to inferences from experimental and from non-experimental data. Inferences to causal structure are often more informative when experimental data is available, not because causation is somehow logically tied to experimental manipulations, but because the experimental setup provides relevant background causal knowledge that is not available about non-experimental data. (See Pearl and Verma 1991 for a similar point.)

There are, of course, besides the argument that randomization eliminates some alternative causal hypotheses from consideration, a variety of other arguments that have been given for randomization. It has been argued that it reduces observer bias; that it warrants the experimenter assigning a probability distribution to the outcomes conditional on the null (causal) hypothesis being true, thereby allowing him to perform a statistical test and calculate the probability of type I error; that for discrete random variables it can increase the power of a test by simulating continuity; and that by bypassing 'nuisance factors' it provides a basis for precise confidence levels. We will not address these arguments for randomization here; for a discussion of these arguments from a Bayesian perspective see e.g. Kadane and Seidenfeld (1990).

Consider three alternative causal structures, and let us suppose for the moment that they exhaust the possibilities and are mutually exclusive: (i) A causes C , (ii) some third variable B causes both A and C , or (iii) C causes A . If by experimental manipulation we can produce a known distribution on A not caused by B or C , and if we can produce a known distribution on C not caused by A or B , we can distinguish these causal structures. In the experiment, all of the edges into A in the causal graph of the non-experimental population are broken, and replaced by an edge from U to A ; furthermore there is no non-empty undirected path between U and any other variable in the graph that does not contain the edge from U to A . Any procedure in which A is caused only by a variable U with these properties we will call a **controlled experiment**. In a controlled experiment we know three useful facts about U : U causes A , there is no common cause of U and C , and if U causes C it does so by a mechanism that is blocked if A is held constant (i.e. in the causal graph if there is a directed path from U to C it contains A). As we noted in Chapter 7, U is not a policy variable and is not included in the combined, manipulated or unmanipulated causal graphs.

The controlled experimental setups for the three alternative causal structures are shown in figure 1, where an A -experiment represents a manipulation of A breaking the edges into A , and a C -experiment represents a manipulation of C breaking edges into C . If we do an A -experiment and find partially oriented inducing path graph (ia*) over $\{A,C\}$ then we know that A causes C because we know that we have broken all edges into A . (We do not include U (or V) in the partially oriented inducing path graphs in figure 1 because including them does not strengthen the conclusions that can be drawn in this case, but does complicate the analysis because of the possible deterministic relationships between U and A .) Similarly, if we perform a C -experiment and find partially oriented inducing path graph (iiic*) then we know that C causes A . If we perform an A -experiment and get (iia*) and a C -experiment and get (iic*) then we know that there is a latent common cause of A and C (assuming that A and C are dependent in the non-experimental population.)

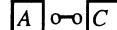
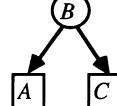
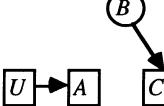
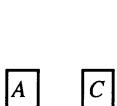
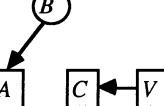
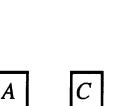
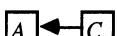
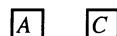
Model	A-Experiment	Partially Oriented Inducing Path Graph	C-Experiment	Partially Oriented Inducing Path Graph
 (i)	 (ia)	 (ia*)	 (ic)	 (ic*)
 (ii)	 (iia)	 (iia*)	 (iic)	 (iic*)
 (iii)	 (iiia)	 (iiia*)	 (iiic)	 (iiic*)

Figure 1

Now suppose that in the non-experimental population there are variables U and V known to bear the same relations to A and C respectively as in the experimental setup. (We assume in the non-experimental population that A is not a deterministic function of U , and C is not a deterministic function of V .) That is, U causes A , there is no common cause of U and C , and if there is any directed path from U to C it contains A ; also, V causes C , there is no common cause of V and A , and if there is any directed path from V to A it contains C . Can we still distinguish (i), (ii), and (iii) from each other without an experiment? The answer is yes. In figure 2, (io*), (iio*) and (iiio*) are the partially oriented inducing path graphs corresponding to (i), (ii), and (iii) respectively. Suppose the FCI algorithm constructs (io*). If it is known that U causes A , then from the fact that the edge between U and A and the edge between A and C do not collide, we can conclude that the edge between A and C is oriented as $A \rightarrow C$ in the inducing path graph. It follows that A causes C . Similarly, if the FCI algorithm constructs (iiio*) ideally we can conclude that C causes A . The partially oriented inducing path graph in (iio*) indicates by Theorem 6.9 that there is a latent common cause of A and C , and by Theorem 6.6 that A does not cause C , and C does not cause A .

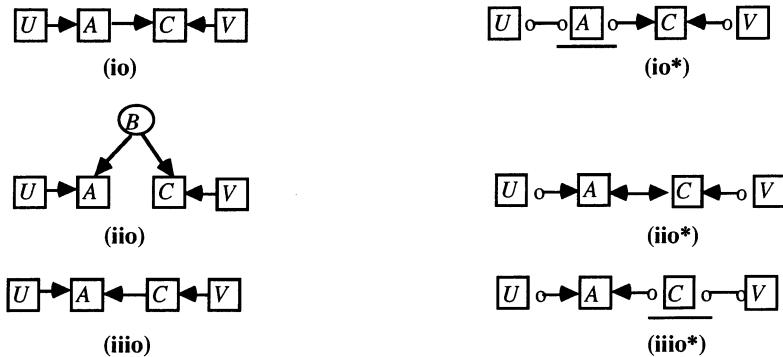


Figure 2

Note that if we had measured variables such as W , U , V , and X in figure 3 then the corresponding partially oriented inducing path graphs would enable us to distinguish (i), (ii), and (iii) without experimentation and without the use of any prior knowledge about the causal relations among the variables.

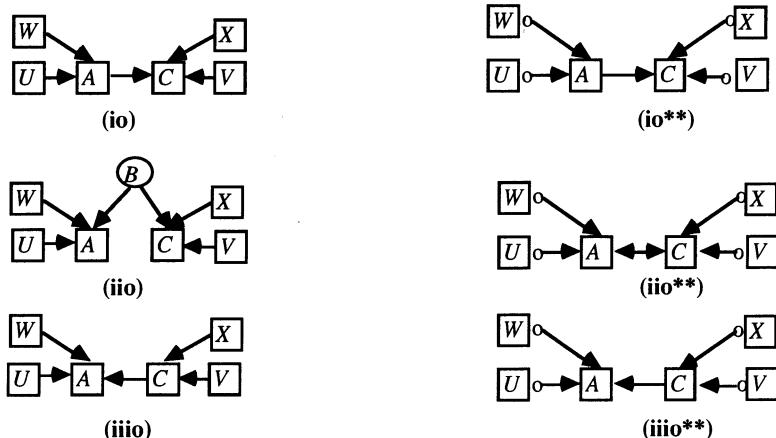


Figure 3

Consider now the more complex cases in which the possibilities are (i) A causes C and there is a latent common cause B of A and C , (ii) there is a latent common cause B of A and C , and (iii) C causes A and there is a latent common cause B of A and C . Each of the structures (i), (ii) and

(iii) can be distinguished from the others by experimental manipulations in which for a sample of systems we break the edges into A and impose a distribution on A and for another sample we break the edges into C and impose a distribution on C . The corresponding graphs are presented in figure 4, and the analysis of the experiment is essentially the same as in the previous case.

Model	A-Experiment	Partially Oriented Inducing Path Graph	C-Experiment	Partially Oriented Inducing Path Graph
 (i)	 (ia)	 (ia*)	 (ic)	 (ic*)
 (ii)	 (iia)	 (iia*)	 (iic)	 (iic*)
 (iii)	 (iiia)	 (iiia*)	 (iiic)	 (iiic*)

Figure 4

The analysis of the corresponding non-experimental case is more complicated. Assume that there is a variable U and it is known that U causes A , there is no common cause of U and A , and if there is any directed path from U to C it contains A , and that there is a variable V and it is known that V causes C , there is no common cause of V and A , and if there is any directed path from V to A it contains C . The directed acyclic graphs and their corresponding partially oriented inducing path graphs are shown in figure 5. Now suppose that the directed acyclic graphs are true of an observed non-experimental population. Can we still distinguish (i), (ii), and (iii) from each other?

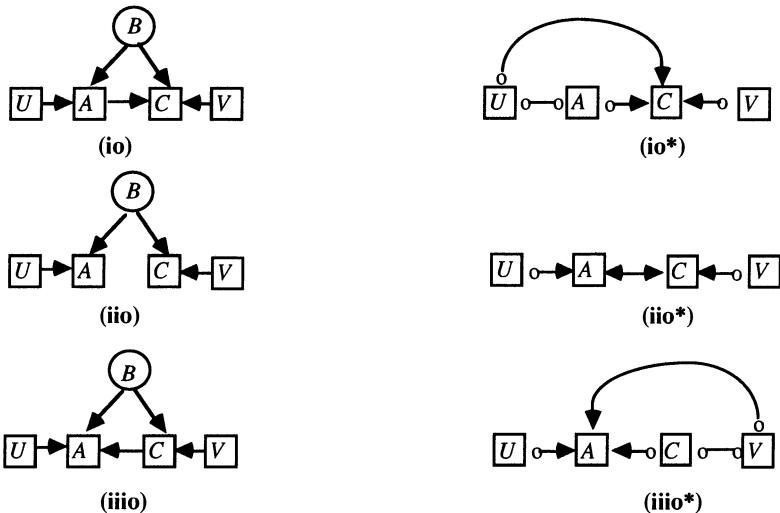


Figure 5

Once again the answer is yes. For example, suppose that an application of the FCI algorithm produces (io*). The existence of the $U \circ \rightarrow C$ edge entails that either there is a common cause of U and C or a directed path from U to C . By assumption, there is no common cause of U and C , so there is a directed path from U to C . Also by assumption, all directed paths from U to C contain A , so there is a directed path from A to C . Given that there is an edge between U and C in the partially oriented inducing path graph, and the same background knowledge, it also follows that there is a latent common cause of A and C . (The proof is somewhat complex and we have placed it in an Appendix to this chapter.) Similarly, if we obtain (iiio*) then we know that C causes A and there is a latent common cause of A and C . If we obtain (iio*) then we know that A and C have a latent common cause but that A does not cause C and C does not cause A . It is also possible to distinguish (i), (ii), and (iii) from each other without any prior knowledge of particular causal relations, but it requires a more complex pattern of measured variables, as shown in figure 6. If we obtain (io**) then we know without using any such prior knowledge about the causal relationships between the variables that A causes C and that there is a latent common cause of A and C , and similarly for (iio**) and (iiio**).

There is an important advantage to experimentation over passive observation in one of these cases. By performing an experiment we can make a quantitative prediction about the consequences of manipulating A in (i), (ii), and (iii). But if (i) is the correct causal model, we

cannot use the Prediction Algorithm to make a quantitative prediction of the effects of manipulating A . (In the linear case, a prediction could be made because U serves as an "instrumental variable.")

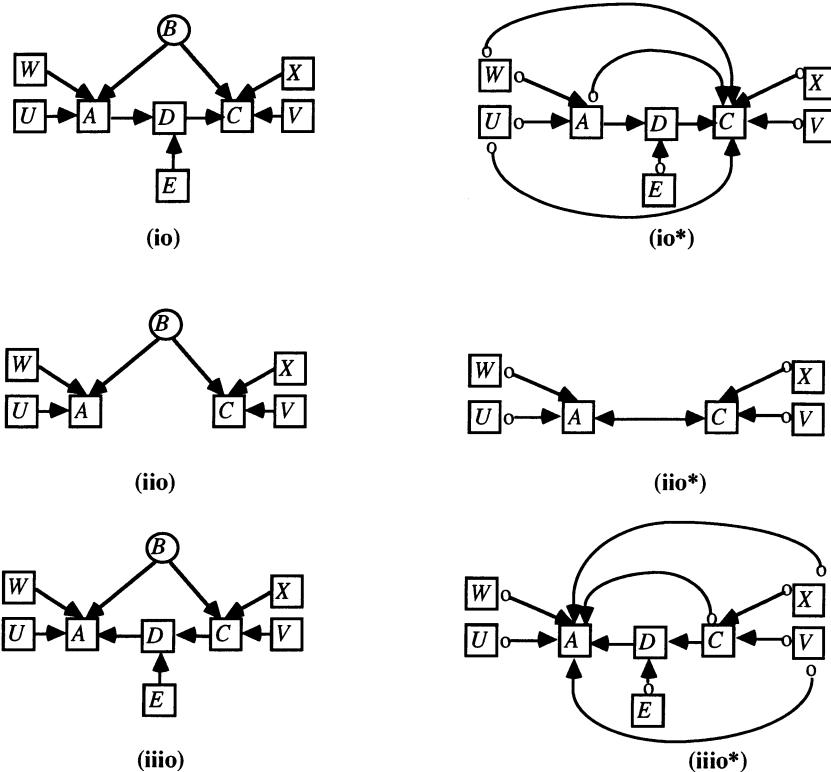


Figure 6

Suppose finally that we want to know whether there are two causal pathways that lead from A to C . More specifically, suppose we want to distinguish which of (i), (ii) and (iii) in figure 7 obtains, remembering again that B is unmeasured.

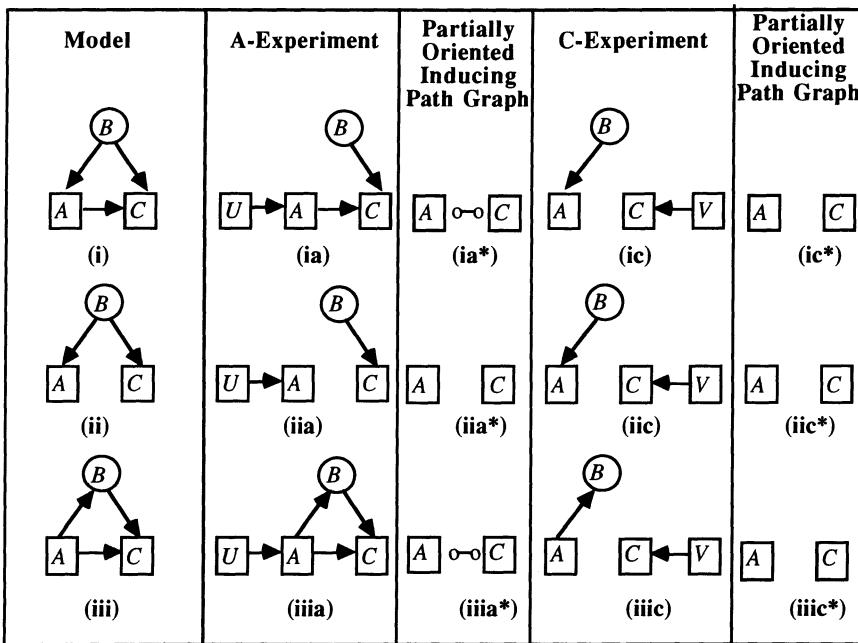


Figure 7

The question is fairly close to Blyth's version of Simpson's paradox. By experimental manipulation that breaks the edges directing into A and imposes a distribution on A , we can distinguish structures (i) and (iii) from structure (ii) but not from one another. Note that in figure 7 the partially oriented inducing path graph (ia*) is identical to (iiia*) and (ic*) is identical to (iiic*).

Assume once again that in a non-experimental population it is known that U causes A , there is no common cause of U and C , and if there is any path from U to C it contains A , and V causes C , there is no common cause of V and A , and if there is any path from V to A it contains C . The directed acyclic graphs and their corresponding partially oriented inducing path graphs are shown in figure 8.

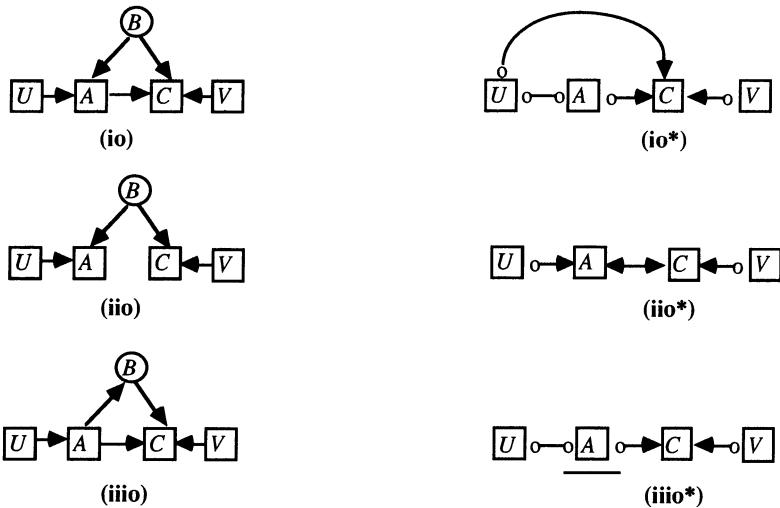


Figure 8

Unlike the controlled experimental case, where (i) and (iii) cannot be distinguished, in the non-experimental case they *can* be distinguished. Suppose we obtain (iiiio*). We know from the background knowledge that U causes A , and from (iiio*) that the edge between U and A does not collide with the edge between A and C . Hence in the corresponding inducing path graph there is an edge from A to C and in the corresponding directed acyclic graph there is a path from A to C . (Of course we cannot tell how many paths from A to C there are; (iiio*) is compatible with a graph like (iio) but in which the $\langle A, B, C \rangle$ path does not exist.) We also know that there is no latent common cause of A and C because (iiio*) together with our background knowledge entails that there is no path in the inducing path graph between A and C that is into A . Suppose on the other hand that we obtain (io*). Recall that the background knowledge together with the partially oriented inducing path graph entail that A is a cause of C and that there is a latent common cause of A and C . (We have placed the proof in an Appendix to this chapter.)

Once again if more variables are measured, it is also possible to distinguish these three cases without any background knowledge about the causal relationships among the variables, as shown in figure 9.

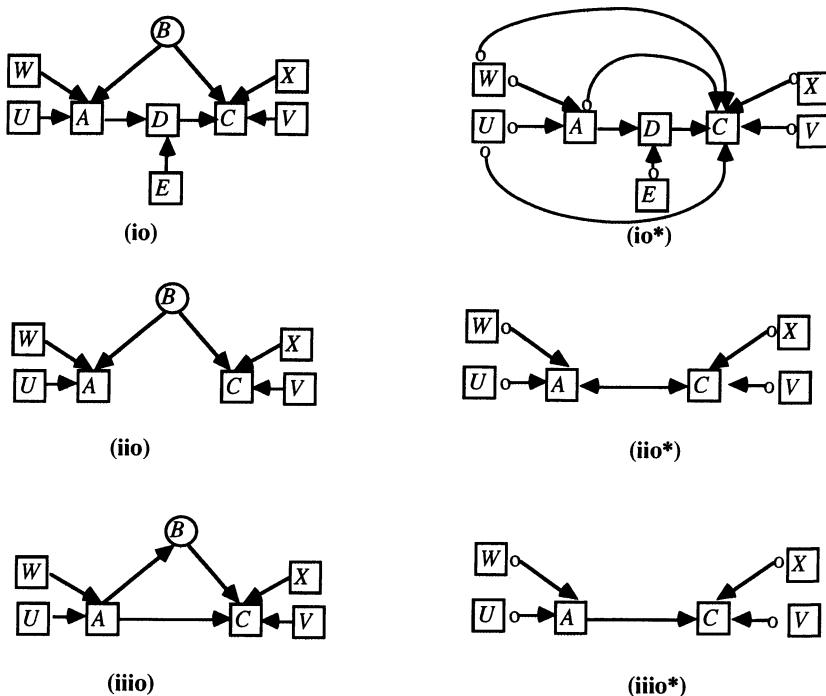


Figure 9

Thus all three structures can be distinguished without experimental manipulation or prior knowledge.

It may seem extraordinary to claim that structure (i) in figure 7 cannot be distinguished from structure (iii) by a controlled experiment, but can be distinguished without experimental control if the structure is appropriately embedded in a larger structure whose variables are measured. It runs against common sense to claim that when A causes C , a controlled experiment cannot distinguish A and C also having an unmeasured common cause from A also having a second mechanism through which it effects B , but that observation without experiment sometimes can distinguish these situations. But controlled experimental manipulation that forces a distribution on A breaks the dependency (in the experimental sample) of A on B in structure (i), and thus information that is essential to distinguish the two structures is lost.

While a controlled experiment alone cannot distinguish (i) from (iii) in figure 7 the combination of a simple observational study and controlled experimentation can distinguish (i) from (iii). We can determine from an A -experiment that there is a path from A to C , and hence no path from C to A . We know if $P(C|A)$ is not invariant under manipulation of A then there is a trek between C and A that is into A . Hence if $P(C|A)$ is different in the non-experimental population and the A -experimental population we can conclude that there is a common cause of A and C . If $P(C|A)$ is invariant under manipulation of A then we know that either there is no common cause of A and C or the particular parameter values of the model "coincidentally" produce the invariance. By combining information from an observational study and an experimental study it is sometimes possible to infer causal relations that cannot be inferred from either alone. This is often done in an informal way. For example, suppose that in both an A -experiment and a C -experiment A and C are independent. This indicates that there is no directed path from A to C or C to A . But it does not distinguish between the case where there is no common cause of A and C (i.e. there is no trek at all between A and C) and the case where there is a common cause of A and C . Of course in practice these two models are distinguished by determining whether A and C are independent in the non-experimental population; assuming faithfulness, there is a trek between A and C if and only if A is not independent of C .

In view of these facts the advantages of experimental procedures in identifying (as distinct from measuring) causal relations need to be recast. There are, of course, well known practical difficulties in obtaining adequate non-experimental random samples without missing values but we are interested in issues of principle. One disadvantage of non-experimental studies is that in order to make the distinctions in structure just illustrated either one has to know something in advance about some of the causal relations of some of the measured variables, or else one must be lucky in actually measuring variables that stand in the right causal relations. The chief advantage of experimentation is that we sometimes know how to *create* the appropriate causal relations. A further advantage to experimental studies is in identifying causal structures in mixed samples. In the experimental population the causal relation between a manipulating variable and a manipulated variable is known to be common to every system so treated. Mixing different causal structures acts like the introduction of a latent variable, which makes inferences about other causal relations from a partially oriented inducing path graph more difficult. Similar conclusions apply to cases in which experimental and statistical controls are combined.

In the "controlled" experiments we have discussed thus far, we have assumed that the experimental manipulation breaks all of the edges into A in the causal graph of the non-experimental population, and that the variable U used to manipulate the value of A has no common cause with C . However, it is possible to do informative experiments that satisfy

neither of these assumptions. Suppose, for example, the causal graph of figure 10 describes a non-experimental population.

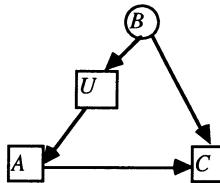


Figure 10

Suppose that in an experiment in which we manipulate A , we force a distribution upon $P(A|U)$. In this case the causal graph of the experimental population is the same as the causal graph of the non-experimental population, although of course the parametrization of the graph is different in the two populations. This kind of experiment does not break the edges into A . More generally, we assume that there is a set of variables U used to influence the value of A such that any direct cause V of A that is not in U is connected to some outcome variable C only by undirected paths that contain some member of U as a non-collider. (This may occur for example, if U is a proper subset of the variables used to fix the value of A , and the other variables used to fix the value of A are directly connected *only* to A .) These are just the conditions that we need in order to guarantee the invariance of the distribution of a variable C given U and A , and hence allows the use of the Prediction Algorithm. (A more extensive discussion of this kind of experiment is given in section 9.4.) With an experiment of this kind, it is possible to distinguish model (i) from model (iii) in figure 7. Of course, with the same background knowledge assumptions it is also possible to distinguish (i) from (iii) in a non-experimental study in which the distribution of $P(A|U)$ is not changed. Indeed with this kind of experiment, the only difference between the analysis of the experimental population and a non-experimental population lies in the background knowledge employed.

9.2 Selecting Variables

The selection of variables is the part of inference that at present depends almost entirely on human judgment. We have seen that poor variable selection will usually not of itself lead to incorrect causal conclusions, but can very well result in a loss of information. Discretizing

continuous variables and using continuous approximations for discrete variables both risk altering the results of statistical decisions about conditional independence.

One fundamental new consideration in the selection of variables is that in the absence of prior knowledge of the causal structure, empirical studies that aim to measure the influence, if any, of A on C or of C on A , should try to measure at least two variables correlated with A that are thought to influence C , if at all, only through A , and likewise for C . As the previous section illustrates, variables with these properties are especially informative about whether A causes C , or C causes A , or neither causes the other but there is a common cause of A and C .

The strategy of measuring every variable that might be a common cause of A and B and conditioning on all such variables is hazardous. If one of the additional variables is in fact an effect of B , or shares a common cause with A and a common cause with B , conditioning on that variable will produce a spurious dependency between A and B . That is not to say that extra variables should not be measured if it is thought that they may be common causes; but if they are measured, they should be analyzed by the methods of chapters 5 and 6 rather than by multiple regression.

Finally, if methods like those described in chapters 5, 6 and 7 are to be employed, we offer the obvious but unusual suggestion that variables be selected for which good conditional independence tests are available. At present our simulations suggest that is a reason to avoid binary variables where possible—for example in psychometric and sociometric test design.

9.3 Sampling

We can view many sampling designs as procedures that specify a property S , which may have two values or several, and from subpopulations with particular S values draw a sample in which the distribution of values of the i^{th} unit drawn is distributed independently of and identically to the distribution of all other sample places from that subpopulation. In the simplest case S can be viewed as a binary variable with the value 1 indicating that a unit has the sample property, which of course does not mean that the unit occurs in any particular actual sample. We distinguish the sample property S from any treatments that might be applied to members of the sample. In sampling according to a property S we obtain information directly not about the general population but about the segments of the population that have various values of S . Our

general questions therefore concern when conditioning on any value of S in the population leaves unaltered the conditional probabilities or conditional independence relations for variables in the causal graph G describing the causal structure of each unit in the population. That is, suppose there is a population in which the causal structure of all units is described by a directed graph G , and let the values of the variables be distributed as P , where P is faithful to G . What are the causal and statistical constraints a sampling property S must satisfy in order that a sub-population consisting of all units with a given value of S will accurately reflect the conditional independence relations in P --and thus the causal structure G --and under what conditions will the conditional probabilities for such sub-populations be as in P ? The answers to these questions bear on a number of familiar questions about sampling, including the appropriateness of retrospective versus prospective sampling and of random sampling as against other sampling arrangements. We will not consider questions about the sampling distributions obtained by imposing various constraints on the distribution of values of S in a sample. Our discussion assumes that S (which may be identical to one of the variables in G) is not determined by any subset of the other variables in G .

We assume in our discussion that S is defined in such a way that if the sampling procedure necessarily excludes any part of the population from occurring in a sample, then the excluded units have the same S value. For example, if a sample is to be drawn from the sub-population of people over six feet tall, then we will assume that $S = 0$ corresponds to people six feet tall or under and $S = 1$ corresponds to people over 6 feet tall.

The causal graph G relating the variables of interest can be expanded to a graph $G(S)$ that includes S and whatever causal relations S and the other variables realize. We assume a distribution $P(S)$ faithful to $G(S)$ whose marginal distribution summing over S values will of course be P . We suppose that the sampling distribution is determined by the conditional distribution $P(\cdot | S)$. Our questions are then, more precisely, when this conditional distribution has the same conditional probabilities and conditional independence relations as P . We require, moreover, that the answer be given in terms of the properties of the graph $G(S)$. The following theorem is obvious and will not be proved.

Theorem 9.1 If $P(S)$ is faithful to $G(S)$, and \mathbf{X} and \mathbf{Y} are sets of variables in $G(S)$ not containing S , then $P(\mathbf{Y}|\mathbf{X}) = P(\mathbf{Y}|\mathbf{X}, S)$ if and only if \mathbf{X} d-separates \mathbf{Y} and S in $G(S)$.

Our sampling property should not be the direct or indirect cause or effect of \mathbf{Y} save through a mechanism blocked by \mathbf{X} , and \mathbf{X} should not be the effect, direct or indirect of both \mathbf{Y} and the sampling property. (The second clause in effect guarantees that Simpson's paradox is avoided

in a faithful distribution). The theorem is essentially the observation that $P(\mathbf{Y}|\mathbf{X} \cup \mathbf{Z}) = P(\mathbf{Y}|\mathbf{X} \cup \mathbf{Z} \cup \{S\})$ if and only if in P \mathbf{Y} and S are independent conditional on $\mathbf{X} \cup \mathbf{Z}$. It entails, for example, that if we wish to estimate the conditional probability of \mathbf{Y} on \mathbf{X} from a sample of units with an S property (say, $S = 1$), we should try to ensure that there is

- (i) no direct edge between any Y in \mathbf{Y} and S ,
- (ii) no trek between any Y in \mathbf{Y} and S that does not contain some X in \mathbf{X} , and
- (iii) no pair of directed paths from any Y in \mathbf{Y} to an X in \mathbf{X} and from S to X .

Figure 11 illustrates some of the ways estimation from the sampling property can bias estimates of the conditional probability of \mathbf{Y} given \mathbf{X} and \mathbf{Z} .

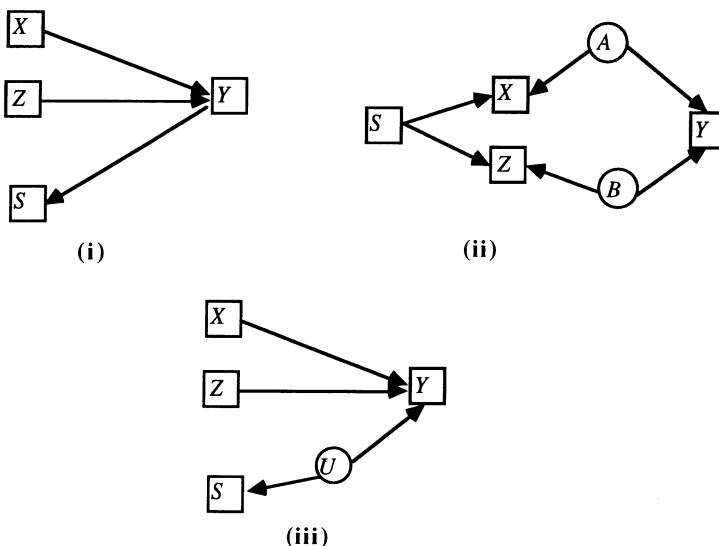


Figure 11

Cases (i) and (iii) are typical of retrospective designs. In case (ii) the sampling property biases estimates of $P(Y|X, Z)$ because Y and the sample property S are dependent conditional on $\{X, Z\}$. Theorem 9.1 amounts to a (very) partial justification of the notion that: "prospective" sampling is more reliable than "retrospective" sampling, if by the former is meant a procedure that selects by a property that causes or is caused by Y , the effect, if at all only through X , the cause, and by the latter is meant a procedure that selects by a property that causes or is caused by X only through Y . In a prospective sampling design in which X is the only direct cause of

S , and S does not cause any variable, the estimate of $P(Y|X,Z)$ is not biased. But case (ii) shows that under some conditions prospective samples can bias estimates as well.

Similar conclusions should be drawn about random sampling. Suppose as before that the goal is to estimate the conditional probability $P(Y|X)$ in distribution P . In drawing a random sample of units from P we attempt to sample according to a property S that is entirely disconnected with the variables of interest in the system. If we succeed in doing that then we ensure that S has no causal connections that can bias the estimate of the conditional probability. Of course even a random sample may fail if the very property of being selected for a study (a property, it should be noted, different from having some particular value of S) affects the outcome, which is part of the reason for blinding treatments. Further, *any* property that has the same causal disconnection will do as a basis for sampling; there is nothing special in this respect about randomization, except that a random S is believed to be causally disconnected from other variables.

When the aim is only to determine the causal structure, and not to estimate the distribution P or the conditional probabilities in P , the asymmetry between prospective and retrospective sampling vanishes.

In models (i) and (iii) of figure 11, which are examples of retrospective design, for any three disjoint sets of variables A , B , and C not containing S , A is d-separated from B given C if and only if A is d-separated from B given $C \cup S$. So these cases in which conditional probability in P cannot be determined from S samples are nonetheless cases in which conditional independence in P , and hence causal structure, can be determined from S samples.

Theorem 9.2 states conditions under which the set of conditional independence relations true in the population as a whole is different from the set of conditional independence relations true in a subpopulation with a constant value of S . In Theorem 9.2 let Z be any set of variables in G not including X and Y .

Theorem 9.2: For a joint distribution P , faithful to graph G , exactly one of $\langle Y \perp\!\!\!\perp X|Z \rangle$; $\langle Y \perp\!\!\!\perp X|Z \cup \{S\} \rangle$ is true in P if and only if the corresponding member and only that member of $\langle Z \text{ d-separates } X; Y; Z \cup \{S\} \text{ d-separates } X, Y \rangle$ is true in G .

Although Theorem 9.2 is no more than a restatement of Theorem 3.3, its consequences are rather intricate. Suppose that X and Y are independent conditional on Z in distribution P . When will sample property S make it appear that X and Y are instead dependent conditional on Z ? The answer is exactly when X, Y are dependent conditional on $Z \cup S$ in $P(S)$. This

circumstance--conditional independence in P and conditional dependence in $P(S)$ --can occur for faithful distributions when and only when there exists an undirected path U from X in \mathbf{X} to Y in \mathbf{Y} such that

- i. no noncollider on U is in $\mathbf{Z} \cup \{S\}$;
- ii. every collider on U has a descendant in $\mathbf{Z} \cup \{S\}$;
- iii. some collider on U does not have a descendant in \mathbf{Z} .

The converse error involves conditional dependence in P and conditional independence in $P(S)$. That can happen in a faithful distribution when and only when there exists an undirected path U from X to Y such that

- i. every collider on U has a descendant in \mathbf{Z} ;
- ii. no noncollider in U is in \mathbf{Z} ;

and S is a noncollider on every such path. Again, asymptotically both of these errors can be avoided by sampling randomly, or by any property S that is unconnected with the variables of interest.

In experimental designs the aim is sometimes to sample from an ambient population, apply a spectrum of treatments to the sampled units, and then infer from the outcome the effect a policy of treatment would have if applied to the general population. In the next section we consider some relations between experimental design, policy prediction, and causal reasoning.

9.4 Ethical Issues in Experimental Design

Clinical trials of alternative therapies have at least two ethical problems. (1) In the course of the trials (or even beforehand) suspicion may grow to near certainty that some treatments are better than others; is it ethical to assign people to treatments, or to continue treatments, found to be less efficacious? (2) In clinical trials, whether randomized or other, patients are generally *assigned* to treatment categories; if the patients were not part of an experimental design, presumably they would be free to choose their treatment (free, that is, if they could pay for it, or persuade their insurer to); is it ethical to ask or induce patients to forego choosing? Suppose the answer one gives to each of these questions is negative. Are there experimental designs for clinical trials that avoid or mitigate the ethical problems but still permit reasonable predictions of the effects of treatment throughout the population from which the experimental subjects are obtained?

Kadane and Sedransk (1980) describes a design (jointly proposed by Kadane, Sedransk, and Seidenfeld) to meet the first problem. Their design has been used in trials of drugs for post-operative heart patients and in other applications. The inferences Kadane and Seidenfeld (1990) make in explaining the design are in accord with the Markov Condition, and indeed follow from it, and the case nicely illustrates the role of causal reasoning in experimental design. Furthermore, combining the Markov and Faithfulness Conditions, and using the Manipulation Theorem and Theorem 7.1 leads to two novel conclusions:

1. The efficacy of treatments in an experiment can be reliably assessed in an experimental procedure that takes patient preference into account in allocating treatment, but except in special cases the knowledge so acquired could not be used to predict the effects of a general policy of treatment.
2. Perhaps of more practical interest, given comparatively weak causal knowledge on the part of the experts, another design in which treatment allocation depends on patient preference can be used to determine whether or not patient self-assignment in the experiment will confound prediction of the outcome of a treatment policy in the general population. When all influences are linear, the effects of treatment policy can be predicted even if confounding occurs.

9.4.1 The Kadane/Sedransk/Seidenfeld Design

In the Kadane/Sedransk/Seidenfeld experimental design (described in Kadane and Seidenfeld (1990)), for each member of a panel of experts, degrees of belief are elicited about the outcome O of each treatment T conditional on each profile of values of $X_1 \dots X_n$. The elicited judgments are used to specify some prior distribution over parameters in a model of the treatment process. For each experimental subject, the panel of experts receives information on the variables X_1, \dots, X_n . Nothing else about the patient is known to the experts. Based on the values of X_1, \dots, X_n each expert i recommends a preferred treatment $p_i(X)$ to the patient, and the patient is assigned to treatment by some rule $T = h(X, p_1, \dots, p_k)$ that is a function of the X values and the experts' treatment preferences (p_1, \dots, p_k) for patients described by X , and perhaps some random factor. The rule guarantees that no patient is given a treatment unless at least one expert recommends it for patients with that profile. The model determines the likelihood, for each vector of parameter values, of outcomes conditional on X and T values. As data are collected on patients, the prior distribution over the parameters is updated by conditioning. If the evidence reaches a stage at which all the experts agree that some treatment T for patients with profile X is

not the best treatment for the patient, then treatment T is suspended for such patients. As evidence accrues, the experts' degrees of belief about the parameter values of the likelihood model should converge.

Let X_j be a vector of observed characteristics of the j^{th} patient, "including those that are used as a basis for deciding what treatment each patient is to receive, and possibly other characteristics as well." (We do not place X_j in boldface in order to match Kadane and Seidenfeld's notation.) Let T_j be the treatment assigned to patient j . Let O_j be the outcome for patient j . Let $P_j = (O_j, T_j, X_j, O_{j-1}, T_{j-1}, X_{j-1}, \dots, X_1)$ be the past evidence up to and including what is known about patient j . Let θ be a vector of the parameters of interest, those that determine the probabilities of outcomes O_j for a patient j given characteristics X_j and treatment T_j . For example, the degrees of belief of an expert might be represented by a mixture of linear models parametrized by exogenous variances, means and linear coefficients. A unique value of these parameters then "determines" the probability of an outcome given X values. For reasons that will become clear, it is essential to the definition of θ that alternative values for the parameter not give alternative specifications of the distribution of X variables.

The expression $f_\theta(P_j)$ represents the expert's conditional degree of belief, given θ , that the total evidence is P_j . Kadane and Seidenfeld add that "It is part of the definition of θ as the parameter that

$$f_\theta(O_j|T_j, X_j, P_{j-1}) = f_\theta(O_j|T_j, X_j) \quad (1 \leq j \leq J) \quad (1)$$

What this means is that θ contains all the information contained in P_{j-1} that might be useful for predicting O_j from T_j and X_j . The factorization of degree of belief,

$$f_\theta(P_J) = \left[\prod_{j=1}^J f_\theta(O_j|T_j, X_j, P_{j-1}) \right] \left[\prod_{j=1}^J f_\theta(T_j|X_j, P_{j-1}) \right] \left[\prod_{j=1}^J f_\theta(X_j|P_{j-1}) \right]$$

1 2 3

follows by the definition of conditional probability. The terms are marked 1, 2, and 3. Kadane and Seidenfeld claim that term 3 does not depend on θ if one believes that the features, treatments and outcomes for earlier subjects in the experimental trial have no influence on "the kinds of people" who subsequently become subjects. (Recall that parameters relevant to the distribution of $X_1 \dots X_n$ are not included in θ .) Kadane and Seidenfeld also say that term 2 does not depend on θ because there is a fixed rule for treatment assignment as a function of X values and the history of the experimental outcomes.

It follows from (1) that

$$\prod_{j=1}^J f_{\theta}(O_j|T_j, X_j, P_{j-1}) = \prod_{j=1}^J f_{\theta}(O_j|T_j, X_j)$$

Kadane and Seidenfeld say the proportionality given by this term:

$$f_{\theta}(P_J) \propto \prod_{j=1}^J f_{\theta}(O_j|T_j, X_j)$$

"is the form that we use to evaluate the results of a clinical trial of the kind considered here." That is for each value θ_i of θ , multiplying $f_{\theta_i}(P_J)$ by the prior density of θ_i gives a quantity proportional to the posterior density of θ_i . The ratios of the posterior densities of two values of θ_i can therefore be found.

Now for a new case, each value of θ determines a probability of treatment outcome given an X profile and a treatment T , and so the posterior distribution of θ yields, for any one expert, degrees of belief in the outcomes of various treatment regimes to various classes of patients. Although Kadane and Seidenfeld say nothing explicit about predicting the effects of policies of treatment, these degrees of belief may be transformed into expected values if outcome is somehow quantified. In any case, an expert who began believing that a rule of treatment given by $T = k(X)$ would most often result in a successful outcome, may come instead to predict that a different rule of treatment, say $T = g(X)$ will more often be successful.

Why can't the experiment let the subjects simply choose their own treatments, and seek any advice that they want? Kadane and Seidenfeld give two reasons. One is that if patients were to determine their own treatment, the argument that term 2 in the factorization does not depend on θ would no longer hold. The other is that "It would now be necessary to explain statistically the behavior of patients in choosing their treatments, and there might well be contamination between these choices and the effect of the treatment itself." We will consider the force of these considerations in the next subsection.

9.4.2 Causal Reasoning in the Experimental Design

What is it that the experts believe that warrants this analysis of the experiment, or the derivation of any predictions? The expert surely entertains the possibility that some unknown common causes U may influence both the X features of a patient and the outcome of the patient's treatment. And yet the analysis assumes that in the expert's degrees of belief, treatment is independent of any such U conditional on X values. That is implicit in the claim that term 2 in the factorization is independent of θ . *Why should treatment T and unknown causes U be conditionally independent given X ?* The reason, clearly, is that in the experiment the only factors that influence the treatment a patient receives are the X values for that patient and P_{j-1} ; any such U , should it exist, has no influence on T except through X . A causal fact is the basis for independence of probabilities.

Aspects of the expert's understanding of the experimental set-up are pictured in figure 12 (where we have made X_j a single variable.)

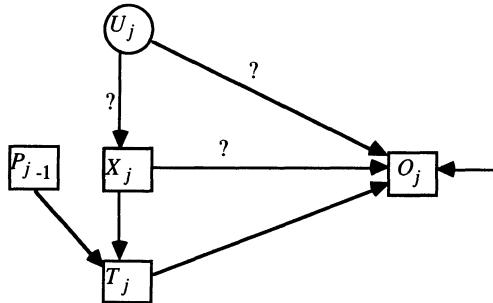


Figure 12

The expert may not be at all sure that the edges with "?" correspond to real influences, but she is sure there is *no* influence of the kind in figure 13 in boldface.

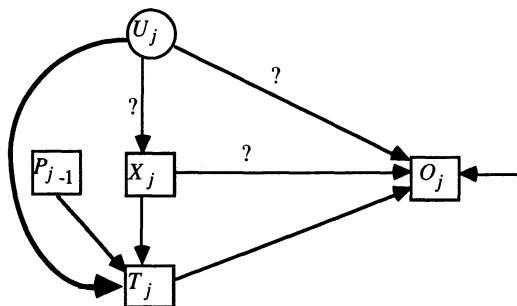


Figure 13

The experimental design, which makes treatment assignment a function of the X variables and P_{j-1} only, is contrived to exclude such influences. The expert's thought seems to be that if U influences T only through X , then U and T are independent conditional on X . That thought is an instance of the Markov Condition. The probabilities in the Markov Condition can be understood either objectively or subjectively. But in the Kadane/Sedransk/Seidenfeld design the probability in the Markov Condition cannot be the expert's unconditional degrees of belief, because those probabilities are mixtures over θ of distributions conditional on θ , and mixtures of distributions satisfying the Markov Condition do not always satisfy the Markov Condition. We will assume the distributions conditional on θ do so.

Consider another feature of the idealized expert belief. An idealized expert in Kadane and Seidenfeld's experiment changes his probability distribution for a parameter whose values specify a model of the experimental process. At the end of the experiment the expert has a view not only about the outcome to be expected for a new patient with profile X if that patient were assigned treatment according to the rule $T = h(X, P_{j-1})$ used in the experiment, but also about the outcome to be expected for a new patient with profile X if that patient were assigned treatment according to the rule $T = g(X)$ that the expert now, in light of the evidence, prefers. In principle, the expert's probabilities for outcomes if the new patient were treated by the experimental rule $T = h(X, P_{j-1})$ is easy to compute because that probability is determinate for each value of θ , and we know the expert's posterior distribution of θ . *But what determines the expert's probability for outcomes if the patient with profile X is now treated according the preferred rule, $T = g(X)$?* Why doesn't changing the rule change the dependence of O on X and T ? The sensible answer implicit in Kadane and Seidenfeld's analysis, is that the outcome for any patient depends on the X profile of the patient and the treatment given to the patient, but not on the "rule" by which treatments are assigned. Changing the assignment rule changes the

probability of treatment T given profile X , but has no effect on other relevant conditional probabilities; the probability O given T and X is unaltered. We can derive this more formally in the following way.

If for a fixed value of θ the distribution $f_{\theta}(O_j, T_j, X_j, P_{j-1})$ satisfies the Markov condition for graphs of the type in figure 12, then Theorem 7.1 entails that $f_{\theta}(O_j|T_j, X_j)$ is invariant under a manipulation of T_j . According to Theorem 7.1, in a distribution that satisfies the Markov condition for graphs of the type in figure 12, $f_{\theta}(O_j|T_j, X_j)$ is invariant under a manipulation of T_j if there is no path that d-connects O_j and X_j given T_j that is into T_j . Every undirected path between T_j and O_j that contains some X_j variable satisfies this condition because some member of X_j is a non-collider on such a path. There are no undirected paths between T_j and O_j that contain P_{j-1} . Hence $f_{\theta}(O_j|T_j, X_j)$ is invariant under manipulation of T_j .

But does the Markov condition reasonably apply in an experiment designed according to the Kadane/Sedransk/Seidenfeld specifications in which $f_{\theta}(O_j, T_j, X_j, P_{j-1})$ does not represent frequencies but an expert's opinions? We are concerned with the circumstance in which the experiment is concluded, and the expert's degrees of belief, we will suppose, have converged so far as they will. The expert is uncertain as to whether there are common causes of X and outcome, or how many there are, but all of the causal structures she entertains are like figure 12 and none are like figure 13. Conditional on θ and any particular causal hypothesis we suppose the Markov condition is satisfied, but the expert's actual degrees of belief are some mixture over different causal structures. Should $f_{\theta}(O_j|T_j, X_j)$ be invariant under manipulation of T_j in the opinion of the expert when her distribution for a given value of θ is a mixture of several different causal hypotheses? The answer is yes, as the following argument shows.

Let us call the experimental (unmanipulated in the sense of Chapter 7) population Exp , and the hypothetical population subjected to some policy based on the results of the experiment Pol . Let $f_{\theta, Exp}(O_j, T_j, X_j, P_{j-1})$ represent the expert's degrees of belief about O_j , T_j , X_j , and P_{j-1} conditional on θ in the experimental population, and $f_{\theta, Pol}(O_j, T_j, X_j, P_{j-1})$ represent the expert's degrees of belief about O_j , T_j , X_j , and P_{j-1} conditional on θ in the hypothetical population subjected to some policy. Let CS be a random variable that denotes a causal structure. We have already noted that it follows from Theorem 7.1 that

$$f_{\theta, Exp}(O_j|T_j, X_j, CS) = f_{\theta, Pol}(O_j|T_j, X_j, CS)$$

Because θ determines the density of O_j conditional on T_j and X_j ,

$$\begin{aligned} f_{\theta, \text{Exp}}(O_j | T_j, X_j, CS) &= f_{\theta, \text{Exp}}(O_j | T_j, X_j) \\ f_{\theta, \text{Pol}}(O_j | T_j, X_j, CS) &= f_{\theta, \text{Pol}}(O_j | T_j, X_j) \end{aligned}$$

Hence,

$$f_{\theta, \text{Exp}}(O_j | T_j, X_j) = f_{\theta, \text{Pol}}(O_j | T_j, X_j)$$

Consider next the question of "bias" raised by Kadane and Seidenfeld. The very notion requires us to consider not just degrees of belief but also some facts and some potential facts. We suppose there is really a correct (or nearly correct) value for the parameters in the likelihood model, and the true values describe features of the process that go on in the experiment. We suppose the expert converges to the truth, so that her posterior distribution is concentrated around the true values. What the public that pays for these experiments cares about is whether the expert's views about the best treatment are correct: *Would a policy that puts in place the expert's preferred rule of treatment, say $T = g(X)$, result in better outcomes than alternative policies under consideration?* One way to look at that question is to ask if the expert's expected values for outcome conditional on X profile and treatment roughly equal the population mean for outcome under these conditions. If degrees of belief accord with population distributions, that is just to ask when the frequency of O conditional on T and X that would result if every relevant person in the population were treated on the basis of the experimental assignment rule $T = h(X, P_{j-1})$ would be roughly the same as the frequency of O conditional on T and X for the general population if a revised rule $T = g(X)$ for assigning treatments were used. In other words: *Will the frequency of O conditional on T and X be invariant under a direct manipulation of T ?*

As we have just seen for this case, the Markov Condition and Theorem 7.1 *entail* that for the graph in figure 12, and all others like it (those in which every trek whose source is a common cause of O_j and T_j contains an X_j variable, O_j does not cause T_j , and every common cause of an X_j variable and T_j is an X_j variable) the probability of O_j on T_j and X_j is invariant under a direct manipulation of T_j . No other assumptions are required. The example is a very special case of a general sufficient condition for the invariance of conditional probabilities under a direct manipulation.

Consider next why the experimental design forbids that treatment assignment depend directly on "unrecorded" features of the patients, such as Y . Suppose such assignments were allowed; Kadane and Seidenfeld say the outcome might be "contaminated" which we understand to mean

that some unrecorded causes of the patient preference, and hence of T_j , might also be causes of O_j . So that we have the causal picture in figure 14.

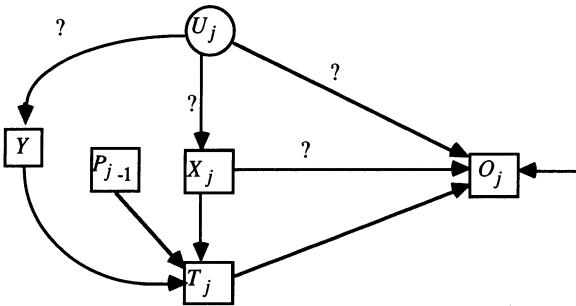


Figure 14

The question marks indicate that, we (or the expert), are uncertain about whether the corresponding causal influences exist. Suppose the directed edges from U_j to Y and from U_j to O_j exist. Then there is an undirected path between O_j and T_j that contains Y . In this case the Markov condition entails that the probability of O_j conditional on T_j and the X_j variables is *not* invariant under a direct manipulation of T_j except for "coincidental" parameter values. So if unrecorded Y values were allowed to influence assignments then for all we or the experts know, the expert's prediction of the effects of his proposed rule $T = g(X)$ would be wrong.

Let us now return to the question of why patient preference cannot be used to influence treatment. The reason why patient preferences cannot be used to determine treatment assignment in the experiment is not only because there may be, for all one knows, a causal interaction between patient preference and treatment outcome. It is true that if it were known that no such confounding occurs, then patient preference could be used in treatment assignment, but why cannot such assignment be used even if there is confounding? In order to make treatment assignments depend on patient preference (and presumably also on other features, such as the X_j variables) the patients' preferences must be ascertained. If the preferences are known, why not conditionalize outcome on *Preference*, T and the X_j variables, just as we have conditionalized outcome on T and the X_j variables? The probability of O_j conditional on *Preference*, the X_j variables, and T_j has no formally different role than the probability of O_j conditional on the X_j variables and T_j . If in figure 14 we make $Y = \text{Preference}$, then according to the Markov condition and Theorem 7.1, the probability of O_j conditional on T_j , X_j and *Preference* is invariant under a manipulation of T_j . Of course some precautions would have to

be taken in the course of an experiment that allows patient preference to influence treatment assignment. There is not much point to allowing a patient to choose his or her treatment unless the choice is informed. In the experimental setting, it would be necessary to standardize the information and advice that each patient received.

Could this design actually be used to predict the effects of a policy of treatment? At the end of the study the expert has a density function of outcomes given T , X , and *Preference*; subsequent patients' preferences for treatment would have to be recorded (but not necessarily used in determining treatment). If the announced results of the study alter *Preference*, the probability of outcome conditional on *Preference*, X and T depends on whether influences represented by the edges adjacent to U in figure 12 exist. But if patients are informed of the experimental results we must certainly expect in many cases that their preferences will be changed, i.e. announcing the result of the experiment is a direct manipulation of *Preference*. Reliable predictions could only be made if the experimental outcome were kept secret!

The reason that patient preferences cannot be used for assigning treatment is therefore, not just because their preferences might have complicated interactions with the outcome of the treatment--so might the X variables. No analysis that does not also consider how a policy changes variables that were relevant to outcome in an experimental study can give a complete account of when predictions can be relied upon. As Kadane has pointed out¹, it is more likely that the causes of *Preference* in the experimental population are different from the causes of *Preference* in the non-experimental population than it is that the causes of X in the experimental population are different from the causes of X in the non-experimental population. In the case we are considering, announcements of experimental results (or of recommendations) about policies that use patient preferences for assigning treatment can generally be expected to directly change those very preferences--whereas policies that use the X variables for assigning treatment do not generally change the values of the X variables for people in the population. (Of course, there may be instances where the results of a study that does not base treatment on patient preference also directly manipulates the distribution of the X variables, in which case the prediction of outcome conditional on treatment and the X variables would also be unreliable. Suppose, fancifully, in experimental trials that assign treatment as a function of cholesterol levels, it were found that a certain drug is very effective against cancer for subjects with low cholesterol.)

¹Personal communication.

The Kadane/Sedransk/Seidenfeld design thus reveals an ethical conundrum that conventional methodological prejudices against non-randomized trials has hidden. There is an obligation to find the most effective and cost effective treatments, and an obligation to take into account in treatment the preferences of people who participate as subjects in clinical trials. Both can be satisfied. But there is also an obligation fully to inform patients about the relevant scientific results that bear on decisions about their treatment. This obligation is incompatible with the others.

9.4.3 Towards Ethical Trials

Finally, we can use the causal analysis to obtain some more optimistic results about patient selection of treatment in experimental trials. Suppose in an experiment treatment assignment is a function $T = h(X_j, \text{Preference}, P_{j-1})$, and every undirected path between *Preference* and O_j contains some member of X_j as a non-collider. (If this is the case then we will say that *Preference* is not confounded with O .) Then it can be shown strictly as a consequence of the Markov and Faithfulness Conditions that the probability of O_j conditional on T_j and X_j is invariant under a direct manipulation of T_j ; we may or may not conditionalize on *Preference*, or take *Preference* into account in the treatment rule used in policy, and in that case whether or not the announced experimental results changes the distribution of preferences is irrelevant to the accuracy of predictions. Now in some cases it may very well be that patient preference is not confounded with treatment outcome. If investigators could in fact discover that *Preference* is not confounded with O , then they could let such preferences be a factor both in treatment assignments in the experimental protocol and in the recommended policy. Kadane and Seidenfeld say that such dependencies, if they exist, are undetectable. If the experts are completely ignorant about what factors do not influence the outcome of treatment, Kadane and Seidenfeld are right; but if the experts know something, anything, that varies with patients and that has no effect on outcome except through treatment and has no common cause with outcome, we disagree. The something could be the phase of the moon on the patient's last birthday, the angular separation of the sun and Jupiter on the celestial sphere on the day of the patient's mother's day of birth, or simply the output of some randomizing device assigned to each patient. How is that?

In any distribution faithful to the graph of figure 15, E and C are dependent conditional on B . The relation is necessary in linear models, without assuming faithfulness.

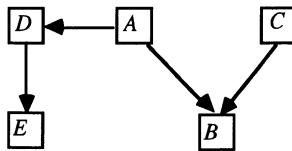


Figure 15

Now, returning to our problem, let Z be any feature whatsoever that varies from patient to patient, and that the experts agree in regarding as independent of patients' preferences for treatments and as affecting outcome only through treatment. Adopt a rule in the experiment that makes treatment a function of *Preference*, the patient's X profile, P_{j-1} , and the patient's Z value. Then the expert view of the causal process in the experiment looks something like figure 16:

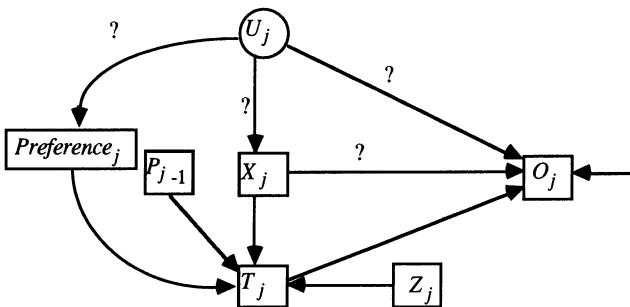


Figure 16

If O_j and Z_j are independent conditional on T_j and X_j then there is (assuming faithfulness) no path d-connecting $Preference_j$ and O_j given X_j that is into $Preference_j$. A confounding relation between $Preference_j$ and O_j , if it exists, can be discovered from the experimental data. (Similarly, if T_j and O_j are dependent given X_j because the experimental population consists of a mixture of causal structures, then O_j and Z_j are dependent conditional on T_j and X_j unless some particular set of parameter values produces a "coincidental" independence.) Indeed, on the rather brave assumption that all dependencies are linear, Z_j is an instrumental variable (Bowden and Turkington, 1984) and the linear coefficient representing the influence of T conditional on O and X can be calculated from the correlations and partial correlations.

This suggests that it is possible to do a pilot study to determine whether *Preference* is confounded with O in the experimental population. In the pilot study, *Preference* can be a factor

influencing, but not completely determining, T . If the results of the pilot study indicate that *Preference* is not confounded with O in the experimental population, a larger study in which *Preference* completely determines T can be done; otherwise, the Kadane/Sedransk/Seidenfeld design can be employed.

The goal of a medical experiment might be to predict outcomes in a population where a policy of assigning treatments without consulting patient preference is adopted. For example, the question might be "What would the death rate be if only halothane were used as a general anesthetic?" In this case, patient preference has little or nothing to do with the assigned treatment. If patient preference is not used to assign treatment in the policy population there is no reason to think that predictions of $P(O|X,T)$ in the policy population will be inaccurate when based upon experiments in which patients choose (or at least influence) their treatment, and *Preference* and O are not confounded.

It might be, however, that the goal of the experiment is to predict $P(O|X,T)$ in the policy population, and to let the patients choose or at least influence the choice of the treatment they receive. For example, in choosing between lumpectomy and mastectomy, patient preference may be the deciding factor. In this case there are a number of reasons to question the accuracy of a prediction of $P(O|X,T)$ in the policy population based upon the design we propose. But in this case every design meets the same difficulties, whether or not patients have assigned themselves in experimental treatments. These are equally good reasons for questioning the accuracy of a prediction of $P(O|X,T)$ (interpreted as frequencies or propensities and not as degrees of belief) in the policy population based upon the Kadane/Sedransk/Seidenfeld or a classical randomized design. The fundamental problem is that there are any number of plausible ways in which the causal relationships among preference and other variables in the experimental population may be different from the causal relationships in the policy population.² For example, in the experimental population the assignment of treatment will not depend on the patient's income. However, in the actual population, the choice of treatment may very well depend upon income. There could easily be a common causal pathway connecting income and outcome that does not contain any variable in the patient's X profile. Again, in the experimental population the information and advice patients receive can be standardized. We can also try and ensure that the advice given is a function only of the patient's X profile. In the policy population, however, the advice and information that patients receive cannot be controlled in this way. If this is the case, the determination of preference may be a mixture of different causal structures in the policy population. Finally, the determination of patient preference in the policy

²We thank Jay Kadane for pointing out that the causal relationship between *Preference* and other variables might be different in the experimental and non-experimental populations, even if *Preference* is not directly manipulated.

population could easily be unstable. There are fads and fashions among patients, and also fads and fashions among doctors. New information could be released, or an intensive advertising campaign introduced. Any of these might create a trek between *Preference* and O , and hence between T and O , that does not contain any member of the X profile as a non-collider.

So even if in the experimental population *Preference* and O are not confounded, they very well might be in the policy population. If they are confounded in the policy population then $P(O|X,T)$ will not be the same in the experimental and policy populations (unless the parameters of the different causal structures coincidentally have values that make them equal.) Note that the same is true of predictions of $P(O|X,T)$ based on the Kadane/Sedransk/Seidenfeld design or of a prediction of $P(O|T)$ based upon a randomized experiment. This does not mean that no useful predictions can be made in situations where patient preference will be used to influence treatment in the policy population. It is still possible to inform the patient what $P(O|X,T)$ would be if a particular treatment were given *without* patient choice. The patient can use this information to help make an informed decision. And with the design we have proposed, this (counterfactual) prediction is accurate as long as *Preference* and O are not confounded in the *experimental* population, regardless of how *Preference* is causally connected to O in the policy population.

Suppose then that we are merely trying to predict $P(O|X,T)$ in a population in which everyone is *assigned* a treatment. Is the design we have suggested practical? One potential problem is the obligation to give patients who are experimental subjects advice and information about their treatments; if this were not done the experiment would be unethical. If patients have access to advice from physicians, the advice is likely to be based upon their X profile. Even if the only information subjects receive is that all of the experts agree that they should not choose treatment T_1 , their X profile is a cause of their preference. Will giving this advice and information make it unlikely that $P(O|X,T)$ is invariant under manipulation of T ? It is true that the variables in the X profiles were chosen to be variables thought to be causes of O or to have common causes with O . Hence advice of this kind is very likely to *create* a common cause of *Preference* and O in the experimental population. Hence, it is likely that in the experimental population there will be a trek between T and O that is into T and contains *Preference*. However, such a trek would *not* d-connect T and O given X because it would also contain some member of X as a non-collider. (See figure 17.) Hence such a trek does not invalidate the invariance of $P(O|X,T)$ under manipulation of T . Moreover, there is no problem with changing the advice as the experiment progresses, under the assumption that P_{j-1} is causally connected to O_j only through *Preference_j*.

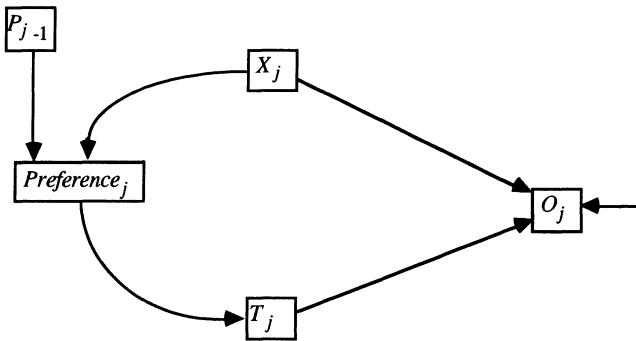


Figure 17

Can we let patients choose their own treatment, or merely influence the choice of treatment? As long as we are merely trying to predict $P(O|X,T)$ in a population in which everyone is *assigned* a treatment we can let patients choose their own treatment, as long as this doesn't result in all patients with a particular X profile always failing to choose some treatment T_1 . In that case, $P(O|X,T = T_1)$ is undefined in the experimental population and cannot be used to predict $P(O|X,T = T_1)$ in a population where that quantity is defined.

In summary, so long as the goal is to predict $P(O|X,T)$ in a population where everyone is assigned a treatment, and there is enough variation of choice of treatments among the patients, and the experimental population is not a mixture of causal structures, and *Preference* and O are not confounded in the experimental population (an issue that must be decided empirically) it is possible to make accurate predictions from an experimental population in which informed patients choose their own treatment. If it is really important to let patient preferences influence their treatment in experiments, then it is worth risking some cost to realize that condition if it is possible to do so consistent with reliable prediction. How much it is worth, either in money or in degradation in confidence about the reliability of predictions, is not for us to say. But a simple modification of the Kadane/Sedransk/Seidenfeld design which has initial trials base treatment assignment on X_j , P_{j-1} , Z_j , and *Preference_j*, and then allows patient self-assignment if *Preference_j* and O_j are discovered to be unconfounded, would in some cases permit investigators to conduct clinical trials that conform to ethical requirements of autonomy and informed consent.

9.5 An Example: Smoking and Lung Cancer

The fascinating history of the debates over smoking and lung cancer illustrates the difficulties of causal inference and prediction from policy studies, and also illustrates some common mistakes. Perhaps no other hypothetical cause and effect relationship has been so thoroughly studied by non-experimental methods or has so neatly divided the professions of medicine and statistics into opposing camps. The theoretical results of this and the preceding chapters provide some insight into the logic and fallacies of the dispute.

The thumbnail sketch is as follows: In the 1950s a retrospective study by Doll and Hill (1952) found a strong correlation between cigarette smoking and lung cancer. That initial research prompted a number of other studies, both retrospective and prospective, in the United States, the United Kingdom, and soon after in other nations, all of which found strong correlations between cigarette smoking and lung cancer, and more generally between cigarette smoking and cancer and between cigarette smoking and mortality. The correlations prompted health activists and some of the medical press to conclude that cigarette smoking causes death, cancer, and most particularly, lung cancer. Sir Ronald Fisher took very strong exception to the inference, preferring a theory in which smoking behavior and lung cancer are causally connected only through genetics. Fisher wrote letters, essays, and eventually a book against the inference from the statistical dependencies to the causal conclusion. Neyman ventured a criticism of the evidence from retrospective studies. The heavyweights of the statistical profession were thus allied against the methods of the medical community. A review of the evidence containing a response to Fisher and Neyman was published in 1959 by Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin, and Wynder. The Cornfield paper became part of the blueprint for the Report of the Surgeon General on Smoking and Health in 1964, which effectively established that as a political fact smoking would be treated as an unconfounded cause of lung cancer, and set in motion a public health campaign that is with us still. Brownlee (1965) reviewed the 1964 report in the *Journal of the American Statistical Association* and rejected its arguments as statistically unsound for many of the reasons one can imagine Fisher would have given. In 1979, the Surgeon General published a second report on smoking and health, repeating the arguments of the first report but with more extensive data, but offering no serious response to Brownlee's criticisms. The report made strong claims from the evidence, in particular that cigarette smoking was the largest preventable cause of death in the United States. The foreword to the report, by Joseph Califano, was downright vicious, and claimed that any criticism of the conclusions of the report was an attack on science itself. That did not stop P. Burch (1983), a physicist turned theoretical biologist turned statistician, from publishing a lengthy criticism of

the second report, again on grounds that were detailed extensions of Fisher's criticisms, but buttressed as well by the first reports of randomized clinical trials of the effects of smoking intervention, all of which were either null or actually suggested that intervention programs increased mortality. Burch's remarks brought a reply by A. Lilienfeld (1983), which began and ended with an *ad hominem* attack on Burch.

Fisher's criticisms were directed against the claim that uncontrolled observations of a correlation between smoking and cancer, no matter whether retrospective or prospective, provided evidence that smoking causes lung cancer, as against the alternative hypothesis that there are one or more common causes of smoking and lung cancer. His strong views can be understood in the light of features of his career. Fisher had been largely responsible for the introduction of randomized experimental designs, one of the very points of which was to obtain statistical dependencies between a hypothetical cause and effect that could not be explained by the action of unmeasured common causes. Another point of randomization was to insure a well-defined distribution for tests of hypotheses, something Fisher may have doubted was available in observational studies. Throughout his adult life Fisher's research interests had been in heredity, and he had been a strong advocate of the eugenics movement. He was therefore disposed to believe in genetic causes of very detailed features of human behavior and disease. Fisher thought a likely explanation of the correlation of lung cancer and smoking was that a substantial fraction of the population had a genetic predisposition both to smoke and to get lung cancer.

One of Fisher's (1959, p. 8) fundamental criticisms of these epidemiological arguments was that correlation underdetermines causation: besides smoking causing cancer, wrote Fisher "there are two classes of alternative theories which any statistical association, observed without the precautions of a definite experiment, always allows--namely, (1) that the supposed effect is really the cause, or in this case that incipient cancer, or a pre-cancerous condition with chronic inflammation, is a factor in inducing the smoking of cigarettes, or (2) that cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype." Not even Fisher took (1) seriously. To these must be added others Fisher did not mention, for example that smoking and lung cancer have several distinct unmeasured common causes, or that while smoking causes cancer, something unmeasured also causes both smoking and cancer.

If we interpret "statistical association" as statistical dependence, Fisher is correct that given observation only of a statistical dependence between smoking and lung cancer in an uncontrolled study, the possibility that smoking does not cause lung cancer cannot be ruled out. However, he does not mention the possibility that this hypothesis, if true, could have been

established without experimentation by finding a factor associated with smoking but independent, or conditionally independent (on variables other than smoking) of cancer. By the 1960s a number of personal and social factors associated with smoking had been identified, and several causes of lung cancer (principally associated with occupational hazards and radiation) potentially independent of smoking had been identified, but their potential bearing on questions of common causes of smoking and lung cancer seems to have gone unnoticed. The more difficult cases to distinguish are the hypotheses that smoking is an unconfounded cause of lung cancer versus the joint hypotheses that smoking causes cancer and that there is also an unmeasured common cause--or causes--of smoking and cancer.

Fisher's hypothesis that genotype causes both smoking behavior and cancer was speculative, but it wasn't a will-o-the-wisp. Fisher obtained evidence that the smoking behavior of monozygotic twins was more alike than the smoking behavior of dizygotic twins. As his critics pointed out, the fact could be explained on the supposition that monozygotic twins are more encouraged by everyone about them to do things alike than are dizygotic twins, but Fisher was surely correct that it could also be explained by a genetic disposition to smoke. On the other side, Fisher could refer to evidence that some forms of cancer have genetic causes.

The paper by Cornfield, et al. (including Lilienfeld) argued that while lung cancer may well have other causes besides, cigarette smoking causes lung cancer. This view had already been announced by official study groups in the United States and Great Britain. Cornfield's paper is of more scientific interest than the Surgeon General's report five years later, in part because the former is not primarily a political document. Cornfield, et al. claimed the existing data showed several things:

1. Carcinomas of the lung found at autopsy had systematically increased since 1900, although different studies gave different rates of increase. Lung cancers are found to increase monotonically with the amount of cigarette smoking and to be higher in current than in former cigarette smokers. In large prospective studies diagnoses of lung cancer may have an unknown error rate, but the total death rate also increases monotonically with cigarette smoking.
2. Lung cancer mortality rates are higher in urban than in rural populations, and rural people smoke less than city people, but in both populations smokers have higher death rates from lung cancer than do non-smokers.

3. Men have much higher death rates from lung cancer than women, especially among persons over 55, but women smoked much less and as a class had taken up the habit much later than men.
4. There are a host of causes of lung cancer, including a variety of industrial pollutants and unknown circumstances associated with socioeconomic class, with the poorer and less well off more likely than the better off to contract the disease, but no more likely to smoke. Cornfield, et al. emphasize that "The population exposed to established industrial carcinogens is small, and these agents cannot account for the increasing lung-cancer risk in the remainder of the population. Also, the effects associated with socioeconomic class and related characteristics are smaller than those noted for smoking history, and the smoking class differences cannot be accounted for in terms of these other effects." (p.179). This passage states that the difference in cancer rates for smokers and non-smokers could not be explained by socioeconomic differences. While this claim was very likely true, no analysis was given in support of it, and the central question of whether smoking and lung cancer were independent or nearly independent conditional on all subsets of the known risk factors *that are not effects* of smoking and cancer--area of residence, exposure to known carcinogens, socioeconomic class, and so on, was not considered. Instead, Cornfield et al. note that different studies measured different variables and "The important fact is that in all studies when other variables are held constant, cigarette smoking retains its high association with lung cancer."
5. Cigarette smoking is not associated with increased cancer of the upper respiratory tract, the mouth tissues or the fingers. Carcinoma of the trachea, for example, is a rarity. But, Cornfield, et al. point out, "There is no a priori reason why a carcinogen that produces bronchogenic cancer in man should also produce neoplastic changes in the anspharynx or in other sites." (p. 186).
6. Experimental evidence shows that cigarette smoke inhibits the action of the cilia in cows, rats and rabbits. Inhibition of the cilia interferes with the removal of foreign material from the surface of the bronchia. Damage to ciliated cells is more frequent in smokers than in nonsmokers.
7. Application of cigarette tar directly to the bronchia of dogs produced changes in the cells, and in some but not other experiments applications of tobacco tar to the skin of mice produced cancers. Exposure of mice to cigarette smoke for up to 200 days produced cell changes but no cancers.

8. A number of aromatic polycyclic compounds have been isolated in tobacco smoke, and one of them, the α form of benzopyrene, was known to be a carcinogen.

Perhaps the most original technical part of the argument was an a kind of sensitivity analysis of the hypothesis that smoking causes lung cancer. Cornfield, et al. considered a single hypothetical binary latent variable causing lung cancer and statistically dependent on smoking behavior. They argued such a latent cause would have to be almost perfectly associated with lung cancer and strongly associated with smoking to account for the observed association. The argument neglected, however, the reasonable possibility of multiple common causes of smoking and lung cancer, and had no clear bearing on the hypothesis that the observed association of smoking and lung cancer is due both to a direct influence and to common causes.

In sum, Cornfield, et al. thought they could show a mechanism for smoking to cause cancer, and claimed evidence from animal studies, although their position in that regard tended to trip over itself (compare items 5 and 7). They didn't put the statistical case entirely clearly, but their position seems to have been that lung cancer is also caused by a number of measurable factors that are not plausibly regarded as effects of smoking but which may cause smoking, and that smoking and cancer remain statistically dependent conditional on these factors. Against Fisher they argued as follows:

The difficulties with the constitutional hypothesis include the following considerations:
(a) changes in lung-cancer mortality over the last half century; (b) the carcinogenicity of tobacco tars for experimental animals; (c) the existence of a large effect from pipe and cigar tobacco on cancer of the buccal cavity and larynx but not on cancer of the lung; (d) the reduced lung-cancer mortality among discontinued cigarette smokers. No one of these considerations is perhaps sufficient by itself to counter the constitutional hypothesis, ad hoc modification of which can accommodate each additional piece of evidence. A point is reached, however, when a continuously modified hypothesis becomes difficult to entertain seriously. (p. 191)

Logically, Cornfield et al. visited every part of the map. The evidence was supposed to be inconsistent with a common cause of smoking and lung cancer, but also consistent with it. Objections that a study involved self-selection--as Fisher and company would object to (d)--was counted as an "ad hoc modification" of the common cause hypothesis. The same response was in effect given to the unstated but genuine objections that the time series argument ignored the combined effects of dramatic improvements in diagnosis of lung cancer, a tendency of physicians to bias diagnoses of lung cancer for heavy smokers and to overlook such a diagnosis

for light smokers, and the systematic increase in the same period of other factors implicated in lung cancer, such as urbanization. The rhetoric of Cornfield, et al. converted reasonable demands for sound study designs into *ad hoc* hypotheses. In fact none of the evidence adduced was inconsistent with the "constitutional hypothesis."

A reading of the Cornfield paper suggests that their real objection to a genetic explanation was that it would require a very close correlation between genotypic differences and differences in smoking behavior and liability to various forms of cancer. Pipe and cigar smokers would have to differ genotypically from cigarette smokers; light cigarette smokers would have to differ genotypically from heavy cigarette smokers; those who quit cigarette smoking would have to differ genotypically from those who did not. Later the Surgeon General would add that Mormons would have to differ genotypically from non-Mormons and Seventh Day Adventists from non-Seventh Day Adventists. The physicians simply didn't believe it. Their skepticism was in keeping with the spirit of a time in which genetic explanations of behavioral differences were increasingly regarded as politically and morally incorrect, and the moribund eugenics movement was coming to be viewed in retrospect as an embarrassing bit of racism.

In 1964 the Surgeon General's report reviewed many of the same studies and arguments as had Cornfield, but it added a set of "Epidemiological Criteria for Causality," said to be sufficient for establishing a causal connection and claimed that smoking and cancer met the criteria. The criteria were indefensible, and they did not promote any good scientific assessment of the case. The criteria were the "consistency" of the association, the "strength" of the association, the "specificity" of the association, the temporal relationship of the association and the "coherence" of the association.

All of these criteria were left quite vague, but no way of making them precise would suffice for reliably discriminating causal from common causal structures. Consistency meant that separate studies should give the "same" results, but in what respects results should be the same was not specified. Different studies of the relative risk of cigarette smoking gave very different multipliers depending on the gender, age and nationality of the subjects. The results of most studies were the same in that they were all positive; they were plainly not nearly the same in the seriousness of the risk. Why stronger associations should be more likely to indicate causes than weaker associations was not made clear by the report. Specificity meant the putative cause, smoking, should be associated almost uniquely with the putative effect, lung cancer. Cornfield, et al. had rejected this requirement on causes for good reason, and it was palpably violated in the smoking data presented by the Surgeon General's report. "Coherence" in the jargon of the report meant that no other explanation of the data was possible, a criterion the observational data

did not meet in this case. The temporal issue concerned the correlation between increase in cigarette smoking and increase in lung cancer, with a lag of many years. Critics pointed out that the time series were confounded with urbanization, diagnostic changes and other factors, and that the very criterion Cornfield, et al. had used to avoid the issue of the unreliability of diagnoses, namely total mortality, was, when age-adjusted, uncorrelated with cigarette consumption over the century.

Brownlee (1965) made many of these points in his review of the report in the *Journal of the American Statistical Association*. His contempt for the level of argument in the report was plain, and his conclusion was that Fisher's alternative hypothesis had not been eliminated or even very seriously addressed. In Brownlee's view, the Surgeon General's report had only two arguments against a genetic common cause: (a) the genetic hypothesis would allegedly have to be very complicated to explain the dose/response data, and (b) the rapid historical rise in lung cancer following by about 20 years a rapid historical rise in cigarette smoking. Brownlee did not address (a), but he argued strongly that (b) is poor evidence because of changes in diagnostics, changes in other factors of known and unknown relevance, and because of changes in the survival rate of weak neonates whom, as adults, might be more prone to lung cancer.

One of the more interesting aspects of the review was Brownlee's "very simplified" proposal for a statistical analysis of " E_2 causes E_1 " which was that E_1 and E_2 be dependent conditional on every possible vector of values for all other variables of the system. Brownlee realized, of course, that his condition did not separate " E_2 causes E_1 " from E_1 causes E_2 ," but that was not a problem with smoking and cancer. But even ignoring the direction of causation, Brownlee's condition--perhaps suggested to him by the fact that the same principle is erroneously used in regression--is quite wrong. It would be satisfied, for example, if, E_1 and E_2 had no causal connection whatsoever provided some measured variable E_j were a direct effect of both E_1 and E_2 .

Brownlee thought his way of considering the matter was important for prediction and intervention:

If the inequality holds only for, say, one particular subset E_j, \dots, E_k , and for all other subsets equality holds, and if the subset E_j, \dots, E_k occurs in the population with low probability, then $\Pr\{E_1|E_2\}$, while not strictly equal to $\Pr\{E_1|E_2^C\}$, will be numerically close to it, and then E_2 as a cause of E_1 may be of small practical importance. These considerations are related to the Committee's responsibility for assessment of the

magnitude of the health hazard (page 8). Further complexities arise when we distinguish between cases in which one of the required secondary conditions E_j, \dots, E_k is, on the one hand, presumably controllable by the individual, e.g., the eating of parsnips, or uncontrollable, e.g., the presence of some genetic property. In the latter case, it further makes a difference whether the genetic property is identifiable or non-identifiable: for example it could be brown eyes which is the significant subsidiary condition E_j , and we could tell everybody with not-brown eyes it was safe for *them* to smoke. (p. 725)

No one seems to have given any better thought than this to the question of how to predict the effects of public policy intervention against smoking. Brownlee regretted that the Surgeon General's report made no explicit attempt to estimate the expected increase in life expectancy from not smoking or from quitting after various histories.

Fifteen years later, in 1979, the second Surgeon General's Report on Smoking and Health was able to report studies that showed a monotonic increase in mortality rates with virtually every feature of smoking practice that increased smoke in the lungs: number of cigarettes smoked per day, number of years of smoking, inhaling versus not inhaling, low tar and nicotine versus high tar and nicotine, length of cigarette habitually left unsmoked. The monotonic increase in mortality rates with cigarette smoking had been shown in England, the continental United States, Hawaii, Japan, Scandinavia and elsewhere, for whites and blacks, for men and women. The report dismissed Fisher's hypothesis in a single paragraph by citing a Scandinavian study (Cederlof, Friberg and Lundman, 1977) that included monozygotic and dizygotic twins:

When smokers and nonsmokers among the dizygotic pairs were compared, a mortality ratio of 1.45 for males and 1.21 for females was observed. Corresponding mortality ratios for the monozygotic pairs were 1.5 for males and 1.222 for females. Commenting on the constitutional hypothesis and lung cancer, the authors observed that "the constitutional hypothesis as advanced by Fisher and still supported by a few, has here been tested in twin studies. The results from the Swedish monozygotic twin series speak strongly against the constitutional hypothesis."

The second Surgeon General's report claimed that tobacco smoking is responsible for 30% of all cancer deaths; cigarette smoking is responsible for 85% of all lung cancer deaths.

A year before the report appeared, in a paper for the British Statistical Association P. Burch (1978) had used the example of smoking and lung cancer to illustrate the problems of distinguishing causes from common causes without experiment. In 1982 he published a full

fledged assault on the second Surgeon General's report. The criticisms of the argument of the report were similar to Brownlee's criticisms of the 1964 report, but Burch was less restrained and his objections more pointed. His first criticism was that while all of the studies showed a increase in risk of mortality with cigarette smoking, the degree of increase varied widely from study to study. In some studies the age adjusted multiple regression of mortality on cigarettes, beer, wine and liquor consumption gave a smaller partial correlation with cigarettes than with beer drinking. Burch gave no explanation of why the regression model should be an even approximately correct account of the causal relations. Burch thought the fact that the apparent dose/response curve for various culturally, geographically and ethnically distinct groups were very different indicated that the effect of cigarettes was significantly confounded with environmental or genetic causes. He wanted the Surgeon General to produce a unified theory of the causes of lung cancer, with confidence intervals for any relevant parameter estimates: Where, he asked, did the 85% figure come from?

Burch pointed out, correctly, that the cohort of 1487 dizygotic and 572 monozygotic twins in the Scandinavian study born between 1901 and 1925 gave no support at all to the claim that the constitutional explanation of the connection between smoking and lung cancer had been refuted, despite the announcements of the authors of that study. The study showed that of the dizygotes exactly 2 nonsmokers or infrequent smokers had died of lung cancer and 10 heavy smokers had died of lung cancer; of the monozygotes, 2 low non smokers and 2 heavy smokers had died of the disease. The numbers were useless, but if they suggested anything, it was that if genetic variation was controlled there is no difference in lung cancer rates between smokers and nonsmokers. The Surgeon General's report of the conclusion of the Scandinavian study was accurate, but not the less misleading for that.

Burch also gave a novel discussion of the time series data, arguing that it virtually refuted the causal hypothesis. The Surgeon General and others had used the time series in a direct way. In the U.K. for example, male cigarette consumption per capita had increased roughly a hundredfold between 1890 and 1960, with a slight decrease thereafter. The age-standardized male death rate from lung cancer began to increase steeply about 1920, suggesting a thirty year lag, consistent with the fact that people often begin smoking in their twenties and typically present lung cancer in their fifties. According to Burch's data, the onset of cigarette smoking for women lagged behind males by some years, and did not begin until the 1920s. The Surgeon General's report noted that the death rate from lung cancer for women had also increased dramatically between 1920 and 1980. Burch pointed out that the autocorrelations for the male series and female series didn't mesh: there was no lag in death rates for the women. Using U.K. data, Burch plotted the *percentage change* in the age-standardized death rate from lung

cancer for both men and women from 1900 to 1980. The curves matched perfectly until 1960. Burch's conclusion is that whatever caused the increase in death rates from lung cancer affected both men and women at the same time, from the beginning of the century on, although whatever it is had a smaller absolute effect on women than on men. But then the whatever-it-was could not have been cigarette smoking, since increases in women's consumption of cigarettes lagged twenty to thirty years behind male increases.

Burch was relentless. The Surgeon General's report had cited the low occurrence of lung cancer among Mormons. Burch pointed out that Mormon's in Utah not only have lower age-adjusted incidences of cancer than the general population, but also have higher incidences than non-Mormon nonsmokers in Utah. Evidently their lower lung cancer rates could not be simply attributed to their smoking habits.

Abraham Lilienfeld, who only shortly before had written a textbook on epidemiology and who had been involved with the smoking and cancer issue for more than twenty years, published a reply to Burch that is of some interest. Lilienfeld gives the impression of being at once defensive and disdainful. His defense of the Surgeon General's report began with an *ad hominem* attack, suggesting that Burch was so out of fashion as to be a crank, and ended with another *ad hominem*, demanding that if Burch wanted to criticize others' inferences from their data he go get his own. The most substantive reply Lilienfeld offered is that the detailed correlation of lung cancer with smoking habits in one subpopulation after another makes it seem very implausible that the association is due to a common cause. Lilienfeld said, citing himself, that the conclusion that 85% of lung cancer deaths are due to cigarettes is based on the relative risk for cigarette smokers and the frequency of cigarette smoking in the population, predicting, in effect, that if cigarette smoking ceased the death rate from lung cancer would decline by that percentage. (The prediction would only be correct, Burch pointed out in response, provided cigarette smoking is a completely unconfounded cause of lung cancer.) Lilienfeld challenged the source of Burch's data on female cigarette consumption early in the century, which Burch subsequently admitted were estimates.

Both Burch and Lilienfeld discussed a then recent report by Rose et al. (1982) on a ten year randomized smoking intervention study. The Rose study, and another that appeared at nearly the same time with virtually the same results, illustrates the hazards of prediction. Middle-aged male smokers were assigned randomly to a treatment or non-treatment group. The treatment group was encouraged to quit smoking and given counseling and support to that end. By self-report, a large proportion of the treatment group either quit or reduced cigarette smoking. The difference in self-reported smoking levels between the treatment and non-treatment groups was

thus considerable, although the difference declined toward the end of the ten year study. To most everyone's dismay, Rose found that there was no statistically significant difference in lung cancer between the two groups after ten years (or after five), but there was a difference in overall mortality--the group that had been encouraged to quit smoking, and had in part done so, suffered higher mortality.

Fully ignoring their own evidence, the authors of the Rose study concluded nonetheless that smokers should be encouraged to give up smoking, which makes one wonder why they bothered with a randomized trial. Burch found the Rose report unsurprising; Lilienfeld claimed the numbers of lung cancer deaths in the sample are too small to be reliable, although he did not fault the Surgeon General's report for using the Scandinavian data, where the numbers are even smaller, and he simply quoted the conclusion of the report, which seems almost disingenuous. To Burch's evident delight, as Lilienfeld's defense of the Surgeon General appeared so did yet further experimental evidence that intervening in smoker's behavior has no benign effect on lung cancer rates. The Multiple Risk Factor Intervention Trial Research Group (1982) reported the results after six years of a much larger randomized experimental intervention study producing roughly three times the number of lung cancer deaths as in the Rose study. But the intervention group showed more lung cancer deaths than the usual care group! The absolute numbers were small in both studies but there could be no doubt that nothing like the results expected by the epidemiological community had materialized.

The results of the controlled intervention trials illustrate how naive it is to think that experimentation always produces unambiguous results, or frees one from requirements of prior knowledge. One possible explanation for the null effects of intervention on lung cancer, for example, is that the reduced smoking produced by intervention was concentrated among those whose lungs were already in poor health and who were most likely to get lung cancer in any case. (Rose, et al. gave insufficient information for an analysis of the correlation of smoking behavior and lung cancer within the intervention group.) This possibility could have been tested by experiments using blocks more finely selected by health of the subjects.

In retrospect the general lines of the dispute were fairly simple. The statistical community focused on the want of a good scientific argument against a hypothesis given prestige by one of their own; the medical community acted like Bayesians who gave the "constitutional" hypothesis necessary to account for the dose/response data so low a prior that it did not merit serious consideration. Neither side understood what uncontrolled studies could and could not determine about causal relations and the effects of interventions. The statisticians pretended to an understanding of causality and correlation they did not have; the epidemiologists resorted to

informal and often irrelevant criteria, appeals to plausibility, and in the worst case to *ad hominem*.

Fisher's prestige as well as his arguments set the line for statisticians, and the line was that uncontrolled observations cannot distinguish among three cases: smoking causes cancer, something causes smoking and cancer, or something causes smoking and cancer and smoking causes cancer. The most likely candidate for the "something" was genotype. Fisher was wrong about the logic of the matter, but the issue never was satisfactorily clarified, even though some statisticians, notably Brownlee and Burch, tried unsuccessfully to characterize more precisely the connection between probability and causality. While the statisticians didn't get the connection between causality and probability right, the Surgeon General's "epidemiological criteria for causality" were an intellectual disgrace, and the level of argument in defense of the conclusions of the Surgeon General's Report was sometimes more worthy of literary critics than scientists. The real view of the medical community seems to have been that it was just too implausible to suppose that genotype strongly influenced how much one smoked, whether one smoked at all, whether one smoked cigarettes as against a cigar or pipe, whether one was a Mormon or a Seventh day Adventist, and whether one quit smoking or not. After Cornfield's survey the medical and public health communities gave the common cause hypothesis more invective than serious consideration. And, finally, in contrast to Burch, who was an outsider and maverick, leading epidemiologists, such as Lilienfeld, seem simply not to have understood that if the relation between smoking and cancer is confounded by one or more common causes, the effects of abolishing smoking cannot be predicted from the "risk ratios," i.e., from sample conditional probabilities. The subsequent controlled smoking intervention studies gave evidence of how very bad were the expectations based on uncontrolled observation of the relative risks of lung cancer in those who quit smoking compared to those who did not.

9.6 Appendix

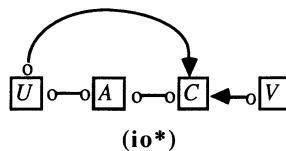


Figure 18

We will prove that the partially oriented inducing path graph (io*) in figure 18, together with the assumptions that U causes A , that there is no common cause of U and C , and that every directed path from U to C contains A , entail that A causes C and that there is a latent common cause of A and C . We assume that A is not a deterministic function of U .

Let $\mathbf{O} = \{A, C, U, V\}$, and G be the directed acyclic graph that generated (io*). The $U \rightarrow C$ edge in (io*) entails that in the inducing path graph of G either $U \rightarrow C$ or $U \leftrightarrow C$. If there is a $U \leftrightarrow C$ edge, then there is a latent common cause of U and C , contrary to our assumption. Hence the inducing path graph contains a $U \rightarrow C$ edge. It follows that in G there is a directed path from U to C . Because every directed path from U to C contains A , there is a directed path from A to C in G . Hence A causes C .

The $U \rightarrow C$ edge in the inducing path graph of G entails that there is an inducing path Z relative to \mathbf{O} that is out of U and into C in G . If Z does not contain a collider then Z is a directed path from U to C and hence it contains A . But then A is a non-collider on Z , and Z is not an inducing path relative to \mathbf{O} , (because Z contains a member of \mathbf{O} , namely A , that is a non-collider on Z) contrary to our assumption. Hence Z contains a collider.

We will now show that no collider on Z is an ancestor of U . Suppose, on the contrary that there is a collider on Z that is an ancestor of U ; let M be the closest such collider on Z to C . No directed path from M to U contains C , because there is a directed path from U to C , and hence no directed path from C to U . There are two cases.

Suppose first that there is no collider between M and C . Then there is a variable Q on Z , such that $Z(Q, C)$ is a directed path from Q to C and $Z(Q, M)$ is a directed path from Q to M . (As in the proofs in Chapter 13, we adopt the convention that on an acyclic path Z containing Q and C , $Z(Q, C)$ represents the subpath of Z between Q and C .) $U \neq M$ because M is a collider on Z and U is not. U does not lie on $Z(Q, C)$ or $Z(Q, M)$ because Z is acyclic. The concatenation of $Z(Q, M)$ and a directed path from M to U contains a directed path from Q to U that does not contain C . $Z(Q, C)$ is a directed path from Q to C that does not contain U . Q is a non-collider on Z , and because Z is an inducing path relative to \mathbf{O} , Q is not in \mathbf{O} . Hence Q is a latent common cause of U and C , contrary to our assumption.

Suppose next that there is a collider between M and C , and N is the collider on Z closest to M and between M and C . Then there is a variable Q on Z , such that $Z(Q, N)$ is a directed path from Q to N and $Z(Q, M)$ is a directed path from Q to M . $U \neq M$ because M is a collider on Z and U is not. U does not lie on $Z(Q, N)$ or $Z(Q, M)$ because Z is acyclic. The concatenation of $Z(Q, M)$

and a directed path from M to U contains a directed path from Q to U that does not contain C . There is a directed path from N to C , and by hypothesis no such directed path contains U . The concatenation of $Z(Q,N)$ and a directed path from N to C contains a directed path from Q to C that does not contain U . Q is a non-collider on Z , and because Z is an inducing path relative to \mathbf{O} , Q is not in \mathbf{O} . Hence Q is a latent common cause of U and C , contrary to our assumption.

It follows that no collider on Z is an ancestor of U .

Let X be the collider on Z closest to U . There is a directed path from X to C . $Z(U,X)$ is a directed path from U to X . The concatenation of $Z(U,X)$ and a directed path from X to C contains a directed path from U to C . By assumption, such a path contains A . A does not lie between U and X on Z , because every vertex between U and X on Z is a non-collider, and if A occurs on Z it is a collider on Z . Hence A lies on every directed path from X to C . Hence there exists a collider on Z that is the source of a directed path to C that contains A . Let R be the collider on Z closest to C such that there is a directed path D from R to C that contains A . There are again two cases.

If there is no collider between R and C on Z , then there is a vertex Q on Z such that $Z(Q,C)$ is a directed path from Q to C and $Z(Q,R)$ is a directed path from Q to R . A does not lie on $Z(Q,C)$ because no vertex on $Z(Q,C)$ is a collider on Z . C does not lie on $D(R,A)$ because the directed graph is acyclic. $C \neq Q$ because Z has an edge into C but not Q . $C \neq R$ because R is a collider on Z and C is not. Hence, C does not lie on $Z(Q,R)$ because Z is acyclic. The concatenation of $Z(Q,R)$ and $D(R,A)$ contains a directed path from Q to A that does not contain C . Q is not a collider on Z , so it not in \mathbf{O} . Hence Q is a latent common cause of A and C .

Suppose next that there is a collider between R and C on Z , and N is the closest such collider to R on Z . Then there is a vertex Q on Z such that $Z(Q,N)$ is a directed path from Q to N and $Z(Q,R)$ is a directed path from Q to R . $Q \neq N$ because by hypothesis there is a path from N to C that does not contain A . A does not lie on $Z(Q,N)$ because no vertex on $Z(Q,N)$ except N is a collider on Z . There is a directed path from N to C , but it does not contain A by hypothesis. Hence the concatenation of $Z(Q,N)$ and a directed path from N to C contains a directed path that does not contain A . C does not lie on $D(R,A)$ because the directed graph is acyclic. $C \neq Q$ because Z has an edge into C but not Q . $C \neq R$ because R is a collider on Z and C is not. Hence, C does not lie on $Z(Q,R)$ because Z is acyclic. The concatenation of $Z(Q,R)$ and $D(R,A)$ contains a directed path from Q to A that does not contain C . Q is not a collider on Z , so it not in \mathbf{O} . Hence Q is a latent common cause of A and C .

Hence in either case, A and C have a latent common cause in G .

Chapter 10

The Structure of the Unobserved

10.1 Introduction

Many theories suppose there are variables that have not been measured but that influence measured variables. In studies in econometrics, psychometrics, sociology and elsewhere the principal aim may be to uncover the causal relations among such "latent" variables. In such cases it is usually assumed that one knows that the measured variables (e.g., responses to questionnaire items) are not themselves causes of unmeasured variables of interest (e.g., attitude), and the measuring instruments are often designed with fairly definite ideas as to which measured items are caused by which unmeasured variables. Survey questionnaires may involve hundreds of items, and the very number of variables is ordinarily an impediment to drawing useful conclusions about structure. Although there are a number of procedures commonly used for such problems, their reliability is doubtful. A common practice, for example, is to form aggregated scales by averaging measures of variables that are held to be proxies for the same unmeasured variable, and then to study the correlations of the scales. The correlations thus obtained have no simple systematic connection with causal relations among the unmeasured variables.

What can a mixture of substantive knowledge about the measured indicators and statistical observations of those indicators reveal about the causal structure of the unobserved variables? And under what assumptions about distributions, linearity, etc.? This chapter begins to address these questions. The procedures for forming scales, or "pure measurement models," that we will describe in this chapter have found empirical application in the study of large psychometric data sets (Callahan and Sorensen 1992).

10.2 An Outline of the Algorithm

Consider the problem of determining the causal structure among a set of unmeasured variables of interest in linear pseudo-indeterministic models, commonly called "structural equation models with latent variables." Assume the distributions are linearly faithful. Structural equation models with latent variables are sometimes presented in two parts: the "measurement model", and the "structural model" (see figure 1). The structural model involves only the causal connections among the latent variables; the remainder is the measurement model. From a mathematical point of view, the distinction marks only a difference in the investigator's interests and access and not any distinction in formal properties. The same principles connecting graphs, probabilities and causes apply to the measurement model as to the structural model. In figure 1 we give an example of a latent variable model in which the measured variables (Q_1-Q_{12}) might be answers to survey questions.

Let \mathbf{T} be a set of latent variables and \mathbf{V} a set of measured variables. We will assume that \mathbf{T} is causally sufficient, although that is clearly not the general case. We let \mathbf{C} denote the set of "nuisance" latent common causes, that is, unobserved common causes, not in \mathbf{T} , of two or more variables in $\mathbf{T} \cup \mathbf{V}$. Call a subgraph of G that contains all of the edges in G except for edges between members of \mathbf{T} a **measurement model** of G .

In actual research the set \mathbf{V} is often chosen so that for each T_i in \mathbf{T} , a subset of \mathbf{V} is intended to measure T_i . In Kohn's (1969) study of class and attitude in America, for example, various questionnaire items were chosen with the intent of measuring the same attitude; factor analysis of the data largely agreed with the clustering one might expect on intuitive grounds. Accordingly, we suppose the investigator can partition \mathbf{V} into $\mathbf{V}(T_i)$, such that for each i the variables in $\mathbf{V}(T_i)$ are direct effects of T_i . We then seek to eliminate those members of $\mathbf{V}(T_i)$ that are impure measures of T_i , either because they are also the effects of some other unmeasured variable in \mathbf{T} , because they are also causes or effects of some other measured variable, or because they share an unmeasured common cause in \mathbf{C} with another measured variable.

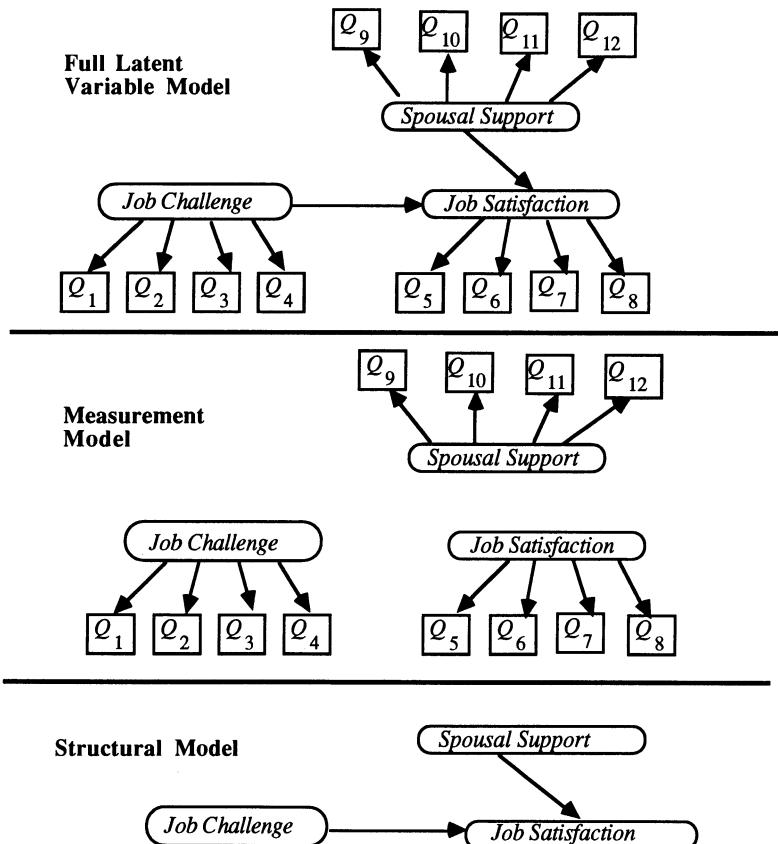


Figure 1

In the class of models we are considering, a measured variable can be an impure measure for four reasons, which are exhaustive:

- If there is a directed edge from some T_i in \mathbf{T} to some V in $\mathbf{V}(T_i)$ but also a trek between V and T_j that does not contain T_i or any member of \mathbf{V} except V then V is **latent-measured impure**.
- If there is a trek between a pair of measured variables V_1, V_2 from the same cluster $\mathbf{V}(T_i)$ that does not contain any member of \mathbf{T} then V_1 and V_2 are **intra-construct impure**.

- (iii) If there is a trek between a pair of measured variables V_1, V_2 from distinct clusters $\mathbf{V}(T_i)$ and $\mathbf{V}(T_j)$ that does not contain any member of \mathbf{T} then we say V_1 and V_2 are **cross-construct impure**.
- (iv) If there is a variable in \mathbf{C} that is the source of a trek between T_i and some member V of $\mathbf{V}(T_i)$ we say V is **common cause impure**.

In figure 2, for example, if $\mathbf{V}(T_1) = \{X_1, X_2, X_3\}$ and $\mathbf{V}(T_2) = \{X_4, X_5, X_6\}$ then X_4 is latent-measured impure, X_1 and X_2 are intra-construct impure, X_2 and X_5 are cross-construct impure, and X_6 is common cause impure. Only X_3 is a pure measure of T_1 .

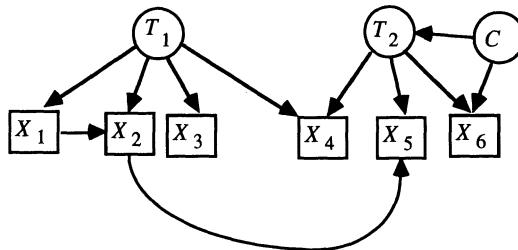


Figure 2

We say that a measurement model is **almost pure** if the only kind of impurities among the measured variables are common cause impurities. An **almost pure latent variable graph** is a directed acyclic graph with an almost pure measurement model. In an almost pure latent variable graph we continue to refer throughout the rest of this chapter to the set of measured variables as \mathbf{V} , a subset of the latent variables as \mathbf{T} , and the "nuisance" latent variables that are common causes of members of \mathbf{T} and \mathbf{V} as \mathbf{C} .

The strategy that we employ has three steps:

- (i) Eliminate measured variables until the variables that remain form the largest almost pure measurement model with at least two indicators for each latent variable.
- (ii) Use vanishing tetrad differences among variables in the measurement model from (i) to determine the zero and first order independence relations among the variables in \mathbf{T} .
- (iii) Use the PC algorithm to construct a pattern from the zero and first order independence relations among the variables in \mathbf{T} .

The next section describes a procedure for identifying the appropriate measured variables. The details are rather intricate; the reader should bear in mind that the procedures have all been

automated, that they work very well in simulation tests, and they all derive from fundamental structural principles. Given the population correlations the inference techniques would be reliable (in large samples) for any conditions under which the Tetrad Representation Theorem holds. The statistical decisions involve a substantial number of joint tests, and no doubt could be improved. We occasionally resort to heuristics for cases in which each latent variable has a large number of measured indicators.

10.3 Finding Almost Pure Measurement Models

If G is the true model over $\mathbf{V} \cup \mathbf{T} \cup \mathbf{C}$ with measurement model G_M , then our task in this section is to find a subset \mathbf{P} of \mathbf{V} (the larger the better) such that the sub-model of G_M on vertex set $\mathbf{P} \cup \mathbf{T} \cup \mathbf{C}$ is an almost pure measurement model, if one exists with at least two indicators per latent variable. Our strategy is to use different types of foursomes of variables to sequentially eliminate impure measures.

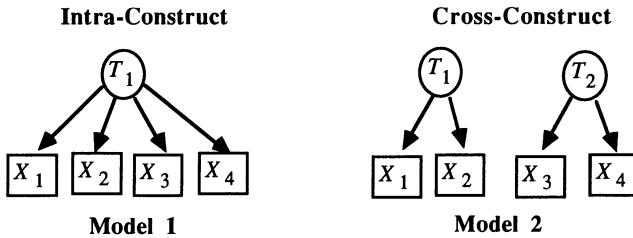


Figure 3

As in figure 3, we call four measured variables an **intra-construct foursome** if all four are in $\mathbf{V}(T_i)$ for some T_i in \mathbf{T} ; otherwise call it a **cross-construct foursome**:

10.3.1 Intra-Construct Foursomes

In this section we discuss what can be learned about the measurement model for T_i from $\mathbf{V}(T_i)$ alone. We take advantage of the following principle, which is a consequence of the Tetrad Representation Theorem.

(P-1) If a directed acyclic graph linearly implies all tetrad differences among the variables in $\mathbf{V}(T_i)$ vanish, then no pair of variables in $\mathbf{V}(T_i)$ is intra-construct impure.

So given a set, $\mathbf{V}(T_i)$, of variables that measure T_i , we seek the largest subset, $\mathbf{P}(T_i)$, of $\mathbf{V}(T_i)$ such that all tetrad differences are judged to vanish among $\mathbf{P}(T_i)$. The number of subsets of $\mathbf{V}(T_i)$ is $2^{|\mathbf{V}(T_i)|}$, so it is not generally feasible to examine each of them. Further, in realistic samples we won't find a sizable subset in which all tetrad differences are judged to vanish. A more feasible strategy is to prune the set iteratively, removing at each stage the variable that improves the performance of the remaining set $\mathbf{P}(T_i)$ on easily computable heuristic criteria derived from principle P-1. In practice, if the set $\mathbf{V}(T_i)$ is large, some small subset of $\mathbf{V}(T_i)$ may by chance do well on these two criteria. For example, if $\mathbf{V}(T_i)$ has 12 variables, then there are 495 subsets of size 4, each of which has only 3 possible vanishing tetrad differences. There are 792 subsets of size 5, but there are 15 possible tetrad differences that must all be judged to vanish among each set instead of 3. Because the larger the size of $\mathbf{P}(T_i)$ the more unlikely it is that all tetrad differences among $\mathbf{P}(T_i)$ will be judged to vanish by chance, and because we might eliminate variables from $\mathbf{P}(T_i)$ later in the process, we want $\mathbf{P}(T_i)$ to be as large as possible. On the other hand, no matter how well a set $\mathbf{P}(T_i)$ does on the first criterion above, some subset of it will do at least as well or better. Thus, in order to avoid always choosing the smallest possible subsets we have to penalize smaller sets.

We use the following simple algorithm. We initialize $\mathbf{P}(T_i)$ to $\mathbf{V}(T_i)$. If the set of tetrad differences among variables in $\mathbf{P}(T_i)$ passes a statistical test, we exit. (We count a set of n tetrad differences as passing a statistical test at a given significance level Sig if each individual tetrad difference passes a statistical test at significance level Sig/n . The details of the statistical tests that we employ on individual tetrad differences are described in Chapter 11.) If the set does not pass a statistical test, we look for a variable to eliminate from $\mathbf{P}(T_i)$. We score each measured variable X in the following way. For each tetrad difference t among variables in $\mathbf{P}(T_i)$ in which X appears we give X credit if t passes a statistical test, and discredit if t fails a statistical test. We then eliminate the variable with the lowest score from $\mathbf{P}(T_i)$. We repeat this process until we arrive at a set $\mathbf{P}(T_i)$ that passes the statistical test, or we run out of variables.

10.3.2 Cross-Construct Foursomes

Having found, for each latent variable T_i , a subset $\mathbf{P}(T_i)$ of $\mathbf{V}(T_i)$ in which no variables are intra-construct impure, we form a subset \mathbf{P} of \mathbf{V} such that

$$\mathbf{P} = \bigcup_{T_i \in \mathbf{T}} \mathbf{P}(T_i).$$

We next eliminate members of \mathbf{P} that are cross-construct impure.

2x2 foursomes involve two measured variables from $\mathbf{P}(T_i)$ and two from $\mathbf{P}(T_j)$, where i and j are distinct. A 2x2 foursome in a pure latent variable model linearly implies exactly one tetrad equation, *regardless of the nature of the causal connection between T_i and T_j in the structural model*. For example, the graph in figure 4 linearly implies the vanishing tetrad difference $\rho_{XY}\rho_{WZ} - \rho_{XW}\rho_{YZ} = 0$. Graphs in which T_j causes T_i and graphs in which T_i and T_j are not causally connected (i.e. there is no trek between them) also linearly imply $\rho_{XY}\rho_{WZ} - \rho_{XW}\rho_{YZ} = 0$.

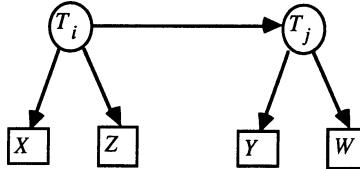


Figure 4

If one variable in $\mathbf{V}(T_i)$ is latent-measured impure because of a trek containing T_j , and one variable in $\mathbf{V}(T_j)$ is latent-measured impure because of a trek containing T_i , then the tetrad differences among the foursome are not linearly implied to vanish by the graph. If T_i and T_j are connected by some trek and a pair of variables in $\mathbf{V}(T_i)$ and $\mathbf{V}(T_j)$ respectively are cross-construct impure then again the tetrad difference is not linearly implied to vanish by the graph. (The case where T_i and T_j are not connected by some trek is considered below.) In figure 5, for example, model (i) implies the tetrad equation $\rho_{XY}\rho_{WZ} = \rho_{XW}\rho_{YZ}$ but models (ii) and (iii) do not.

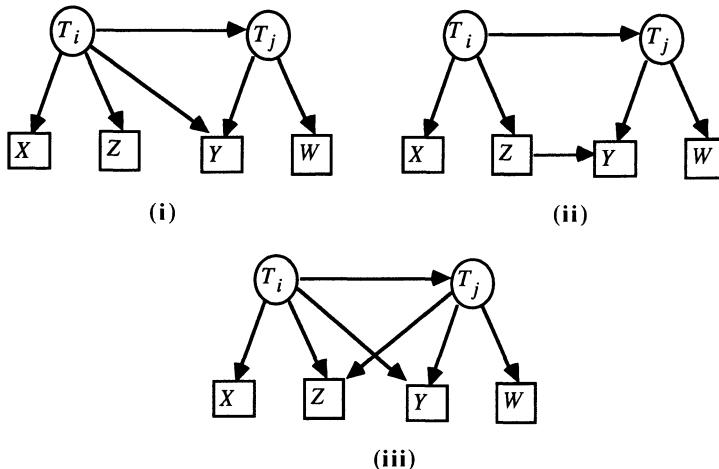


Figure 5

So if we test a 2×2 foursome F_1 and the hypothesis that the appropriate tetrad difference vanishes can be rejected, then we know that in at least one of the four pairs in which there is a measured variable from each construct, both members of the pair are impure. We don't yet know which pair. We can find out by testing other 2×2 foursomes that share variables with F_1 . Suppose the largest subgraph of the true model containing $\mathbf{P}(T_1)$ and $\mathbf{P}(T_2)$ is the graph in figure 6.

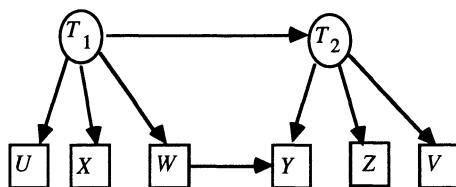


Figure 6

Only 2×2 foursomes involving the pair $\langle W, Y \rangle$ will be recognizably impure. When we test vanishing tetrad differences in the foursome $F_1 = \langle X, W, Y, Z \rangle$, we won't know which of the pairs $\langle W, Z \rangle$, $\langle X, Y \rangle$, $\langle X, Z \rangle$, $\langle W, Y \rangle$ is impure. When we test the foursome $F_2 = \langle X, W, Z, V \rangle$, however, we find that no pair among $\langle X, Z \rangle$, $\langle X, V \rangle$, $\langle W, Z \rangle$, or $\langle W, V \rangle$ is impure. We know therefore that the pairs $\langle X, Z \rangle$ and $\langle W, Z \rangle$ are not impure in F_1 . By testing

the foursome $F_3 = \langle U, X, Y, Z \rangle$ we find that $\langle X, Y \rangle$ is not impure, entailing $\langle W, Y \rangle$ is impure in F_1 . If there are at least two pure indicators within each construct, then we can detect exactly which of the other indicators are impure in this way.

By testing all the 2×2 foursomes in \mathbf{P} , we can in principle eliminate all variables that are cross-construct impure. We cannot yet eliminate all the variables that are latent-measured impure, because if there is only one such variable it is undetectable from 2×2 foursomes.

Foursomes that involve three measured variables from $\mathbf{P}(T_i)$ and one from $\mathbf{P}(T_j)$, where i and j are distinct, are called **3x1 foursomes**. All 3x1 foursomes in a pure measurement model linearly imply all three possible vanishing tetrad differences (see model (i) in figure 7 for example), *no matter what the causal connection between T_i and T_j* . If the variable from $\mathbf{P}(T_j)$ in a 3x1 foursome is impure because it measures both latents (model (ii) in figure 7), then T_i is still a choke point and all three equations are linearly implied. If a variable Z from $\mathbf{P}(T_i)$ is impure because it measures both latents (model (iii) in figure 7), however, then the latent variable model does not linearly imply that the tetrad differences containing the pair $\langle Z, W \rangle$ vanish. This entails that a non-vanishing tetrad differences among the variables in a 3x1 foursome can identify a unique measured variable as latent-measured impure.

Also if T_i and T_j are not trek-connected and a pair of variables V_1 and V_2 in $\mathbf{P}(T_i)$ and $\mathbf{P}(T_j)$ respectively are cross-construct impure, then the correlation between V_1 and V_2 does not vanish, and a tetrad difference among a 3x1 foursome that contains V_1 and V_2 is not linearly implied to vanish; hence the impure member of $\mathbf{P}(T_j)$ will be recognized.

If there are least three variables in $\mathbf{P}(T_i)$ for each i , then when we finish examining all 3x1 foursomes we will have a subset \mathbf{P} of \mathbf{V} such that the sub-model of the true measurement model over \mathbf{P} (which we call G_P) is an almost pure measurement model.

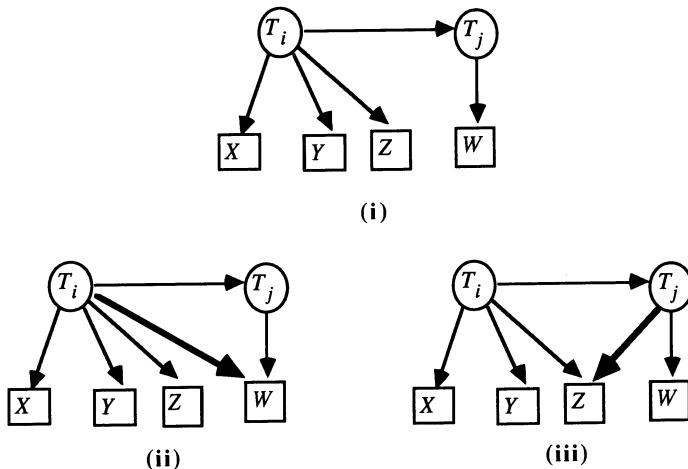


Figure 7

10.4 Facts about the Unobserved Determined by the Observed

In an almost pure latent variable model constraints on the correlation matrix among the *measured* variables determine

- (i) for each pair A, B , of latent variables, whether A, B are uncorrelated,
- (ii) for each triple A, B, C of latent variables, whether A and B are d-separated given $\{C\}$.

Part (i) is obvious: two measured variables are uncorrelated in an almost pure latent variable model if and only if they are effects of distinct unmeasured variables that are not trek connected (i.e. there is no trek between them) and hence are d-separated given the empty set of variables. Part (ii) is less obvious, but in fact certain d-separation facts are determined by vanishing tetrad differences among the measured variables.

Theorem 10.1 is a consequence of the Tetrad Representation Theorem:

Theorem 10.1: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, and each latent variable in T has at least two measured indicators, then latent variables T_1 and T_3 , whose measured indicators include J and L respectively, are d-separated given latent variable T_2 , whose measured indicators include I and K , if and only if G linearly implies $\rho_{JIPLK} = \rho_{JLPKI} = \rho_{JKPIL}$.

For example, in the model in figure 8, the fact that T_1 and T_3 are d-separated given T_2 is entailed by the fact that for all m, n, o , and p between 1 and 3, where o and p are distinct:

$$\rho_{A_m D_n} \rho_{B_o B_p} = \rho_{A_m B_o} \rho_{D_n B_p} = \rho_{A_m B_p} \rho_{D_n B_o}$$

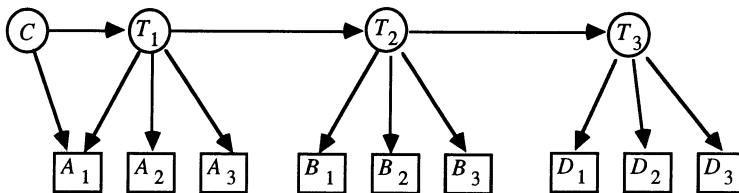
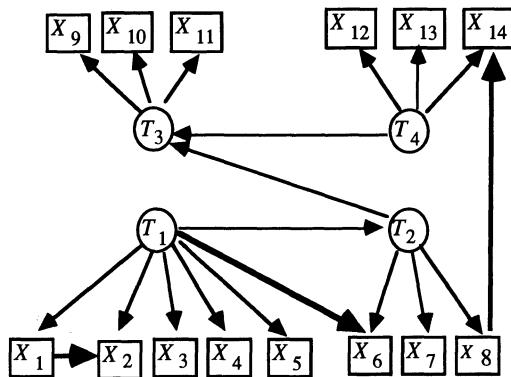


Figure 8

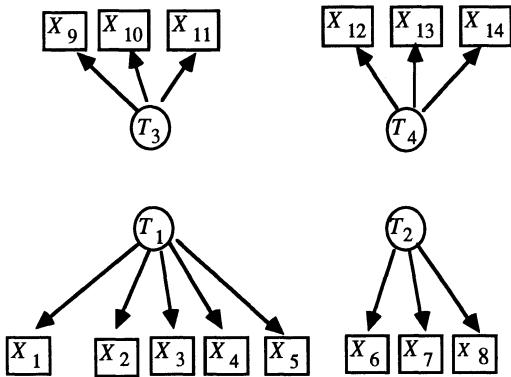
By testing for such vanishing tetrad differences we can test for first order d-separability relations among the unmeasured variables in an almost pure latent variable model. (If A and B are d-separated given D , we call the number of variables in D the **order** of the d-separability relation.) These zero and first order d-separation relations can then be used as input to the PC algorithm, or to some other procedure, to obtain information about the causal structure among the latent variables. In the ideal case, the pattern among the latents that is output will always contain the pattern that would result from applying the PC algorithm directly to d-separation facts among the latents, but it may contain extra edges and fewer orientations.

10.5 Unifying the Pieces

Suppose the true but unknown graph is shown in figure 9.

**Figure 9: True Causal Structure**

We assume that a researcher can accurately cluster the variables in the specified measurement model, e.g., figure 10.

**Figure 10: Specified Measurement Model**

The actual measurement model is then the graph in figure 11:

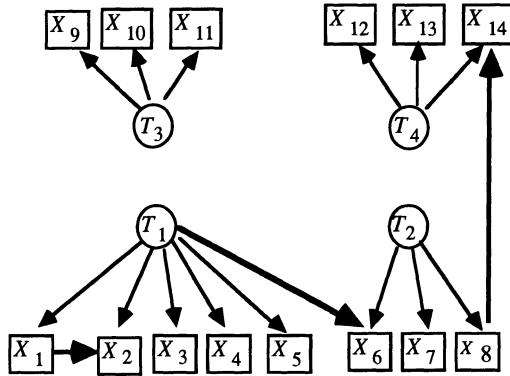


Figure 11: Actual Measurement Model

Figure 12 shows a subset of the variables in G (one that leaves out X_1, X_6 and X_{14}) that do form an almost pure measurement model.

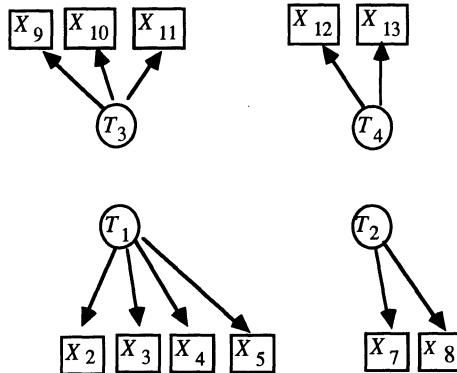


Figure 12: Almost Pure Measurement Model

Assuming the sequence of vanishing tetrad difference tests finds such an almost pure measurement model, a sequence of tests of $1 \times 2 \times 1$ vanishing tetrad difference tests then decides some d-separability facts for the PC or other algorithm through Theorem 10.1. Since in figure 12 there are many $1 \times 2 \times 1$ tetrad tests with measured variables drawn respectively from the clusters for T_1, T_2 and T_3 , the results of the tests must somehow be aggregated. For each $1 \times 2 \times 1$ tetrad difference among variables in $V(T_1), V(T_2)$, and $V(T_3)$ we give credit if the tetrad

difference passes a significance test and discredit if it fails a significance test; if the final score is greater than 0, we judge that T_1 and T_3 are d-separated by T_2 .

With two slight modifications, the PC algorithm can be applied to the zero and first order d-separation relations determined by the vanishing tetrad differences. The first modification is of course that the algorithm never tries to test any d-separation relation of order greater than 1 (i.e. in the loop in step B) of the PC Algorithm the maximum value of n is 1.) The second is that in step D) of the PC algorithm we do not orient edges to avoid cycles.

Without all of the d-separability facts available, the PC algorithm may not find the correct pattern of the graph. It may include extra edges and fail to orient some edges. However, it is possible to recognize from the pattern that some edges are definitely in the graph that generated the pattern, while others may or may not be. We add the following step to the PC algorithm to label with a "?" edges that may or may not be in the graph. Y is a **definite non-collider** on an undirected path U in pattern Π if and only if either $X \dashv\vdash Y \rightarrow Z$, or $X \dashv\vdash Y \leftarrow Z$ are subpaths of U , or X and Z are not adjacent and not $X \rightarrow Y \leftarrow Z$ on U .

- E.) Let \mathbf{P} be the set of all undirected paths in Π between X and Y of length ≥ 2 . If X and Y are adjacent in Π , then mark the edge between X and Y with a "?" unless either
- (i) no paths are in \mathbf{P} , or
 - (ii) every path in \mathbf{P} contains a collider, or
 - (iii) there exists a vertex Z such that Z is a definite non-collider on every path in \mathbf{P} , or
 - (iv) every path in \mathbf{P} contains the same subpath $\langle A, B, C \rangle$.

We refer to the combined procedure as the Multiple Indicator Model Building (MIMBuild) Algorithm.

Theorem 10.2: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , and Π is the output of the MIMBuild Algorithm then:

- A-1) If X and Y are not adjacent in Π , then they are not adjacent in G .
- A-2) If X and Y are adjacent in Π and the edge is not labeled with a "?", then X and Y are adjacent in G .
- O-1) If $X \rightarrow Y$ is in Π , then every trek in G between X and Y is into Y .
- O-2) If $X \rightarrow Y$ is in Π and the edge between X and Y is not labeled with a "?", then $X \rightarrow Y$ is in G .

The algorithm's complexity is bounded by the number of tetrad differences it must test, which in turn is bounded by the number of foursomes of measured variables. If there are n measured variables the number of foursomes is $O(n^4)$. We do not test each possible foursome, however, and the actual complexity depends on the number of latent variables and how many variables measure each latent. If there are m latent variables and s measured variables for each, then the number of foursomes is $O(m^3 \times s^4)$. Since $m \times s = n$, this is $O(n^3 \times s)$.

10.6 Simulation Tests

The procedure we have sketched has been fully automated in the TETRAD II program, with sensible but rather arbitrary weighting principles where required. To test the behavior of the procedure we generated data from the causal graph in figure 13, which has 11 impure indicators.

The distribution for the exogenous variables is standard normal. For each sample, the linear coefficients were chosen randomly between .5 and 1.5.

We conducted 20 trials each at sample sizes of 100, 500, and 2000. We counted errors of commission and errors of omission for detecting uncorrelated latents (0-order d-separation) and for detecting 1st-order d-separation. In each case we counted how many errors the procedure could have made and how many it actually made. We also give the number of samples in which the algorithm identified the d-separations perfectly. The results are shown in Table 1, where the proportions in each case indicate the number of errors of a given kind over all samples divided by the number of possible errors of that kind over all samples.

Extensive simulation tests with a variety of latent topologies for as many as six latent variables, and 60 normally distributed measured variables, show that for a given sample size the reliability of the procedure is determined by the number of indicators of each latent and the proportion of indicators that are confounded. Increased numbers of almost pure indicators make decisions about d-separability more reliable, but increased proportions of confounded variables makes identifying the almost pure indicators more difficult. For large samples with ten indicators per latent the procedure gives good results until more than half of the indicators are confounded.

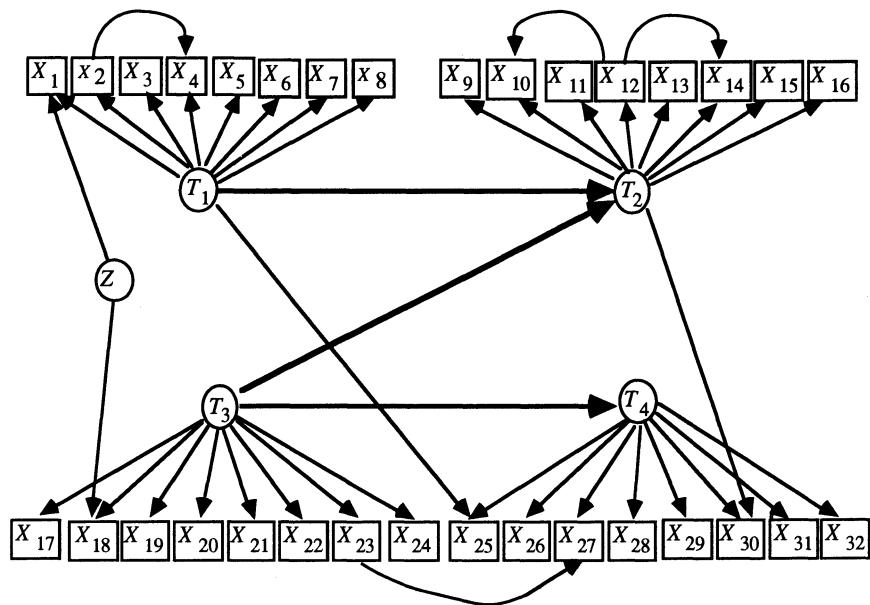


Figure 13 Impure Indicators = $\{X_1, X_2, X_4, X_{10}, X_{12}, X_{14}, X_{18}, X_{23}, X_{25}, X_{27}, X_{30}\}$

Table 1

Sample Size	0-order Commission	0-order Omission	1st-Order Commission	1st-Order Omission	Perfect
100	2.50%	0.00%	3.20%	5.00%	65.00%
500	1.25%	0.00%	0.90%	0.00%	95.00%
2000	0.00%	0.00%	0.00%	0.00%	100.00%

10.7 Conclusion

Alternative strategies are available. One could, for example, purify the measurement sets, and specify a "theoretical model" in which each pair of latent variables is directly correlated. A maximum likelihood estimate of this structure will then give an estimate of the correlation matrix for the latents. The correlation matrix could then be used as input to the PC or FCI algorithms. The strategy has two apparent disadvantages. One is that these estimates typically depend on an assumption of normality. The other is that in preliminary simulation studies with normal variates and using LISREL to estimate the latent correlations, we have found the strategy less reliable than the procedure described in this chapter. Decisions about d-separation facts among latent variables seem to be more reliable if they are founded on a weighted average of a number of decisions about vanishing tetrad differences based on measured correlations than if they are founded on decisions about vanishing partial correlations based on estimated correlations.

The MIMBuild algorithm assumes \mathbf{T} is causally sufficient; an interesting open question is whether there are reliable algorithms that do not make this assumption. In addition, although the algorithm is correct, it is incomplete in a number of distinct ways. There is further orientation information linearly implied by the zero and first order vanishing partial correlations. Further, we do not know whether there is further information about which edges definitely exist (i.e. should not be marked with a "?") that is linearly implied by the vanishing zero and first order partial correlations. Moreover, it is sometimes the case that for each edge labeled with a "?" in the MIMBuild output there exists a pattern compatible with the vanishing zero and first order partial correlations that does not contain that edge, but no pattern compatible with the vanishing zero and first order partial correlations that does not contain two or more of the edges so labeled.

Finally, and most importantly, the strategy we have described is not very informative about latent structures that have multiple causal pathways among variables. An extension of the strategy might be more informative and merits investigation. In addition to tetrad differences, one could test for higher-order constraints on measured correlations (e.g. algebraic combinations of five or more correlations) and use the resulting decisions to determine higher-order d-separation relations among the latent variables. The necessary theory has not been developed.

Chapter 11

Elaborating Linear Theories with Unmeasured Variables¹

11.1 Introduction

In many cases investigators have a causal theory in which they place some confidence, but they are unsure whether the model contains all important causal connections, or they believe it to be incomplete but don't know which dependencies are missing. How can further unknown causal connections be discovered? The same sort of question arises for the output of the PC or FCI algorithms when, for example, two correlated variables are disconnected in the pattern; in that case we may think that some mechanism not represented in the pattern accounts for the dependency, and the pattern needs to be elaborated. In this chapter we consider a special case of the "elaboration problem," confined to linear theories with unmeasured common causes each having one or more measured indicators. The general strategy we develop for addressing the elaboration problem can be adapted to models without latent variables, and also to models for discrete variables. Other strategies than those we consider here are also promising; the Bayesian methods of Cooper and Herskovits, in particular, could be adapted to the elaboration problem.

The problem of elaborating incomplete "structural equation models" has been addressed in at least two commercial computer packages, the LISREL program (Joreskog and Sorbom 1984) and the EQS program (Bentler 1985). We will describe detailed tests of the reliabilities of the automated search procedures in these packages. Generally speaking, we find them to be very unreliable, but not quite useless, and the analysis of why they fail when other methods succeed suggests an important general lesson about computerized search in statistics: in specification search computation matters, and in large search spaces it matters far more than does using tests that would, were computation free, be optimal.

¹This chapter is an abbreviated version of Spirtes, Scheines and Glymour (1990), and is reprinted with the permission of Sage Publications.

We will compare the EQS and LISREL searches with a search procedure based on tests of vanishing tetrad differences. In principle, the collection of tetrad tests is less informative than maximum likelihood tests of an entire model used by the LISREL and EQS searches. In practice, this disadvantage is overwhelmed by the computational advantages of the tetrad procedure. Under some general assumptions, the procedure we describe gives correct (but not necessarily fully informative) answers if correct decisions are made about vanishing tetrad differences in the population. We demonstrate that for many problems the procedure obtains very reliable conclusions from samples of realistic sizes.

11.2 The Procedure

The procedure we will describe is implemented in the TETRAD II program. It takes as input:

- (i) a sample size,
- (ii) a correlation or covariance matrix, and
- (iii) the directed acyclic graph of an initial linear structural equation model.

A number of specifications of internal parameters can also be input. The graph is given to the program simply by specifying a list of paired causes and effects. The algorithm can be divided into two parts, a scoring procedure and a search procedure.

11.2.1 Scoring

The procedure uses the following methodological principles.

Falsification Principle: Other things being equal, prefer models that do not linearly imply constraints that are judged not to hold in the population.

Explanatory Principle: Other things being equal, prefer models that linearly imply constraints that are judged to hold in the population.

Simplicity Principle: Other things being equal, prefer simpler models.

The intuition behind the Explanatory Principle is that an explanation of a constraint based on the causal structure of a model is superior to an explanation that depends upon special values of the free parameters of a model. This intuition has been widely shared in the natural sciences; it was used to argue for the Copernican theory of the solar system, the General Theory of Relativity, and the atomic hypothesis. A more complete discussion of the Explanatory Principle can be found in Glymour et. al. (1987), Scheines (1988), and Glymour (1983). As with vanishing partial correlations, the set of values of linear coefficients associated with the edges of a graph that generate a vanishing tetrad difference not linearly implied by the graph has Lebesgue measure zero.

Unfortunately, the principles can conflict. Suppose, for example, that model M' is a modification of model M , formed by adding an extra edge to M . Suppose further that M' linearly implies fewer constraints that are judged to hold in the population, but also linearly implies fewer constraints that are judged not to hold in the population. Then M' is superior to M with respect to the Falsification Principle, but inferior to M with respect to the Simplicity and Explanatory Principles. The procedure we use introduces a heuristic scoring function that balances these dimensions.²

In order to calculate the *Tetrad-score* we first calculate the **associated probability** $P(t)$ of a vanishing tetrad difference, which is the probability of obtaining a tetrad difference as large or larger than the one actually observed in the sample, under the assumption that the tetrad difference vanishes in the population. Assuming normal variates, Wishart (1928) showed that the variance of the sampling distribution of the vanishing tetrad difference $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK}$ is equal to

$$\frac{D_{12}D_{34}(N+1)}{(N-1)(N-2)} - D$$

where D is the determinant of the population correlation matrix of the four variables I, J, K , and L , D_{12} is the determinant of the two-dimensional upper left-corner submatrix, D_{34} is the determinant of the lower right-corner submatrix and I, J, K , and L , have a joint normal distribution. In calculating $P(t)$ we substitute the sample covariances for the corresponding population covariances in the formula. $P(t)$ is determined by lookup in a chart for the standard normal distribution. An asymptotically distribution free test has been described by Bollen (1990).

² The original TETRAD program (Glymour, Scheines, Spirtes and Kelly, 1987) had no such scoring function. It was left to the user to balance the Explanatory and Falsification principles.

Among any four distinct measured variables I, J, K and L we compute three tetrad differences:

$$\begin{aligned}t_1 &= \rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} \\t_2 &= \rho_{IL}\rho_{JK} - \rho_{IK}\rho_{JL} \\t_3 &= \rho_{IK}\rho_{JL} - \rho_{IJ}\rho_{KL}\end{aligned}$$

and their associated probabilities $P(t_i)$ on the hypothesis that the tetrad difference vanishes. If $P(t_i)$ is larger than the given significance level, the procedure takes the tetrad difference to vanish in the population. If $P(t_i)$ is smaller than the significance level, but the other two tetrad differences have associated probabilities higher than the significance level, then t_i is ignored. Otherwise, if $P(t_i)$ is smaller than the significance level, the tetrad difference is judged not to vanish in the population.

Let **Implied_H** be the set of vanishing tetrads linearly implied by a model M that are judged to hold in the population and **Implied~_H** be the set of vanishing tetrads linearly implied by M that are judged not to hold in the population. Let **Tetrad-score** be the score of model M for a given significance level assigned by the algorithm, and let **weight** be a parameter (whose significance is explained below). Then we define

$$T = \sum_{t \in \text{Implied}_H} P(t) - \sum_{t \in \text{Implied}_{\sim H}} \text{weight} * (1 - P(t))$$

The first term implements the explanatory principle while the second term implements the falsification principle. The simplicity principle is implemented by preferring, among models with identical *Tetrad-scores*, those with fewer free parameters--which amounts to preferring graphs with fewer edges. The weight determines how conflicts between the explanatory and falsification principles are resolved by determining the relative importance of explanation relative to residual reduction.

The scoring function is controlled by two parameters. The *significance level* is used to judge when a given tetrad difference is zero in the population. The *weight* is used to determine the relative importance of the Explanatory and Falsification Principles. The scoring function has several desirable asymptotic properties, but we do not know whether the particular value for *weight* we use is optimal.

11.2.2 Search

The TETRAD II procedure searches a tree of elaborations of an initial model. The search is comparatively fast because there is an easy algorithm for determining the vanishing tetrad differences linearly implied by a graph (using the Tetrad Representation Theorem), because most of the computational work required to evaluate a model can be stored and reused to evaluate elaborations, and because the scoring function is such that if a model can be conclusively eliminated from consideration because of a poor score, so can any elaboration of it.

The search generates each possible one-edge elaboration of the initial model, orders them by the tetrad score, and eliminates any that score poorly. It then repeats this process recursively on each model generated, until no improvements can be made to a model.

The search is guided by a quantity called **T-maxscore**, which for a given model M represents the maximum *Tetrad-score* that could possibly be obtained by any elaboration of M . $T\text{-maxscore}$ is equal to:

$$T\text{-maxscore} = \sum_{t \in \text{Implied}_H} P(t)$$

The use we make of this quantity is justified by the following theorem.

Theorem 11.1. If G is a subgraph of directed acyclic graph G' , than the set of tetrad equations among variables of G that are linearly implied by G' is a subset of those linearly implied by G .

In order to keep the following example small, suppose that there are just 4 edges, e_1, e_2, e_3 , or e_4 which could be added to the initial model. The example illustrates the search procedure in a case where each possible elaboration of the initial model is considered. Node 1 in figure 1 represents the initial model. Each node in the graph represents the model generated by adding the edge next to the node to its parent. For example, node 2 represents the initial model + e_1 ; node 7 represents node 2 + e_4 , which is the initial model + $e_1 + e_4$. We will say that a program **visits** a node when it creates the model M corresponding to the node and then determines whether any elaboration of M has a higher *Tetrad-score* than M . (Note that the algorithm can generate a model M without visiting M if it generates M but does not determine whether any

elaboration of M has a higher *Tetrad-score* than M .) The numbers inside each node indicate the order in which the models are visited. Thus for example, when the algorithm visits node 2, it first generates all possible one edge additions of the initial model + e_1 , and orders them according to their *T-maxscore*. It then first visits the one with the highest *T-maxscore* (in this case, node 3 that represents the initial model + $e_1 + e_2$). Note that the program does not visit the initial model + e_2 (node 10) until after it has visited all elaborations of the initial model + e_1 .

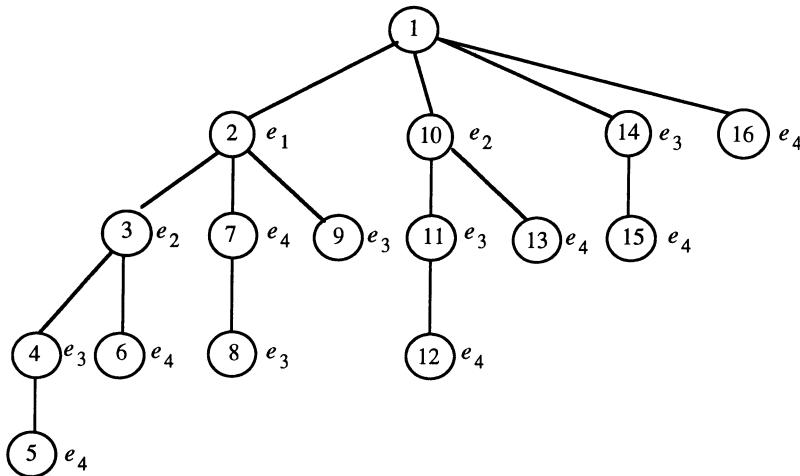
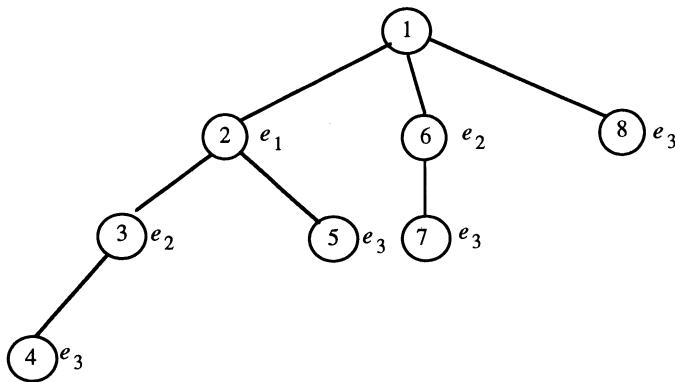


Figure 1

In practice, this kind of complete search could not possibly be carried out in a reasonable amount of time. Fortunately we are able to eliminate many models from consideration without actually visiting them. Addition of edges to a graph may defeat the linear implication of tetrad equations, but in view of Theorem 11.1 will never cause more tetrad equations to be linearly implied by the resulting graph. If the *T-maxscore* of a model M is less than the *Tetrad-score* of some model M' already visited, then we know that neither M nor any elaboration of M has a *Tetrad-score* as high as that of M' . Hence we need never visit M or any of its elaborations. This is illustrated in figure 2. If we find that *T-maxscore* of the initial model + e_4 is lower than the *Tetrad-score* of the initial model + e_1 , we can eliminate from the search all models that contain the edge e_4 .

**Figure 2**

In some cases in the simulation study described later, the procedure described here is too slow to be practical. In those cases the time spent on a search is limited by restricting the depth of search. (We made sure that in every case the depth restriction was large enough that the program had a chance to err by overfitting.) The program adjusts the search to a depth that can be searched in a reasonable amount of time; in many of the Monte Carlo simulation cases no restriction on depth was necessary.³

11.3 The LISREL and EQS Procedures

LISREL VI and EQS are computer programs that perform a variety of functions, such as providing maximum likelihood estimates of the free parameters in a structural equation model. The feature we will consider automatically suggests modifications to underspecified models.

11.3.1 Input and Output

Both programs take as input:

- (i) a sample size,

³ We have also implemented heuristic search procedures that are theoretically less reliable than that described here but are much faster and in practice about equally reliable.

- (ii) a sample covariance matrix,
- (iii) initial estimates of the variances of independent variables,
- (iv) initial estimates of the linear coefficients,
- (v) an initial causal model (specified by fixing at zero the linear coefficient of A in the equation for B if and only if A is not a direct cause of B), in the form of equations (EQS) or a system of matrices (LISREL VI)
- (vi) a list of parameters not to be freed during the course of the search,
- (vii) a significance level, and
- (viii) a bound on the number of iterations in the estimation of parameters.

The output of both programs includes a *single* estimated model that is an elaboration of the initial causal model, various diagnostic information as well as a χ^2 value for the suggested revision, and the associated probability of the χ^2 measure.

11.3.2 Scoring

LISREL VI and EQS provide maximum likelihood estimates of the free parameters in a structural equation model. More precisely, the estimates are chosen to minimize the fitting function

$$F = \log|\Sigma| + \text{tr}(S\Sigma^{-1}) - \log|S| - t$$

where S is the sample covariance matrix, Σ is the predicted covariance matrix, t is the total number of indicators, and if A is a square matrix then $|A|$ is the determinant of A and $\text{tr}(A)$ is the trace of A . In the limit, the parameters that minimize the fitting function F also maximize the likelihood of the covariance matrix for the given causal structure.

After estimating the parameters in a given model, LISREL VI and EQS test the null hypothesis that Σ is of the form implied by the model against the hypothesis that Σ is unconstrained. If the associated probability is greater than the chosen significance level, the null hypothesis is accepted, and the discrepancy is attributed to sample error; if the probability is less than the significance level, the null hypothesis is rejected, and the discrepancy is attributed to the falsity of M . For a "nested" series of models M_1, \dots, M_k in which for all models M_i in the sequence the free parameters of M_i are a subset of the free parameters of M_{i+1} , asymptotically, the *difference* between the χ^2 values of two nested models also has a χ^2 distribution, with degrees of freedom equal to the difference between the degrees of freedom of the two nested models.

11.3.3 The LISREL VI Search⁴

The LISREL VI search is guided by the "modification indices" of the fixed parameters. Each modification index is a function of the derivatives of the fitting function with respect to a given fixed parameter. More precisely, the modification index of a given fixed parameter is defined to be $N/2$ times the ratio between the squared first-order derivative and the second-order derivative (where N is the sample size). Each modification index provides a lower bound on the decrease in the χ^2 obtained if that parameter is freed and all previously estimated parameters are kept at their previously estimated values.⁵ (Note that if the coefficient for variable A in the linear equation for B is fixed at zero, then freeing that coefficient amounts to adding an edge from A to B to the graph of the model.) LISREL VI first makes the starting model the current best model in its search. It then calculates the modification indices for all of the fixed parameters⁶ in the starting model. If LISREL VI estimates that the difference between the χ^2 statistics of M , the current best model, and M' , the model obtained from M by freeing the parameter with the largest modification index, is not significant, then the search ends, and LISREL VI suggests model M' . Otherwise, it makes M' the current best model and repeats the process.

11.3.4 The EQS Search

EQS computes a Lagrange Multiplier statistic, which is asymptotically distributed as χ^2 .⁷ EQS performs univariate Lagrange Multiplier tests to determine the approximate separate effects on the χ^2 statistic of freeing each fixed parameter in a set specified by the user. It frees the parameter that it estimates will result in the largest decrease in the χ^2 value. The program repeats this procedure until it estimates that there are no parameters that will significantly decrease the χ^2 . Unlike LISREL VI, when EQS frees a parameter it does not reestimate the model.⁸

⁴LISREL VII retains the same architecture but with an altered modification index.

⁵ LISREL VI outputs a number of other measures that could be used to suggest modifications to a starting model, but these are not used in the automatic search. See Costner and Herting (1985).

⁶ As long as they are not in the list of parameters not to be freed.

⁷ Since the Lagrange Multiplier statistic, like the modification indices of LISREL VI, estimates the effect on the χ^2 of freeing a parameter, in subsequent sections we will use the term "modification index" to refer to either of these statistics.

⁸ EQS allows the user to specify several different types of searches. We have only described the one used in our Monte Carlo simulation tests.

It should be noted that both LISREL VI and EQS are by now quite complicated programs. An understanding of their flexibility and limitations can only be obtained through experimentation with the programs.

11.4 The Primary Study

Eighty data sets, forty with a sample size of 200 and forty with a sample size of 2,000, were generated by Monte Carlo methods from *each* of nine different structural equation models with latent variables. The models were chosen because they involve the kinds of causal structures that are often thought to arise in social and psychological scientific work. In each case part of the model used to generate the data was omitted and the remainder, together in turn with each of the data sets for that model, was given to the LISREL VI, EQS, and TETRAD II programs. A variety of specification errors are represented in the nine cases. Linear coefficient values used in the true models were generated at random to avoid biasing the tests in favor of one or another of the procedures. In addition, a number of ancillary studies were suggested by the primary studies and bear on the reliability of the three programs.

11.4.1 The Design of Comparative Simulation Studies

To study the reliability of automatic respecification procedure under conditions in which the general structural equation modeling assumptions are met, the following factors should be varied independently.

- (i) the causal structure of the true model;
- (ii) the magnitudes and signs of the parameters of the true model;
- (iii) how the starting model is misspecified;
- (iv) the sample size.

In addition, an ideal study should:

- (i) Compare fully algorithmic procedures, rather than procedures that require judgment on the part of the user. Procedures that require judgment can only adequately be tested by carefully blinding the user to the true model; further, results obtained by one user may not transfer to other users. With fully algorithmic procedures, neither of these problems arises.

- (ii) Examine causal structures that are of a kind postulated in empirical research, or that there are substantive reasons to think occur in real domains.
- (iii) Generate coefficients in the models randomly. Costner and Herting showed that the size of the parameters affects LISREL's performance. Further, the reliability of TETRAD II depends on whether vanishing tetrads hold in a sample because of the particular numerical values of the coefficients rather than because of the causal structure, and it is important not to bias the study either for or against this possibility.
- (iv) Ensure insofar as possible that all programs compared must search the same space of alternative models.

11.4.2 Study Design

11.4.2.1 Selection of Causal Structures

The nine causal structures studied are illustrated in figures 3, 4 and 5. For simplicity of depiction we have omitted uncorrelated error terms in the figures, but such terms were included in the linear models. The heavier directed or undirected lines in each figure represent relationships that were included in the model used to generate simulated data, but were omitted from the models given to the three programs; i.e., they represent the dependencies that the programs were to attempt to recover. The starting models are shown in figure 6. The models studied include a one factor model with five measured variables, seven multiple indicator models each with eight measured variables and two latent variables, and one multiple indicator model with three latent variables and eight measured variables.

One factor models commonly arise in psychometric and personality studies (see Kohn 1969); two latent factor models are common in longitudinal studies in which the same measures are taken at different times (see McPherson et. al. 1977), and also arise in psychometric studies; the triangular arrangement of latent variables is a typical geometry (see Wheaton et. al. 1977).

The set of alternative structures determines the search space. Each program was forced to search the same space of alternative elaborations of the initial model, and the set of alternatives was chosen to be as large as possible consistent with that requirement.

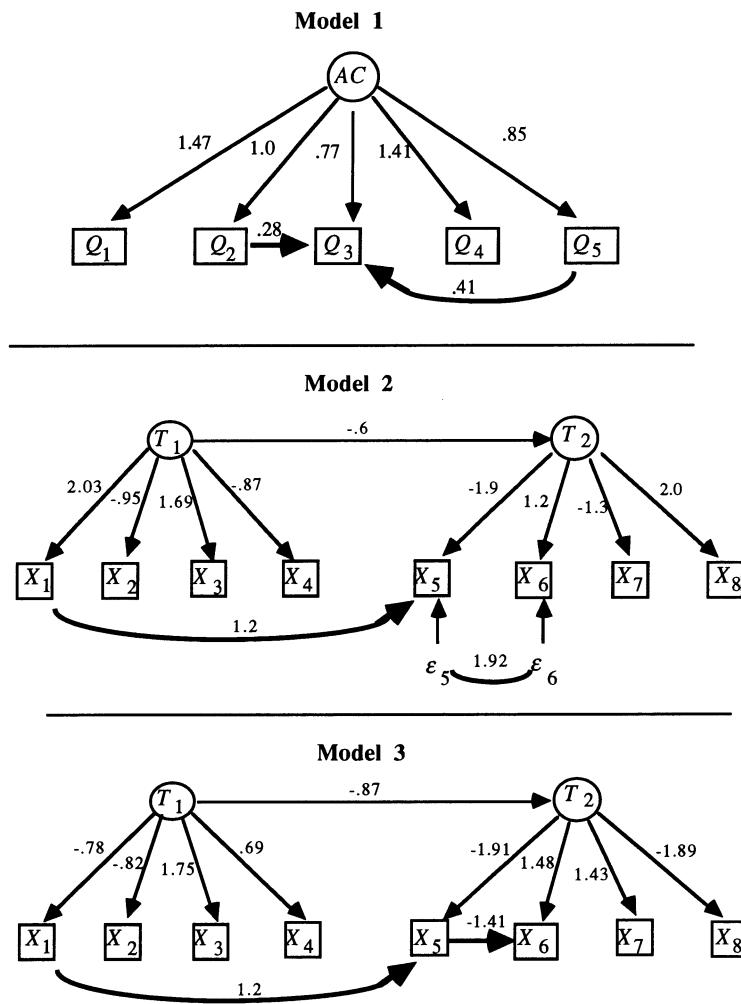
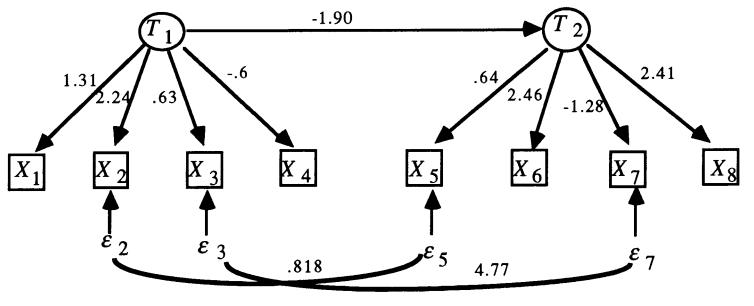
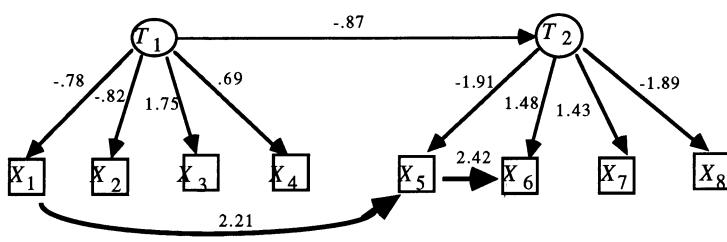
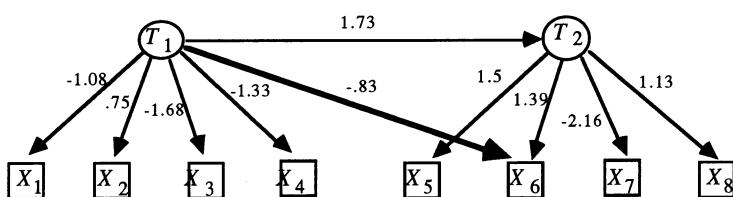
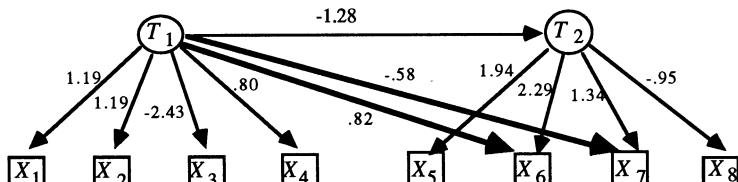
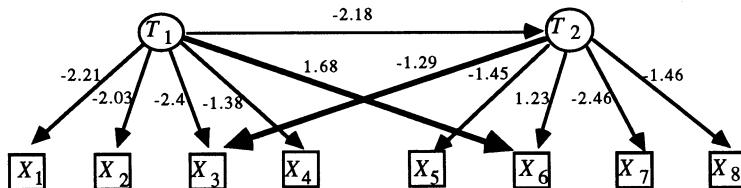
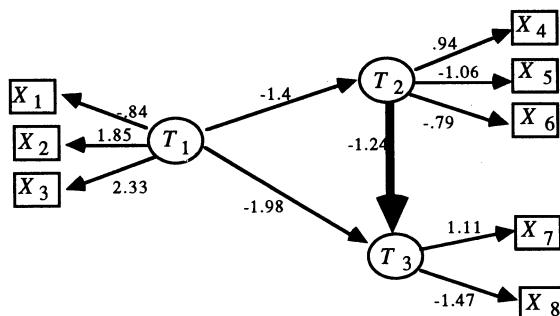
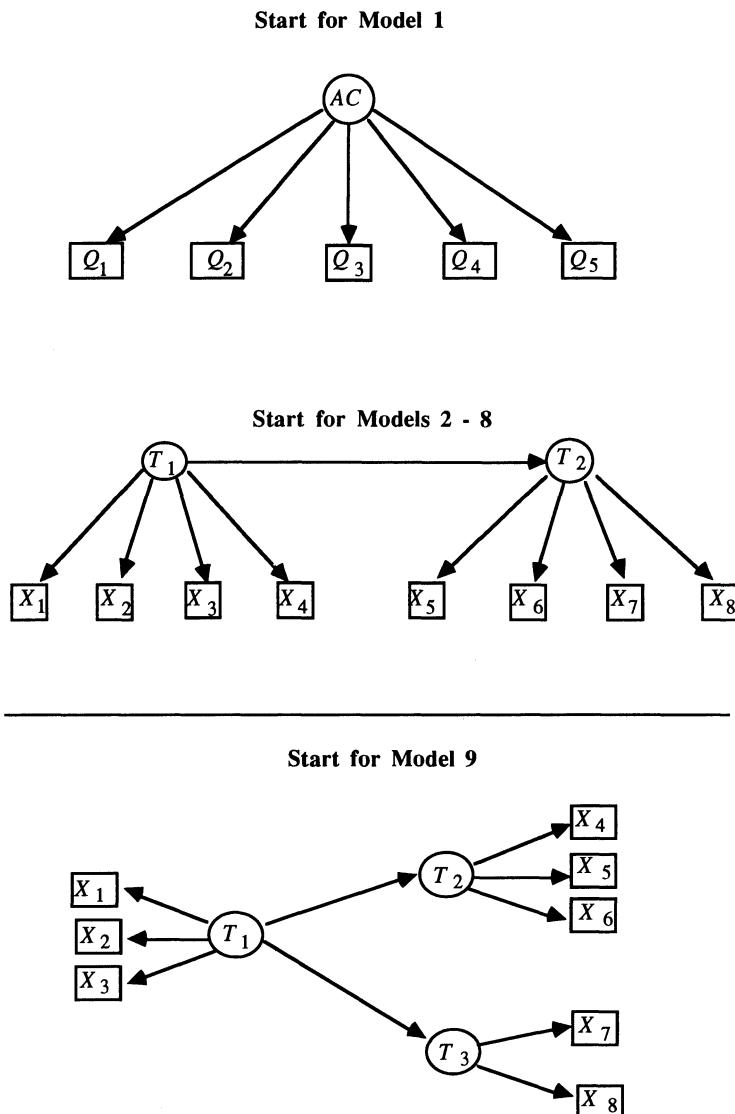


Figure 3

Model 4**Model 5****Model 6****Figure 4**

Model 7**Model 8****Model 9****Figure 5**

**Figure 6**

11.4.2.2 Selection of Connections to be Recovered

The connections to be recovered include:

- (i) Directed edges from latent variables to latent variables; relations of this kind are often the principal point of empirical research. See Maruyama and McGarvey (1980) for an example.
- (ii) Edges from latent variables to measured variables; connections of this kind may arise when measures are impure, and in other contexts. See Costner and Schoenberg (1973) for an example.
- (iii) Correlated errors between measured variables; relationships of this kind are perhaps the most frequent form of respecification.
- (iv) Directed edges from measured variables to measured variables. Such relations cannot obtain, for example, between social indices, but they may very well obtain between responses to survey or psychometric instruments (see Campbell et. al. 1966), and of course between measured variables such as interest rates and housing sales.

We have not included cases that we know beforehand cannot be recovered by one or another of the programs. Details are given in a later section.

11.4.2.3 Selection of Starting Models

Only three starting models were used in the nine cases. The starting models are, in causal modeling terms, pure factor models or pure multiple indicator models. In graph theoretic terms they are *trees*.

11.4.2.4 Selection of Parameters

In the figures showing the true models the numbers next to directed edges represent the values given to the associated linear coefficients. The numbers next to undirected lines represent the values of specified covariances. In all cases, save for models 1 and 5, the coefficients were chosen by random selection from a uniform distribution between .5 and 2.5. The value obtained was then randomly given a sign, positive or negative.

In model 1, all linear coefficients were made positive. The values of the causal connections between indicators were specified non-randomly. The case was constructed to simulate a psychometric or other study in which the loadings on the latent factor are known to be positive, and in which the direct interactions between measured variables are comparatively small.

Model 5 was chosen to provide a comparison with model 3 in which the coefficients of the measured-measured edges were deliberately chosen to be large relative to those in model 3.

11.4.2.5 Generation of Data

For each of the nine cases, twenty data sets with sample size 200 and twenty data sets with sample 2,000 were generated by Monte Carlo simulation methods.

Pseudo-random samples were generated by the method described in Chapter 5. In order to optimize the performance of each of the programs, we assumed that all of the exogenous variables had a standard normal distribution. This assumption made it possible to fix a value for each exogenous variable for each unit in the population by pseudo random sampling from a standard normal distribution. Correlated errors were obtained in the simulation by introducing a *new* exogenous common cause of the variables associated with the error terms.

11.4.2.6 Data Conditioning

The entire study we discuss here was performed twice. In the original study, we gave LISREL VI and EQS positive starting values for all parameters. If either program had difficulty estimating the starting model, we reran the case with the initial values set to the correct sign.

LISREL and EQS employ iterative procedures to estimate the free parameters of a model. These procedures are sensitive to "poorly conditioned" variables and will not perform optimally unless the data are transformed. For example, it is a rule of thumb with these procedures that no two variances should vary by more than an order of magnitude in the measured variables. After generating data in the way we describe above, a small but significant percentage of our covariance matrices were ill conditioned in this way.

To check the possibility that the low reliability we obtained in the first study for the LISREL VI and EQS procedures was due to "ill-conditioned" data, the entire study was repeated. Sample

covariances were transformed into sample correlations by dividing each cell $[I,J]$ in the covariance matrix by $s_I s_J$, where s_I is the *sample* standard deviation of I . To avoid sample variances of widely varying magnitudes, we transformed each cell $[I,J]$ in the sample covariance matrix by dividing it by $\sigma_I \sigma_J$ where σ_I is the *population* standard deviation of I ⁹. We call the result of this transformation the *pseudocorrelation* matrix. The transformation makes all of the variances of the measured variables almost equal, without using a data-dependent transformation. Of course in empirical research, this transformation could not be performed, since the population parameters would not be known.

In practice, we found that conditioning the data and giving the population parameters as starting values did little to change the performance of LISREL VI or EQS. The performance of the TETRAD II procedure was essentially the same in both cases. Conditioning the data improved LISREL VI's reliability very slightly for small samples, and degraded it slightly for large samples.

11.4.2.7 Starting Values for the Parameters

We selected the linear coefficients for our models randomly, allowing some to be negative and some to be positive. Models with negative parameters actually represent a harder case for the TETRAD procedures. If a model implies a vanishing tetrad difference then the signs of its parameters make no difference. If a model does not imply that a tetrad difference vanishes, however, but instead implies that the tetrad difference is equal to the sum of two or more terms, then it is possible, if not all of the model's parameters are positive, that these terms sum to zero. Thus, in data generated by a model with negative parameters, we are more likely to observe vanishing tetrad differences that are *not* linearly implied by the model.

The iterative estimation procedures for LISREL and EQS begin with a vector of parameters θ . They update this vector until the likelihood function converges to a local maximum. Inevitably, the iterative procedures are sensitive to starting values. Given the same model and data, but two different starting vectors θ^i and θ^j , the procedures might converge for one but not for the other. This is especially true when the parameters are of mixed signs. To give LISREL and EQS the best chance possible in the second study, we set the starting values of each parameter to its actual value whenever possible. For the linear coefficients that correspond to edges in the

⁹ We are indebted to Peter Bentler for suggesting this transformation.

generating model left out of the starting model, we assigned a starting value of 0. For all other parameters, however, we started LISREL and EQS with the *exact value in the population*.¹⁰

11.4.2.8 Significance Levels

EQS and LISREL VI continue to free parameters as long as the associated probability of the decrease in χ^2 exceeds the user-specified significance level. For both LISREL and EQS, we set the significance level to .01. (This is the default value for LISREL; the default value for EQS is .05.) The lower the significance level, the fewer the parameters that each program tends to set free. Since both LISREL and EQS both tend to overfit even at .01, we did not attempt to set the significance level any higher. (It may appear in our results that LISREL VI and EQS both underfit more than they overfit, but almost all of the "underfitting" was due to aborted searches that did not employ the normal stopping criterion.)

11.4.2.9 Number of Iterations

The default number of maximum iterations for estimating parameters for LISREL VI on a personal computer is 250. We set the number of maximum iterations to 250 for both our LISREL VI and EQS tests.

11.4.2.10 Specifying Starting Models in LISREL VI

LISREL VI, like previous editions of the program, requires the user to put variables into distinct matrices according to whether they are exogenous, endogenous but unmeasured, measured but dependent on exogenous latent, measured but dependent on endogenous latent, and so forth. Variables in certain of these categories cannot have effects on variables in other categories. When formulated as recommended in the LISREL manual, LISREL VI would be in principle unable to detect many of the effects considered in this study. However, these restrictions can in most cases be overcome by a system of substitutions of phantom variables in

¹⁰We did not provide LISREL or EQS with the values of the parameters in the original models that generated our covariance matrices because the input to LISREL and EQS was a pseudocorrelation matrix, not the original covariance matrix. We therefore provided the programs with the population parameters of transformed models that would generate the pseudo correlation matrices. The detailed transformations are given in Spirtes (1990).

which measured variables are actually represented as endogenous latent variables.¹¹ In the current study, we were not able to get LISREL VI to accept changing ξ variables, which are exogenous and latent, to η variables, which are endogenous and latent. This had the unfortunate effect that LISREL would not consider adding any edges into T_1 (represented by the ξ variable). To ensure a comparable search problem, we restricted TETRAD II and EQS in the same way.

11.4.2.11 Implementation

The LISREL VI runs were performed with the personal computer version of the program, run on a Compaq 386 computer with a math coprocessor. EQS runs were performed on an IBM XT clone with a math coprocessor. All TETRAD II runs were performed on Sun 3/50 workstations. For TETRAD II, which also runs on IBM clones, the processing time for the Compaq 386 and the Sun 3/50 are roughly the same.

11.4.2.12 Specification of TETRAD II Parameters

TETRAD II requires that the user set a value of the weight parameter, a value for the significance level used in the test for vanishing tetrad differences, and a value for a percentage parameter that bounds the search. In all cases we set the significance level at 0.05. At sample size 2000, we set the weight to .1 and the percentage to 0.95.

At smaller sample size the estimates of the population covariances are less reliable, and more tetrad differences are incorrectly judged to vanish in the population. This makes judgments about the Explanatory Principle less reliable. For this reason, at sample size 200, we set the weight to 1, in order to place greater importance upon the Falsification Principle. Less reliable judgments about the Explanatory Principle also make lowering the percentage for small sample sizes helpful. At sample size 200, we set the percentage to 0.90. We do not know if these parameter settings are optimal.

¹¹ For LISREL IV, the details of this procedure are described in *Discovering Causal Structure*. The same procedure works for LISREL VI with the exception of the Beta matrix. See Joreskog and Sorbom (1984).

11.5 Results

For each data set and initial model, TETRAD II produces a set of best alternative elaborations. In some cases that set consists of a single model; typically it consists of two or three alternatives. EQS and LISREL VI, when run in their automatic search mode, produce as output a single model elaborating the initial model. The information provided by each program is scored "correct" when the output contains the true model. But it is important to see how the various programs err when their output is not correct, and we have provided a more detailed classification of various kinds of error. We have classified the output of TETRAD II as follows (where a model is in TETRAD's top group if and only if it is tied for the highest *Tetrad score*, and no model with the same *Tetrad-score* has fewer edges):

Correct--the true model is in TETRAD's top group.

Width--the average number of alternatives in TETRAD's top group.

Errors:

Overfit--TETRAD's top group does not contain the true model but contains a model that is an elaboration of the true model.

Underfit--TETRAD's top group does not contain the true model but does contain a model of which the true model is an elaboration.

Other--none of the previous categories apply to the output.

We have scored the output of the LISREL VI and EQS programs as follows:

Correct--the true model is recommended by the program.

Errors:

In TETRAD's Top Group--the recommended model is not correct, but is among the best alternatives suggested by the TETRAD II program for the same data.

Overfit--the recommended model is an elaboration of the true model.

Underfit--the true model is an elaboration of the recommended model.

Right Variable Pairs--the recommended model is not in any of the previous categories, but it does connect the same pairs of variables as were connected in the omitted parts of the true model.

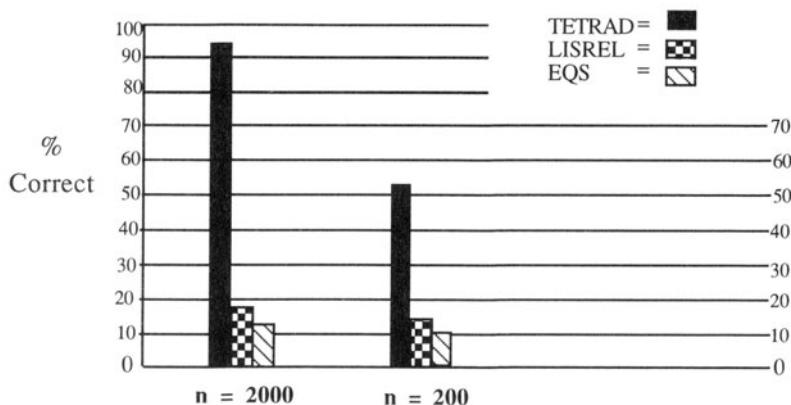
Other--none of the previous categories apply to the output.

In most cases no estimation problems occurred for either LISREL VI or EQS. In a number of data sets for cases 3 and 5, LISREL VI and EQS either issued warnings about estimation problems or aborted the search due to computational problems. Since our input files were built to minimize convergence problems, we ignored such warnings in our tabulation of the results. If either program recommended freeing a parameter, we counted that parameter as freed regardless of what warnings or estimation problems occurred before or after freeing it. If either program failed to recommend freeing any parameters because of estimation problems in the starting model, we counted it as an underfit. The results are shown in the next table and figure.

TABLE 1: Case by Case Width of Set Suggestions

Width, n=2000									
Case	1	2	3	4	5	6	7	8	9
LISREL VI	1	1	1	1	1	1	1	1	1
EQS	1	1	1	1	1	1	1	1	1
TETRAD	4	2.1	2	1	1.1	3	7.1	11.3	2.9

Width, n=200									
Case	1	2	3	4	5	6	7	8	9
LISREL VI	1	1	1	1	1	1	1	1	1
EQS	1	1	1	1	1	1	1	1	1
TETRAD	1.9	3.5	1.5	1	1	3.2	5.9	8.4	3



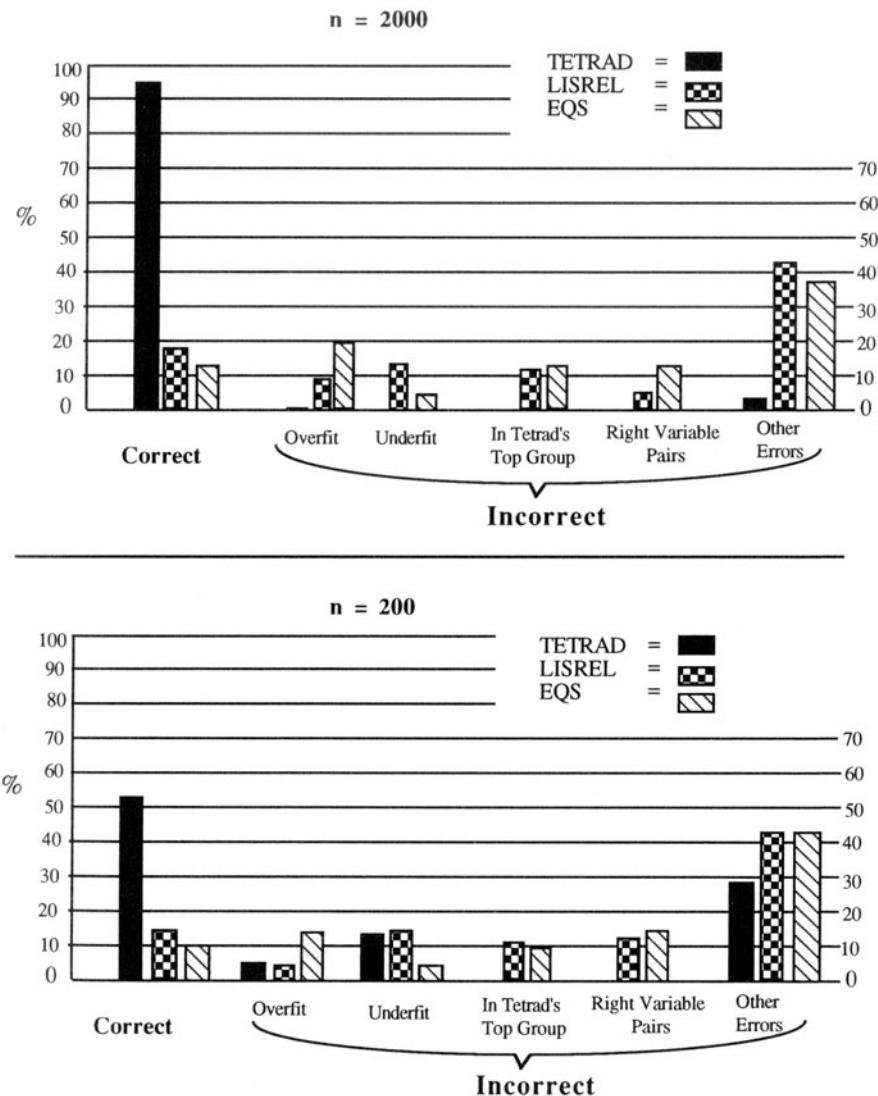


Figure 8

For a sample size of 2000, TETRAD II's set included the correct respecification in 95% of the cases. LISREL VI found the right model 18.8% of the time and EQS 13.3%. For a sample size of 200, TETRAD II's set included the correct respecification 52.2% of the time, while LISREL VI corrected the misspecification 15.0% of the time, and EQS corrected the misspecification 10.0 % of the time. A more detailed characterization of the errors is given in figure 8.

11.6 Reliability and Informativeness

There are two criteria by which the suggestions of each of these programs can be judged. The first is reliability. Let the **reliability** of a program be defined as the probability that its set of suggested models includes the correct one. In these cases, the TETRAD search procedures are clearly more reliable than either LISREL VI and EQS. One can achieve higher reliability simply by increasing the number of guesses. A program that outputs the top million models might be quite reliable, but its suggestions would be uninformative. Thus we call the second criterion boldness. Let the **boldness** of a program's suggestions be the reciprocal of the number of models suggested. On this measure, our procedure does worse than LISREL VI or EQS in seven of the nine cases considered.

Since neither our procedure nor the modification index procedures dominate on both of these criteria, it is natural to ask whether the greater reliability of the former is due simply to reduced boldness. This question can be interpreted in at least two ways:

- (i) If TETRAD II were to increase its boldness to match LISREL VI and EQS, i.e., if it were to output a single model, would it be more or less reliable than LISREL VI or EQS?
- (ii) If LISREL VI or EQS were to decrease their boldness to match TETRAD II, i.e., were they to output a set of models as large as does TETRAD II, would they be more or less reliable than TETRAD II?

If we have no reason to believe that any one model in the TETRAD II output is more likely than any other to be correct, we could simply choose a model at random. We can calculate the **expected single model reliability** of our procedure in the following way. We assume that when TETRAD II outputs a list of n models for a given covariance matrix, the probability of selecting any particular one of the models as the best guess is $1/n$. So instead of counting a list of length n that contains the correct model as a single correct answer, we would count it as $1/n$

correct answers.¹² Then simply divide the expected number of correct answers by the number of trial runs.

Were TETRAD II to be as bold as LISREL VI or EQS, its single model reliability at sample size 2000 would drop from 95% to about 42.3%. On our data, LISREL VI has a reliability of 18.8% and EQS has a reliability of 13.3%. At sample size 200 the TETRAD II single model reliability is 30.2% LISREL has a reliability of 15.0% for sample size 200 and EQS 10.0%. In a more realistic setting one might have substantive reasons to prefer one model over another. If substantive knowledge is worth anything, and we use it to select a single model M , then M is more likely to be true than a model selected at random from TETRAD II's set of suggested models. Thus, in a sense the numbers given in the paragraph above are worst case.

An alternative strategy is to cut down the size of the set before one picks a model. We can often eliminate some of the TETRAD II suggestions by running them through EQS or LISREL VI and discarding those that were not tied for the highest associated probability. There is little effect. We raise the (worst case) single model reliability of TETRAD II at sample size 2000 from 42.3% to about 46%, and at sample size 200 from 30.2 to approximately 32%.

There are a number of good reasons to want a list of equally good suggestions rather than a single guess. All have to do with the reliability and informativeness of the output.

First, it is important for the user of a program to have a good idea of how reliable the output of a program is. At sample size 2000, in the range of cases that we considered, the reliability of the TETRAD II output was very stable, ranging from a low of 90% to a high of 100%. For reasons explained below, the single model output by LISREL VI and EQS is at best in effect a random selection from a list of models that contains all of the models whose associated probabilities are equal to that of the true model (and possibly others of lower associated probabilities as well). Unfortunately, the size of the list from which the suggested model is randomly selected varies a great deal depending on the structure of the model, and is not known to the user. Thus, even ignoring the cases where LISREL VI had substantial computational difficulties, the reliability of LISREL VI's output at sample size 2000 ranged from 0 out of 20 to 11 out of 20. So it is rather difficult for a user of LISREL VI or EQS to know how much confidence to have in the suggested models.

¹²To simplify the calculations, we assumed that the length of the lists output by TETRAD II for all of the covariance matrices generated by a single model was in each case equal to the average length of the lists. This is a fairly good approximation in most cases.

Second, more than one model in a suggested set might lead to the same conclusion. For example, many of the models suggested by TETRAD II might overlap, i.e., they might agree on a substantial number of the causal connections. If one's research concerns are located within those parts of the models that agree, then choosing a single model is not necessary. In this case one need not sacrifice reliability by increasing boldness, because all competitors agree.

Finally, having a well-defined list of plausible alternatives is more useful than a single less reliable suggestion for guiding further research. In designing experiments and in gathering more data it is useful to know exactly what competing models have to be eliminated in order to establish a conclusive result. For example, consider case 3. The correct model contains edges from X_1 to X_5 and X_5 to X_6 . TETRAD II suggests the correct model, as well as a model containing edges from X_1 to X_5 and X_1 to X_6 . An experiment which varied X_1 , and examined the effect on X_6 would not distinguish between these two alternatives (since both predict that varying X_1 would cause X_6 to change), but an experiment which varied X_5 and examined the effect on X_6 would distinguish between these alternatives. Only by knowing the plausible alternatives can we decide which of these experiments is more useful.

If LISREL VI or EQS were to output a set of models as large as does our procedure, would they be as reliable? The answer depends upon how the rest of the models in the set were chosen. In many cases LISREL VI and EQS find several parameters tied, or almost tied, for the highest modification index. Currently both programs select one, and only one, of these parameters to free, on the basis of an arbitrary ordering of parameters. For example, if after evaluating the initial model it found that $X_3 \rightarrow X_5$ and $X_3 \text{ C } X_5$ ¹³ were tied for the highest modification indices, LISREL VI or EQS would choose one of them, (say $X_3 \rightarrow X_5$) and continue until the search found no more parameters to free. Then they would suggest the single model that had the highest associated probability. If LISREL VI or EQS searched all branches corresponding to tied modification indices, instead of arbitrarily choosing one, their reliability would undoubtedly increase substantially. For example, after freeing $X_3 \rightarrow X_5$ and then freeing parameters until no more should be freed, LISREL VI or EQS could return to the initial model, free $X_3 \text{ C } X_5$, and again continue freeing parameters until no more should be freed. They could then suggest all of the models tied for the highest associated probability. This is essentially the search strategy followed by the TETRAD II program.

If the LISREL VI search were expanded in this way on case 1 at sample size 2000, it would increase the number of correct outputs from 3 to 16 out of 20. In other cases, this strategy

¹³ The expression " $X \text{ C } Y$ " means that the error terms for X and Y are correlated, or, equivalently, that there is an additional, common cause of X and Y .

would not improve the performance of LISREL VI or EQS much at all. For example, in case 5 at sample size 2000, LISREL VI was incorrect on every sample in part because of a variety of convergence and computational problems, while TETRAD II was correct in every case. In case 4 at sample size 2000, LISREL VI missed the correct answer on nine samples (while TETRAD II missed the correct answer on only two samples) for reasons having nothing to do with the method of breaking ties.

LISREL VI and EQS would pay a substantial price for expanding their searches; their processing time would increase dramatically. A branching procedure that retained three alternatives at each stage and which stopped on all branches after freeing two parameters in the initial model, would increase the time required by about a factor of 7. In general, the time required for a branching search increases exponentially as the number of alternatives considered at each stage. Could such a search be run in a reasonable amount of time? Without a math coprocessor, a typical LISREL VI run on a Compaq 386 took roughly 20 minutes; with a math coprocessor it took about 4 minutes. EQS runs were done on a LEADING EDGE (an IBM XT clone that is considerably slower than the COMPAQ 386) with a math coprocessor and the average EQS run was about 5 minutes. This suggests that a branching strategy is possible for LISREL VI even for medium-sized models only on relatively fast machines; a branching search is practical on slower machines for the faster, but less reliable EQS search.

11.7 Using LISREL and EQS as Adjuncts to Search

There are two ways in which the sort of search TETRAD II illustrates can profitably be used in conjunction with LISREL VI or EQS. A procedure such as ours can be used to generate a list of alternative revisions of an initial model, which can then be estimated by LISREL or EQS, discarding those alternatives that have very low, or comparatively low associated probabilities.¹⁴ We found that in only three cases could the associated probabilities distinguish among models suggested by TETRAD II. In case 6, one of the three models suggested by TETRAD II had a lower associated probability than the other two. In case 7, one of the six models suggested by TETRAD II had a lower associated probability than the other five. The largest reduction in TETRAD II's suggestions came in case 8, where 8 of the 12 models suggested by TETRAD II had associated probabilities lower than the top four. These results

¹⁴ TETRAD II will, on request, automatically generate EQS input files for all models that it suggests.

were obtained when LISREL VI was given the correct starting values for all of the edges in the true model, and a starting value of zero for edges not in the true model; in previous tests when LISREL VI was not given the true parameters as initial values, it often suffered convergence problems.

It is also instructive to run the both the automatic searches of TETRAD II and LISREL VI or EQS together. When LISREL VI and TETRAD II agree (that is when the model suggested by LISREL VI is in TETRAD II's top group) both programs are correct a higher percentage of times than their respective averages; conversely when they disagree, both programs are wrong a higher percentage of times than their average. The same holds true of EQS when used in conjunction with TETRAD II. Indeed, at sample size 2000, neither EQS nor LISREL VI was *ever* correct when it disagreed with TETRAD II. In contrast, at sample size 2000 LISREL VI was correct 61.8% of the time when it agreed with TETRAD II, and EQS was correct 53.3% of the time when it agreed with TETRAD II. Again, at sample size 2000, TETRAD II was *always* correct when it agreed with either LISREL VI or EQS. At sample size 200, while TETRAD II was correct on average 52.2% of the time, when it agreed with LISREL VI it was correct 75.7% of the time, and when it agreed with EQS it was correct 75.0% of the time. These results are summarized below:

Sample size 2000:

$P(\text{TETRAD correct})$	95.0
$P(\text{LISREL VI correct})$	18.8
$P(\text{EQS correct})$	13.3
$P(\text{TETRAD correct} \mid \text{LISREL VI agree})$	100.0
$P(\text{TETRAD correct} \mid \text{LISREL VI disagree})$	92.1
$P(\text{TETRAD correct} \mid \text{EQS agree})$	100.0
$P(\text{TETRAD correct} \mid \text{EQS disagree})$	92.6
$P(\text{LISREL VI correct} \mid \text{TETRAD II agree})$	61.8
$P(\text{LISREL VI correct} \mid \text{TETRAD II disagree})$	0.0
$P(\text{EQS correct} \mid \text{TETRAD II agree})$	53.3
$P(\text{EQS correct} \mid \text{TETRAD II disagree})$	0.0

Sample size 200:

$P(\text{TETRAD correct})$	52.2
$P(\text{LISREL VI correct})$	15.0
$P(\text{EQS correct})$	10.0
$P(\text{TETRAD correct} \mid \text{LISREL VI agree})$	75.7
$P(\text{TETRAD correct} \mid \text{LISREL VI disagree})$	46.9
$P(\text{TETRAD correct} \mid \text{EQS agree})$	75.0
$P(\text{TETRAD correct} \mid \text{EQS disagree})$	47.2
$P(\text{LISREL VI correct} \mid \text{TETRAD II agree})$	39.4
$P(\text{LISREL VI correct} \mid \text{TETRAD II disagree})$	9.5
$P(\text{EQS correct} \mid \text{TETRAD II agree})$	43.7
$P(\text{EQS correct} \mid \text{TETRAD II disagree})$	2.7

11.8 Limitations of the TETRAD II Elaboration Search

The TETRAD II procedure cannot find the correct model if there are a large number of vanishing TETRAD differences that are not linearly implied by the true model, but hold because of coincidental values of the free parameters. Our study indicates that this occurrence is unusual, at least given the uniform distribution that we placed on the linear coefficients in the models that generated our data, but it certainly does occur. The same results can be expected for any other "natural" distribution on the parameters. Further, the search does not guarantee that it will find all of the models that have the highest *Tetrad-score*. But in many cases, depending upon the size of the model, the amount of background knowledge, the structure of the model, and the sample size, the search space is so large that a search that *guarantees* finding the models with the highest *Tetrad-score* is not practical. One way the procedure limits search is through the application of the simplicity principle. This is a substantive assumption that may be false. The simplicity assumption is not needed for some small models, but in many problems with more variables there may be a large number of models that have maximal scores but contain many redundant edges that do not contribute to the score. Without the use of the simplicity principle, it is often difficult to search this space of models and if it is searched, there may be so many models tied for the highest score that the output is uninformative. If a model with

"redundant" edges is correct, then our procedure will not find it. Typically these structures are underidentified, and so they could not be found by either LISREL VI or EQS.

The search procedure we have described here is practical for no more than several dozen variables. However, for larger numbers of variables, the MIMBuild algorithm described in Chapter 10 may be applicable.

Finally, there exist many latent variable models that cannot be distinguished by the vanishing tetrad differences they imply, but are nonetheless in principle statistically distinguishable. More reliable versions of the LISREL or EQS procedures might succeed in discovering such structures when the TETRAD procedures fail.

11.9 Some Morals for Statistical Search

There were three reasons why the TETRAD II procedure proved more reliable over the problems considered here than either of the other search procedures.

- (i) TETRAD II, unlike LISREL VI or EQS, does not need to estimate any parameters in order to conduct its search. Because the parameter estimation must be performed on an initial model that is wrong, LISREL VI and EQS often failed to converge, or calculated highly inaccurate parameter estimates. This in turn, led to problems in their respective searches.
- (ii) In the TETRAD II search, when the scores of several different models are tied, the program considers elaborations of each model. In contrast, LISREL VI and EQS arbitrarily chose a single model to elaborate.
- (iii) Both LISREL VI and EQS are less reliable than TETRAD II in deciding when to stop adding edges.

The morals for statistical search are evident: avoid iterative numerical procedures wherever possible; structure search so that it is feasible to branch when alternative steps seem equally good; find structural properties that permit reliable pruning of the search tree; for computational efficiency use local properties whenever possible; don't rely on statistical tests as stopping criteria without good evidence that they are reliable in that role.

Statistical searches cannot be adequately evaluated without clarity about the goals of search. We think in the social, medical and psychological uses of statistics the goals are often to find and estimate causal influence. The final moral for search is simple: once the goals are clearly and candidly given, if theoretical justifications of reliability are unavailable for the short run or even the long run, the computer offers the opportunity to subject the procedures to experimental tests of reliability under controlled conditions.

Chapter 12

Open Problems

A number of questions have been raised and not answered in the course of this book. Foremost among these are issues concerning extensions of the reliabilities and informativeness of the search algorithms. We record here a number of other questions that seem important. Some of the problems and questions may be quite easy, or may follow from results already available but unknown to us. Others have been worked at for some time by ourselves or others and appear to be quite difficult. Some are not particularly difficult but require work we have not done. All of the issues seem to us important to help fill out our understanding of the relations between causal structure and probability, and of the possibilities of causal inference and prediction.

12.1 Feedback, Reciprocal Causation, and Cyclic Graphs

"Reciprocal causation" may arise through treating A and B as variables in a population in which some A events cause B events and some B events cause A events. Consider, for example, a population of samples of ideal gasses, some of which have obtained their state by manipulation of temperature holding volume constant, and some by manipulation of volume holding temperature constant. In such cases the population under study is a mixture in the sense used throughout this book. If all causal connections are "reciprocal" in this way, then applying the PC Algorithm to such a population will generally yield a complete graph, and the FCI procedure will generally yield a complete graph with an "o" at each end of each edge. An interesting question is whether graphs with cycles may usefully represent causal systems in which some, but not all variables reciprocally cause each other.

In many applications statistical models are produced with a set of simultaneous equations, usually linear, in which some variable X , is specified as a function of Y and other variables, and reciprocally, Y is specified as a function of X and other variables. Such models are

standardly called "non-recursive." They are a clear indication that the relevant mathematical structure is not given by the algebra and probabilities alone, but involves directed graphical relationships, and in these cases graphs with *cycles*. The cyclic graphs are sometimes given explicitly. In contemporary cognitive science certain models of how humans compute have a related structure. In these theories computation is carried out through cyclic directed graphs whose nodes are random variables, in some cases variables taking only discrete values.

These mathematical structures are intended somehow to represent both the causal relationships among systems with feedback and the "equilibrium" probability relations among the variables of such systems. Just what that means is unclear, and part of the aim of this section is to try to clarify the matter a little. The analyses of representation, indistinguishability and inference developed earlier for acyclic representations of causal processes and probability distributions need to be extended to the cyclic case.

12.1.1 Mason's Theorem

For linear simultaneous equation systems in which each equation is accompanied by an "error term," assumed to be independent of other error terms, the algebra entails that each variable X is a function of various error terms only. The error terms and combinations of linear coefficients that occur in such an equation for a variable X depend on the cycles in which X occurs, the cycles in which variables in cycles with X occur, the cycles in which the variables in these neighboring cycles occur and so on. Nearly forty years ago, Samuel Mason gave a general analysis, entirely in terms of graphical properties, of how each variable in a cyclic graph depends on error variables and linear coefficients. For any linear system associated with a cyclic graph, taking appropriate expectations using Mason's formulas for the variables then gives the correlations entailed by the simultaneous equation system. So for the linear case there is a developed theory that can give us some information about when a cyclic graph linearly implies vanishing correlations or partial correlations. For a useful review and references see Heise (1975).

Mason's results enable us to use the linear case to examine how cyclic graphs may be interpreted as representations of the "equilibrium" or limiting results of time series, and whether the Markov Condition, factorization, Faithfulness, and d-separability are sensible or informative conditions for cyclic graphs.

12.1.2 Time Series and Cyclic Graphs

Since cyclic graphical models are meant to represent equilibria that result from feedback processes, each such model must correspond to the limiting distribution of some class of time series models. What is the correspondence? Consider a simple case:

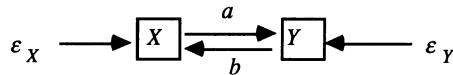


Figure 1

We have the equations

$$\begin{aligned} Y &= aX + \varepsilon_Y \\ X &= bY + \varepsilon_X \end{aligned}$$

which by Mason's rules yield

$$\gamma_{XY} = \frac{a\sigma_{\varepsilon_X}^2 + b\sigma_{\varepsilon_Y}^2}{(1-ab)^2}.$$

Consider the time series

$$\begin{aligned} X_t &= bY_{t-k} + \varepsilon_{X_{t-m}} \\ Y_t &= aX_{t-j} + \varepsilon_{Y_{t-n}} \end{aligned}$$

where for all t , $\varepsilon_{X_{t-m}}$ are i.i.d. and for all t , $\varepsilon_{Y_{t-n}}$ are i.i.d. For simplicity take $k = m = j = n = 1$. A finite segment of the corresponding directed graph is

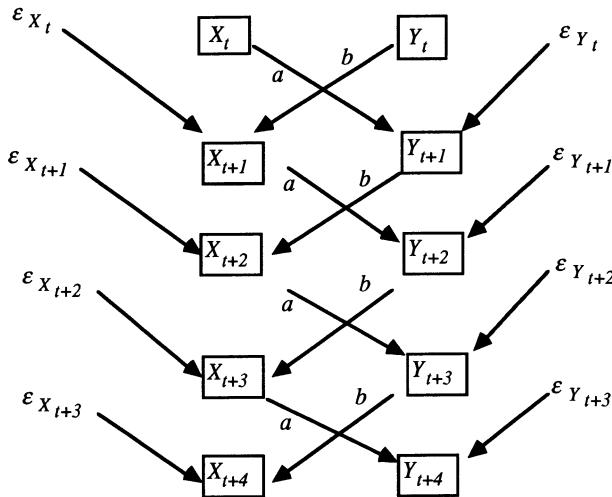


Figure 2

As t increases without bound this graph gives Mason's formula for the correlation of X and Y obtained from the simultaneous equations. If the ε_X and ε_Y -variables are constant, it also gives Mason's formula for the distribution of X and of Y . The same limit will of course result if at some finite time the values (or distributions) of the ε_X and ε_Y variables change but remain the same thereafter.

We can also have a "shock" model in which ε_X and ε_Y represent changes to a system in equilibrium. According to this model, for each individual X_0 and Y_0 have some arbitrary initial value, and ε_X and ε_Y represent changes to the system introduced (independent of X_0 , Y_0 , and each other) at time t_1 .

$$X_1 = X_0 + \varepsilon_X$$

$$Y_1 = Y_0 + \varepsilon_Y$$

$$X_t - X_{t-1} = b(Y_{t-1} - Y_{t-2})$$

$$Y_t - Y_{t-1} = a(X_{t-1} - X_{t-2})$$

The latter two equations we can write in a cyclical form:

$$\Delta Y = a\Delta X$$

$$\Delta X = b\Delta Y$$

Mason's formula entails that the new equilibrium values of X and Y , X_e and Y_e respectively are:

$$X_e = X_0 + \frac{\varepsilon_X + b\varepsilon_Y}{1-ab}$$

$$Y_e = Y_0 + \frac{\varepsilon_Y + a\varepsilon_X}{1-ab}$$

If ε_X and ε_Y have zero covariance it follows that the change $\Delta\gamma_{XY}$ in the covariance of X and Y , resulting from the "shocks" has the following form:

$$\Delta\gamma_{XY} = \frac{a\sigma_{\varepsilon_X}^2 + b\sigma_{\varepsilon_Y}^2}{(1-ab)^2}$$

A finite segment of the corresponding directed graph is depicted in figure 3.

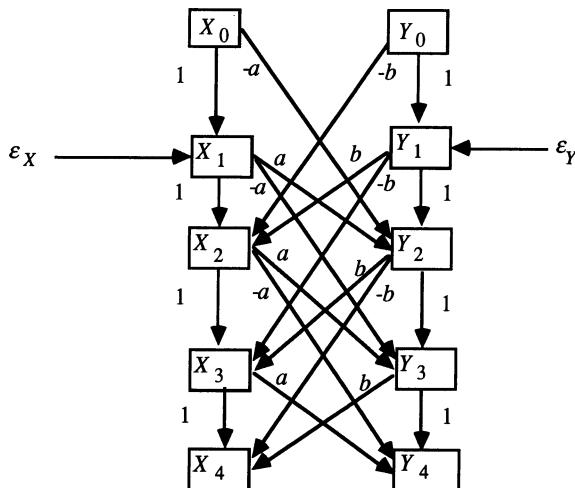


Figure 3

Other representations are also possible.

12.1.3 The Markov Condition, Factorizability and Faithfulness

The Markov Condition makes sense for a cyclic graph, but it may not be informative. Consider the graph:

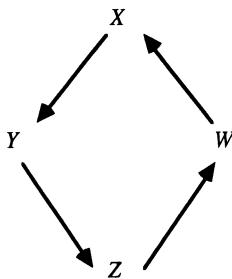


Figure 4

The Markov condition entails no independence or conditional independence relations among the four variables, because each variable is a descendent of each of the others. But the associated undirected independence graph represents two conditional independence claims: $X \perp\!\!\!\perp Z | \{Y, W\}$ and $Y \perp\!\!\!\perp W | \{X, Z\}$. Moreover, if we introduce an error variable adjacent to each of X , Y , Z and W , treat the errors as independent, and write each variable as a linear function of its error term and its parent, the partial correlation of X, Z controlling for Y, W vanishes, and so does the partial correlation of Y, W controlling for X, Z .

In the acyclic case the Markov condition is essentially equivalent to the requirement that d-separation implies conditional independence. The example just considered shows the equivalence does not obtain in cyclic graphs, where the d-separability condition is more informative. The vanishing partial correlations obtained from figure 4 are exactly those that result if we assume that d-separation implies conditional independence, or, in the linear case, vanishing partial correlation. Further

Theorem 12.1: If A and B are d-separated by C in a directed graph (cyclic or acyclic) G , then G linearly implies that the partial correlation of A, B controlling for C vanishes.

These results suggest the following conjecture.

Conjecture 1: if A and B are d-separated by set C in a directed graph (cyclic or acyclic) G , then G linearly implies that the partial correlation of A, B controlling for C vanishes.

The conjecture gains support from the reflection that the translations of cyclic graphs into time series models in the previous section preserves a form of d-separation. If in the finite cyclic graph A and B are d-separated by C , then in the infinite non-cyclic time series graph every A_t and B_{t+k} are d-separated by some occurrences of variables in C , and one would expect the corresponding partial correlation to vanish in the limit. If the conjecture is true, it would provide a strong case for adopting the *convention* for graphs of discrete variables that d-separation entails conditional independence.

Suppose that we adopt the convention that a cyclic directed graph G represents a distribution P if and only if whenever X and Y are d-separated given Z in G , X and Y are independent given Z in P . Unlike the acyclic case, it does not follow from this that for a positive distribution that a distribution represented by G can be written as the product of the conditional distribution of each variable given its parents in G . For example, suppose in G , X is a parent of Y , and Y is a parent of X . G represents any distribution over X and Y . However if the density f satisfies

$$f(XY) = f(X|Y)f(Y|X)$$

then it follows, simply by writing out the definitions of conditional density on the right hand side, that X and Y are independent. (Note that in linear models, if we adopt the convention that a cycle containing X and Y means X is a linear function of Y , and Y is a linear function of X , this does not entail that X and Y are uncorrelated.)

12.1.4 Discovery Procedures

Open Problem 1: Find a computationally feasible algorithm for inferring graphical structure when the measured variable set may be causally insufficient and there may be feedback.

12.2 Indistinguishability Relations

Assume for the moment the convention that d-separation implies conditional independence. For acyclic graphs the Faithfulness Condition is equivalent to the converse: if A and B are conditionally independent on C (or have vanishing partial correlation in the linear case) then C d-separates A, B . Likewise, we understand Faithfulness as a principle connecting conditional independence facts in a distribution with d-separation facts in a graph.

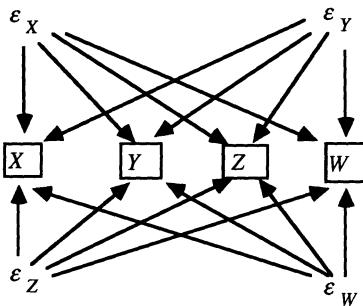
These assumptions and conventions make the Faithfulness Condition intelligible for cyclic as well as for acyclic graphs, and the definition of faithful indistinguishability (f.i.) from Chapter 4 therefore extends to cyclic graphs. Let us say that a cyclic path in a directed graph is **chordless** if and only if for each variable V in the cyclic path there is at most one directed edge out of V and into another variable in the cycle.

Conjecture 2: No directed cyclic graph with a chordless directed cycle of length 4 or more is faithfully indistinguishable from any directed acyclic graph on the same vertex set.

Given a system of linear equations with a cyclic graph, Mason's theorem permits one to write an "equivalent" system of equations in which each variable is a function of exogenous variables, and the corresponding graph is acyclic. But the exogenous variables will include the error terms, and so the acyclic model will have unmeasured common causes. In Chapter 4 we defined indistinguishability relations only for causally sufficient sets of variables. The first problem is therefore to say precisely what "equivalent" means.

Definition: Let G and H be directed graphs, and let \mathbf{O} be a subset of the vertex sets of both graphs. G and H are **faithfully indistinguishable over \mathbf{O}** if and only if for every distribution faithful to G there is a distribution faithful to H with the same \mathbf{O} marginal, and vice versa.

A parallel definition of **linearly faithful indistinguishable over \mathbf{O}** is obvious. Now the acyclic model obtained by applying Mason's theorem to a cyclic graph will not in general be linearly faithfully indistinguishable from the cyclic model. Thus in figure 4, if we give each of the variables a disturbance term, applying Mason's rules we obtain a model in which each measured variable is a function of each unmeasured disturbance, with a corresponding graph:

**Figure 5**

Notice that the graph in figure 5 does not entail the conditional independencies required by the graph in figure 4, and the two graphs are therefore not faithfully indistinguishable over $\mathbf{O} = \{X, Y, Z, W\}$. The graph of figure 5 is indistinguishable over \mathbf{O} from the graph of figure 4 only in the weak sense that for any choice of linear coefficients and distributions on the error variables for figure 4, there exists a choice of linear coefficients and distributions on the ϵ variables of figure 5 producing the same distribution on \mathbf{O} . But while the conditional independence relations among X, Y, Z and W will be stable for the model of figure 4 under a small variation of linear coefficients, they will not be stable for the model of figure 5.

There are cases in which acyclic graphs are faithfully indistinguishable from cyclic graphs. For example graphs 1 and 2 in figure 6 are faithfully indistinguishable.

$$(1) \quad W \xleftarrow{} X \xrightleftharpoons{} Y \xrightarrow{} Z$$

$$(2) \quad W \xleftarrow{} X \xrightarrow{} Y \xrightarrow{} Z$$

Figure 6

On the other hand, the graph in figure 7 is not faithfully indistinguishable from any acyclic graph on the same vertex set. It is, however, faithfully indistinguishable from an acyclic graph with unmeasured common causes.

**Figure 7**

Conjecture 3: Every directed graph without directed cycles of length greater than 2 is faithfully indistinguishable from an acyclic directed graph, possibly with extra vertices.

Open Problem 2: Give necessary and sufficient conditions for two directed acyclic graphs sharing a subset \mathbf{O} of vertices to be faithfully indistinguishable over \mathbf{O} .

Open Problem 3: Give necessary and sufficient conditions for two directed graphs, cyclic or acyclic, sharing a subset \mathbf{O} of vertices to be faithfully indistinguishable over \mathbf{O} .

Open Problem 4: Give necessary and sufficient conditions for two directed acyclic graphs sharing a subset \mathbf{O} of vertices to be linearly faithfully indistinguishable over \mathbf{O} .

Open Problem 5: Give necessary and sufficient conditions for two directed graphs, cyclic or acyclic, sharing a subset \mathbf{O} of vertices to be linearly faithfully indistinguishable over \mathbf{O} .

12.3 Time series and Granger Causality

Econometricians often deal with data in which the sample consists of the same variables measured at a discrete sequence of times. A variety of techniques have been developed to reduce inferences from such data to linear regression problems. Each series may be differenced to produce an approximation of stationarity, so that the values obtained at different times can be regarded as samples from the same population. (A stochastic process is said to be **strictly stationary** if the joint distribution of \mathbf{Z}_t is the same as the joint distribution of \mathbf{Z}_{t-k} for all time points t and all time lags k .) Variables may be transformed or data "filtered" in the attempt to approximate linearity and reduce autocorrelation.

The properties relevant to unit t in such a sample are not only the values x_t, y_t, z_t of variables X, Y and Z at t , but also the values these variables had at times $t-k$, for each k , that is, X_{t-k}, Y_{t-k} and Z_{t-k} . An outcome variable of interest, say Y , is then regressed on Y_{t-k}, X_{t-k} and Z_{t-k} ,

although in principle the choices of lags may be different for each series. The causal picture in one such regression model is illustrated in figure 8, where error terms are omitted:

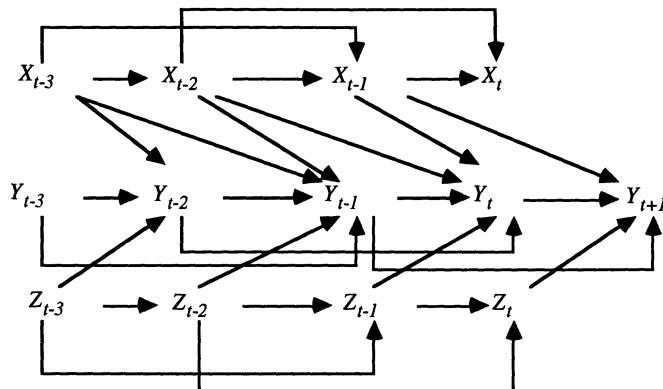


Figure 8

The regression equation in this case specifies Y_t as a function of the lags Y_{t-1} , Y_{t-2} , X_{t-1} , X_{t-2} , Z_{t-1} and an error term. The influence of X on Y , for example, is estimated by testing for vanishing partial correlations of Y_t with the X lags, controlling for the Y lags and the Z lags, or by testing simultaneously for vanishing linear coefficients associated with the X lags (Geweke, Meese and Dent, 1983). When there are multiple series, stepwise regression has been recommended and used for selecting the influences among series variables (Hsiao, 1981).

Granger's (1969) account of the notion of causal influence between time series is sometimes used in econometrics to justify the view that regression procedures result in reasonable estimates of causal influence. Granger assumes the series are stationary, denotes by $P_t(A|B)$ "the optimum, unbiased, least-squares predictor of A using the set of values B_t " and he denotes the prediction error by $\varepsilon_t(A|B) = A_t - P_t(A|B)$, with variance $\sigma^2(A|B)$. "Let U_t be all the information in the universe accumulated since time $t-1$ and let $U_t Y$ denote all this information apart from the specified series Y_t ." (p. 428) His definition is then that Y causes X if $\sigma^2(X|U) < \sigma^2(X|U-Y)$, where $U-Y$ is the U information prior to t absent the information in the Y series prior to t . Granger also proposes that instead of the imaginary set U , the set of all "relevant" information could be used.

The use of regression procedures to estimate causal influence in econometric time series usually realizes Granger's proposal by taking linear least squares estimators to be optimal and by taking

the "relevant" variables to be known a priori or to be determined by stepwise regression. One could quibble with Granger's definition, but the fundamental problem about common methods of causal inference from time series data appears to rest in these further assumptions. The problems with regression as a means for determining causal influence seem essentially the same in time series studies as elsewhere. First, tests of regression parameters waste degrees of freedom at the cost in small samples of power against alternatives. Since in many cases the number of observations is of the order of the number of parameters, whatever can be done to increase reliability should be. Second, it appears that while the time series setting removes ambiguities about the direction of dependencies, or edges, it does not remedy problems about unmeasured common causes of the outcome and regressors, and thus even asymptotically regression may yield significant coefficients for variables that are neither direct nor indirect causes of the outcome. These remarks are, however, entirely informal, and hence the next open question:

Open Problem 6: Do the methods of this book extend to causal relations among stationary time series?

12.4 Model Specification and Parameter Estimation from the Same Data Base

One of the routine objections to specification searches is that the error probabilities for parameter estimates in such models are no longer given by the conventional formulas. By simulation, Freedman, Navidi and Peters (1986) examine the estimates of mean squared error, mean square prediction error, and R^2 for bootstrap, jackknife and cross validation estimators in models specified as a function of sample data. On the whole, they find cross-validation works best but not especially well. Their case is extreme: they use 75 variables and 100 data points in each trial. A subset of the variables is selected by choosing the significant regressors in a multiple regression at the .25 level of significance.

It would be difficult to form any warranted conclusion about methodology for most applications from this single, nearly worst-case example, but the simulation does raise an interesting question. Freedman, et al, vary the success criteria and the estimation methods in their simulations, but they do not vary the method of model specification. (In fact they use a method that is not correct for causal inference unless prior assumptions about the causal structure are

made which are often not justified in applications.) The error probabilities of tests of parameter hypotheses are functions of the information in the data, represented by the number of data points and the number of free parameters in the model; intuitively, the model specification uses up some of that information. But not all specification searches must use the data in the same way or to the same extent. For example, if the true graph is sparse many of the methods we have described will use only low order tests of conditional independence, or tests of vanishing partial correlations and vanishing tetrad differences.

Open Problem 7: How do error probabilities of estimates vary for parameters in models specified by the PC or other algorithms, as a function of sample size and sparseness of the true graph?

A similar question applies to the search procedures for multiple indicator models using tests of vanishing tetrad differences.

12.5 Conditional Independence Tests

The examples in this book have used either approximately continuous variables assumed to be linearly related, or discrete variables, but never both, and never non-linear functional forms. The reasons for these restrictions are entirely statistical: the availability of relevant tests of conditional independence. Many empirical studies measure both discrete and continuous variables, and the continuous variables may sometimes be regarded as causes, and sometimes as effects, of the discrete variables. For the special case in which continuous variables never cause discrete variables, directed graphical treatments have been given in terms of joint distributions that are "conditionally Gaussian," that is the continuous variables are normally distributed conditional on each vector of values of their discrete parents. But in many instances where logistic regression is applied, and in many psychometric models, the cause is continuous and the effect discrete. Expanding the portfolio of tests of conditional independence to permit reliable decisions about conditional independence would likewise expand the range of applications for which causal inferences can reliably be made.

Chapter 13

Proofs of Theorems

We will adopt the following notational conventions. "w.l.g." abbreviates "without loss of generality", "r.h.s." abbreviates "right hand side", and "l.h.s." abbreviates "left hand side". Any sum over the empty set is equal to 0 and any product over the empty set is 1. $R(I,J)$ represents a directed path from I to J . If U is an undirected path from A to B , and X and Y occur on U , then we will denote the subpath of U between X to Y as $U(X,Y)$. $T(I,J)$ represents a trek in $\mathbf{T}(I,J)$. The definitions of all technical terms in this chapter that have not been defined in Chapters 2 or 3 have been placed in a glossary following the chapter.

13.1 Theorem 2.1

Theorem 2.1: If $P(\mathbf{V})$ is a positive distribution, then for any ordering of the variables in \mathbf{V} , P satisfies the Markov and Minimality conditions for the directed independence graph of $P(\mathbf{V})$ for that ordering.

Proof. See Pearl (1988).

13.2 Theorem 3.1

Theorem 3.1: If S is an LCT, and S' is a random coefficient LCT with the same directed acyclic graph, the same set of non-coefficient random variables, the same variances for each non-coefficient exogenous variable, and for each random coefficient a'_{IJ} in S' , $E(a'_{IJ}) = a_{IJ}$ in S , then a partial correlation is equal to zero in S if and only if it is equal to 0 in S' .

Let a **linear causal theory** be (LCT) be $\langle\langle R, M, E \rangle, (\Omega, f, P), EQ, L, Err \rangle$ where

- i. (Ω, f, P) is a probability space, where Ω is the sample space, f is a sigma-field over Ω , and P is a probability distribution over f .
- ii. $\langle R, M, E \rangle$ is a directed acyclic graph. R is a set of random variables over (Ω, f, P) .
- iii. The variables in R have a joint distribution. Every variable in R has a non-zero variance. E is a set of directed edges between variables in R . (M is the set of marks that occur in a directed graph, i.e. $\{EM, >\}$).
- iv. EQ is a consistent set of independent homogeneous linear equations in random variables in R . For each X_i in R of positive indegree there is an equation in EQ of the form

$$X_i = \sum_{X_j \in \text{Parents}(X_i)} a_{ij} X_j$$

where each a_{ij} is a non-zero real number and each X_i is in R . This implies that each vertex X_i in R of positive indegree can be expressed as a linear function of all and only its parents. There are no other equations in EQ . A non-zero value of a_{ij} is the **equation coefficient** of X_j in the equation for X_i .

- v. If vertices (random variables) X_i and X_j are exogenous, then X_i and X_j are pairwise statistically independent.
- vi. L is a function with domain E such that for each e in E , $L(e) = a_{ij}$ iff $\text{head}(e) = X_i$ and $\text{tail}(e) = X_j$. $L(e)$ will be called the **label** of e . By extension, the product of labels of edges in any acyclic undirected path U will be denoted by $L(U)$, and $L(U)$ will be called the **label** of U . The label of an empty path is fixed at 1.
- vii. There is a subset of S of R called the **error variables**, each of indegree 0 and outdegree 1. For every X_i in R of $\text{indegree} \neq 0$ there is exactly one error variable with an edge into X_i . We assume that the partial correlations of all orders involving only non-error variables are defined.

Note that the variance of any endogenous variable I conditional on any set of variables that does not contain the error variable of I is not equal to zero.

The definition of a **random coefficient linear causal theory** is the same as that of a linear causal theory except that each linear coefficient is a random variable independent of the set of all other random variables in the model.

A **linear causal form** is an unestimated LCT in which the linear coefficients and the variances of the exogenous variables are real variables instead of constants. This entails that an edge label in an LCF is a real variable instead of a constant (except that the label of an edge from an error variable is fixed at one.) More formally, let a linear causal form (**LCF**) be $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle, \mathbf{C}, \mathbf{V}, \mathbf{EQ}, \mathbf{L}, \mathbf{Err} \rangle$ where

- i. $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle$ is a directed acyclic graph. \mathbf{Err} is a subset of \mathbf{R} called the **error variables**. Each error variable is of indegree 0 and outdegree 1. For every X_i in \mathbf{R} of indegree $\neq 0$ there is exactly one error variable with an edge into X_i .
- ii. c_{ij} is a unique real variable associated with an edge from X_j to X_i , and \mathbf{C} is the set of c_{ij} . \mathbf{V} is the set of variables σ_i^2 , where X_i is an exogenous variable in $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle$ and σ_i^2 is a variable that ranges over the positive real numbers.
- iii. L is a function with domain E such that for each e in E , $L(e) = c_{ij}$ iff $head(e) = X_i$ and $tail(e) = X_j$. $L(e)$ will be called the **label** of e . By extension, the product of labels of edges in any acyclic undirected path U will be denoted by $L(U)$, and $L(U)$ will be called the **label** of U . The label of an empty path is fixed at 1.
- iv. \mathbf{EQ} is a consistent set of independent homogeneous linear equationals in variables in \mathbf{R} . For each X_i in \mathbf{R} of positive indegree there is an equation in \mathbf{EQ} of the form

$$X_i = \sum_{X_j \in \text{Parents}(X_i)} c_{ij} X_j$$

where each c_{ij} is a real variable in \mathbf{C} and each X_i is in \mathbf{R} . There are no other equations in \mathbf{EQ} . c_{ji} is the **equation coefficient** of X_j in the equation for X_i .

An LCT S is an **instance** of an LCF F if and only if the directed acyclic graph of S is isomorphic to the directed acyclic graph of F . In an LCF, a quantity (e.g. a covariance) X is **equivalent to a polynomial in the coefficients and variances of exogenous variables** if and only if for each LCF $F = \langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle, \mathbf{C}, \mathbf{V}, \mathbf{EQ}, \mathbf{L}, \mathbf{Err} \rangle$ and in every LCT $S = \langle\langle \mathbf{R}', \mathbf{M}', \mathbf{E}' \rangle, (\Omega f, P), \mathbf{EQ}', \mathbf{L}', \mathbf{Err}' \rangle$ that is an instance of F , there is a polynomial in the variables in \mathbf{C} and \mathbf{V} such that X is equal to the result of substituting the linear coefficients of S in as values for the corresponding variables in \mathbf{C} , and the variances of the exogenous variables in S as values for the corresponding variables in \mathbf{V} .

In an LCT or LCF S , a variable X_i is **independent** iff X_i has zero indegree (i.e. there are no edges directed into it); otherwise it is **dependent**. Note that the *property* of independence is completely distinct from the *relation* of statistical independence. The context will make clear in which of these senses the term is used. For a directed acyclic graph G , \mathbf{Ind} is the set of

independent variables in G . Given a directed acyclic graph G , $\mathbf{D}(X_i, X_j)$ is the set of all directed paths from X_i to X_j . In an LCF <<R,M,E>, C, V, EQ,L,S> an equation is an **independent equational for a dependent variable** X_j if and only if it is implied by EQ and the variables in R which appear on the r.h.s. are independent and occur at most once. Ind_{IJ} is the coefficient of J in the independent equational for I .

Lemma 3.1.1: In an LCF S , if J is an independent variable, then

$$Ind_{IJ} = \sum_{U \in \mathbf{D}(J,I)} L(U)$$

Proof. This is a special case of Mason's rule for calculating the "total effect" of a variable J on a variable I . See Glymour et. al. (1987). ∴

The following two lemmas show how to calculate the variance of random variables and covariances between random variables in terms of the covariances between other random variables. The proofs of these lemmas can be found in Freund and Walpole (1980). We denote the covariance of I and J by γ_{IJ} , the variance of I by σ_I^2 , the correlation of I and J by ρ_{IJ} , the partial correlation of I and J given the set H by $\rho_{IJ,H}$, and the partial covariance of I and J given H by $\gamma_{IJ,H}$. The correlation of two subscripted variables such as X_i and X_j we will write as ρ_{ij} for legibility, and similarly for partial correlations, etc.

Lemma 3.1.2: If Q is a set of random variables with a joint probability distribution and

$$Y = \sum_{I \in Q} a_{YI} I$$

and

$$Z = \sum_{J \in Q} a_{ZJ} J$$

then

$$\gamma_{YZ} = \sum_{I \in Q} \sum_{J \in Q} a_{YI} a_{ZJ} \gamma_{IJ}$$

Lemmas 3.1.3, 3.1.5, and 3.1.7 are not used in the proof of Theorem 3.1, but they are used in later theorems, and we include them here because they follow easily from the other lemmas in this section.

Lemma 3.1.3: If \mathbf{Q} is a set of random variables with a joint probability distribution and

$$Y = \sum_{I \in \mathbf{Q}} a_{YI} I$$

then

$$\sigma_Y^2 = \sum_{I \in \mathbf{Q}} \sum_{J \in \mathbf{Q}} a_{YI} a_{YJ} \gamma_{IJ}$$

In an LCF S , \mathbf{U}_X is the set of all independent variables that are the source of a directed path to X . (Note that if X is independent then $X \in \mathbf{U}_X$ since there is an empty path from every vertex to itself.) In an LCF S , \mathbf{U}_{XY} is $\mathbf{U}_X \cap \mathbf{U}_Y$.

Lemma 3.1.4: If S is an LCF,

$$Y = \sum_{I \in \mathbf{Ind}} {}^{Ind} a_{YI} I$$

and

$$Z = \sum_{I \in \mathbf{Ind}} {}^{Ind} a_{ZI} I$$

then

$$\gamma_{YZ} = \sum_{I \in \mathbf{U}_{YZ}} {}^{Ind} a_{YI} {}^{Ind} a_{ZI} \sigma_I^2$$

Proof. \mathbf{Ind} is a set of independent variables. It follows that γ_{IJ} is equal to 0 if $I \neq J$, and γ_{II} is equal to σ_I^2 if $I = J$. Substituting these values for γ_{IJ} into the r.h.s. of the equation for γ_{YZ} in lemma 3.1.2 shows that

$$(1) \quad \gamma_{YZ} = \sum_{I \in \mathbf{Ind}} {}^{Ind} a_{YI} {}^{Ind} a_{ZI} \sigma_I^2$$

If I is in \mathbf{Ind} , but I is not in \mathbf{U}_{YZ} then there is no pair of directed acyclic paths from I to Y and Z . By lemma 3.1.1, if there is no pair of directed acyclic paths from I to Y and Z , then the coefficient of I in the independent equation for either Y or Z is zero. So, the only non-zero terms in equation 1 are for $I \in \mathbf{U}_{YZ}$. ∴

Lemma 3.1.5: If S is an LCF,

$$Y = \sum_{I \in \text{Ind}}^{Ind} a_{YI} I$$

then

$$\sigma_Y^2 = \sum_{I \in U_Y}^{Ind} a_{YI}^2 \sigma_I^2$$

Proof. Ind is a set of independent random variables. It follows that γ_{IJ} is equal to 0 if $I \neq J$, and γ_{II} is equal to σ_I^2 if $I = J$. Substituting these values for γ_{IJ} into the r.h.s. of the equation for σ_Y^2 in lemma 3.1.3 proves that

$$(2) \quad \sigma_Y^2 = \sum_{I \in \text{Ind}}^{Ind} a_{YI}^2 \sigma_I^2$$

If I is in Ind , but I is not in U_Y , then there is no directed path from I to Y . It follows from lemma 3.1.1 that a_{YI} is zero. Hence the only non-zero terms in equation 2 come from $I \in U_Y$.

∴

Lemma 3.1.6: If S is an LCF,

$$\gamma_{IJ} = \sum_{K \in U_{IJ}} \sum_{R \in D(K, I)} \sum_{R' \in D(K, J)} L(R)L(R') \sigma_K^2$$

Proof. This follows immediately from lemmas 3.1.2 and 3.1.4. ∴

Lemma 3.1.7: If S is an LCF,

$$\sigma_I^2 = \sum_{K \in U_I} \left(\left(\sum_{R \in D(K, I)} L(R) \right)^2 \sigma_K^2 \right)$$

Proof. This follows immediately from lemmas 3.1.1 and 3.1.5. ∴

Theorem 3.1: If S is an LCT, and S' is a random coefficient LCT with the same directed acyclic graph, the same set of non-coefficient random variables, the same variances for each exogenous variable, and for each random coefficient a'_{IJ} in S' , $E(a'_{IJ}) = a_{IJ}$ in S , then a partial correlation is equal to zero in S if and only if it is equal to 0 in S' .

Proof. Because S is an instance of an LCF, by lemma 3.1.6

$$\gamma_{IJ} = \sum_{K \in \mathbf{U}_{IJ}} \sum_{R \in \mathbf{D}(K, I)} \sum_{R' \in \mathbf{D}(K, J)} L(R)L(R')\sigma_K^2$$

The label of a path is equal to the product of the labels of the edges and because the random coefficients are independent of each other and all the random variables that are not coefficients, it follows that

$$E\left(\prod_{edge \in U} L(edge)\right) = \prod_{edge \in U} E(L(edge))$$

Transform all of the variables so that they have mean 0; this does not affect the value of any of the covariances. In T , $\gamma_{IJ} = E(IJ)$ and

$$\begin{aligned} E(IJ) &= E\left(\sum_{H \in \mathbf{U}_I} \sum_{U \in \mathbf{D}(H, X)} \sum_{F \in \mathbf{U}_J} \sum_{V \in \mathbf{D}(F, Y)} L(U)L(V)HF\right) = \\ &\quad \sum_{H \in \mathbf{U}_{IJ}} \sum_{U \in \mathbf{D}(H, X)} \sum_{V \in \mathbf{D}(H, Y)} E(L(U)L(V)H^2) = \\ &\quad \sum_{H \in \mathbf{U}_{IJ}} \sum_{U \in \mathbf{D}(H, X)} \sum_{V \in \mathbf{D}(H, Y)} E\left(\prod_{edge \in U} L(edge) \prod_{edge \in V} L(edge) H^2\right) = \\ &\quad \sum_{H \in \mathbf{U}_{IJ}} \sum_{U \in \mathbf{D}(H, X)} \sum_{V \in \mathbf{D}(H, Y)} \prod_{edge \in U} E(L(edge)) \prod_{edge \in V} E(L(edge)) E(H^2) \end{aligned}$$

because for exogenous variables $E(HF) = 0$ unless $H = F$.

By hypothesis, $E(L(edge))$ in $S' = L(edge)$ in S . Hence the expression γ_{IJ} is the same for both random and constant coefficients. The partial correlations are a function of the covariance matrix so the partial correlations are the same in S and S' . \therefore

13.3 Theorem 3.2

Theorem 3.2: Let M be an LCF with n free linear coefficients a_1, \dots, a_n and k positive variances v_1, \dots, v_k . Let $M(<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>)$ be the distributions consistent with specifying values $<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>$ for a_1, \dots, a_n and v_1, \dots, v_k . Let Π be the set of probability measures P on the space \Re^{n+k} of values of the parameters of M such that for every subset V of \Re^{n+k} having Lebesgue measure zero, $P(V) = 0$. Let Q be the set of vectors of coefficient and variance values such that for all q in Q every probability distribution consistent with $M(q)$ has a vanishing partial correlation that is not linearly implied by M . Then for all P in Π , $P(Q) = 0$.

Lemma 3.2.1: In an LCF S , $\rho_{ij,X} = 0$ is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables.

Proof. We will prove more generally that a polynomial equation in partial covariances is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables. If X contains n distinct variables, then say $\rho_{ij,X}$ is a partial correlation of order n . Let the pc-order (partial covariance order) of a polynomial in partial covariances be the highest order of any partial covariance appearing in the polynomial. The proof is by induction on the pc-order of the polynomials.

Base Case. If polynomial Q is of pc-order 0, then by lemma 3.1.2, Q is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables.

Induction Case. Suppose that the lemma is true for polynomials of pc-order $n-1$, and let Q be a polynomial of pc-order n . The recursion formula for partial covariances is

$$\gamma_{ij,Y \cup r} = \gamma_{ij,Y} - \frac{\gamma_{ir,Y} \gamma_{jr,Y}}{\gamma_{rr,Y}}$$

Form Q' by using this recursion formula to replace each covariance of pc-order n appearing in Q by an algebraic combination of covariances of pc-order $n-1$. Form Q'' by multiplying Q' by the lowest common denominator of all of the terms in Q' , producing a polynomial of pc-order $n-1$. By the induction hypothesis, Q'' is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables. Hence, a polynomial equation in partial covariances is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables.

By definition,

$$\rho_{ij.X} = \frac{\gamma_{ij.X}}{\sqrt{\gamma_{ii.X}} \sqrt{\gamma_{jj.X}}}$$

so $\rho_{ij.X} = 0$ iff $\gamma_{ij.X} = 0$. Since the latter is a polynomial equation in partial covariances, it is equivalent to a polynomial equation in the linear coefficients and variances of the independent variables. It follows that the former is also equivalent to a polynomial equation in the linear coefficients and variances of the independent variables. ∴

Theorem 3.2: Let M be an LCF with n free linear coefficients a_1, \dots, a_n and k positive variances v_1, \dots, v_k . Let $M(<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>)$ be the distributions consistent with specifying values $<u_1, \dots, u_n, u_{n+1}, \dots, u_{n+k}>$ for a_1, \dots, a_n and v_1, \dots, v_k . Let Π be the set of probability measures P on the space \Re^{n+k} of values of the parameters of M such that for every subset V of \Re^{n+k} having Lebesgue measure zero, $P(V) = 0$. Let Q be the set of vectors of coefficient and variance values such that for all q in Q every probability distribution consistent with $M(q)$ has a vanishing partial correlation that is not linearly implied by M . Then for all P in Π , $P(Q) = 0$.

Proof. For any LCF, each partial correlation is equivalent to a polynomial in the linear coefficients and the variances of the exogenous variables: the rest of the features of the distribution have no bearing on the partial correlation. Hence for a vanishing partial correlation to be linearly implied by the directed acyclic graph of the theory, it is necessary and sufficient that the corresponding polynomial in the linear coefficient and variance parameters vanish identically. Thus any vanishing partial correlation not linearly implied by an LCF represents a polynomial P in variables consisting of the linear coefficients and variances of that theory, and the polynomial does not vanish identically. So the set of linear coefficient and variance values satisfying P is an algebraic variety in \Re^{n+k} . Any connected component of such a variety has Lebesgue measure zero. But an algebraic variety has at most a finite number of connected components (Whitney, 1957). ∴

13.4 Theorem 3.3

Theorem 3.3 $P(V)$ is faithful to directed acyclic graph G with vertex set V if and only if for all disjoint sets of vertices X , Y , and Z , X and Y are independent conditional on Z if and only if X and Y are d-separated given Z .

The "if" portion of the theorem was first proved in Verma (1986) and the "only if" portion of the theorem was first proved in Geiger and Pearl (1989a). The proof produced here is considerably different, but since the bulk of it is a series of lemmas that we also need to prove other theorems, we state it here.

G' is an **inducing path graph over O for directed acyclic graph G** if and only if O is a subset of the vertices in G , there is an edge between variables A and B with an arrowhead at A if and only if A and B are in O , and there is an inducing path in G between A and B relative to O that is into A . (Using the notation of Chapter 2, the set of marks in an inducing path graph is $\{>, EM\}$.) We will refer to the variables in O as **observed** variables. Unlike a directed acyclic graph, an inducing path graph can contain double-headed arrows. However, it does not contain any edges with no arrowheads. If there is an inducing path between A and B in G that is into A , then the edge between A and B in G' is into A . However, if there is an inducing path between A and B in G that is out of A , it does not follow that the edge in G' between A and B is out of A . Only if *no* inducing path between A and B in G is into A is the edge between A and B in G' out of A . The definitions of directed path, d-separability, inducing path, collider, ancestor, and descendant are the same as those for directed graphs, i.e. a directed path in an inducing path graph, as in an acyclic directed graph, contains only directed edges (e.g. $A \rightarrow B$). However, an undirected path in an inducing path graph can contain either directed edges, or bi-directed edges (e.g. $C <-> D$.) Also, if $A <-> B$ in an inducing path graph, A is not a parent of B . Note that if G is a directed acyclic graph, and G' the inducing path graph for G over O , then there are no directed cycles in G' .

Lemma 3.3.1 states a method for constructing a path between X and Y that d-connects X and Y given Z out of a sequence of paths.

Lemma 3.3.1: In a directed acyclic graph G (or an inducing path graph G') over V , if X and Y are not in Z , there is a sequence S of distinct vertices in V from X to Y , and there is a set T of undirected paths such that

- (i). for each pair of adjacent vertices V and W in S there is a unique undirected path in T that d-connects V and W given $\mathbf{Z}\{V,W\}$, and
- (ii). if a vertex Q in S is in \mathbf{Z} , then the paths in T that contain Q as an endpoint collide at Q , and
- (iii). if for three vertices V, W, Q occurring in that order in S the d-connecting paths in T between V and W , and W and Q collide at W then W has a descendant in \mathbf{Z} ,

then there is a path U in G that d-connects X and Y given \mathbf{Z} . In addition, if all of the edges in all of the paths in T that contain X are into (out of) X then U is into (out of) X , and similarly for Y .

Proof. Let U' be the concatenation of all of the paths in T in the order of the sequence S . U' may not be an acyclic undirected path, because it may contain some vertices more than once. Let U be the result of removing all of the cycles from U' . If each edge in U' that contains X is into (out of) X , then U is into (out of) X , because each edge in U is an edge in U' . Similarly, if each edge in U' that contains Y is into (out of) Y , then U is into (out of) Y , because each edge in U is an edge in U' . We will prove that U d-connects X and Y given \mathbf{Z} .

We will call an edge in U containing a given vertex V an endpoint edge if V is in the sequence S , and the edge containing V occurs on the path in T between V and its predecessor or successor in S ; otherwise the edge is an internal edge.

First we prove that every member R of \mathbf{Z} that is on U is a collider on U . If there is an endpoint edge containing R on U then it is into R because by assumption the paths in T containing R collide at R . If an edge on U is an internal edge with endpoint R then it is into R because it is an edge on a path that d-connects two variables A and B not equal to R given $\mathbf{Z}\{A,B\}$, and R is in \mathbf{Z} . All of the edges on paths in T are into R , and hence the subset of those edges that occur on U are into R .

Next we show that every collider R on U has a descendant in \mathbf{Z} . R is not equal to either of the endpoints X or Y , because the endpoints of a path are not colliders along the path. If R is a collider on any of the paths in T then R has a descendant in \mathbf{Z} because it is an edge on a path that d-connects two variables A and B not equal to R given $\mathbf{Z}\{A,B\}$. If R is a collider on two endpoint edges then it has a descendant in \mathbf{Z} by hypothesis. Suppose then that R is not a collider on the path in T between A and B , and not a collider on the path in T between C and D , but after cycles have been removed from U' , R is a collider on U . In that case U' contains an undirected cycle containing R . Because G is acyclic, the undirected cycle contains a collider. Hence R has a descendant that is a collider on U' . Each collider on U' has a descendant in \mathbf{Z} . Hence R has a descendant in \mathbf{Z} . \therefore

Lemma 3.3.2: If G is a directed acyclic graph (or an inducing path graph), R is d-connected to Y given Z by undirected path U , and W and X are distinct vertices on U not in Z , then $U(W,X)$ d-connects W and X given $Z = Z \setminus \{W,X\}$.

Proof. Suppose G is a directed acyclic graph, R is d-connected to Y given Z by undirected path U , and W and X are distinct vertices on U not in Z . Each non-collider on $U(W,X)$ except for the endpoints is a non-collider on U , and hence not in Z . Every collider on $U(W,X)$ has a descendant in Z because each collider on $U(W,X)$ is a collider on U , which d-connects R and Y given Z . It follows that $U(W,X)$ d-connects W and X given $Z = Z \setminus \{W,X\}$. \therefore

Lemma 3.3.3: If G is a directed acyclic graph (or an inducing path graph), R is d-connected to Y given Z by undirected path U , there is a directed path D from R to X that does not contain any member of Z , and X is not on U , then X is d-connected to Y given Z by a path U' that is into X . If D does not contain Y , then U' is into Y if and only if U is.

Proof. Let D be a directed path from R to X that does not contain any member of Z , and U an undirected path that d-connects R and Y given Z and does not contain X . Let Q be the point of intersection of D and U that is closest to Y on U . Q is not in Z because it is on D .

If D does contain Y , then $Y = Q$, and $D(Y,X)$ is a path into X that d-connects X and Y given Z because it contains no colliders and no members of Z .

If D does not contain Y then $Q \neq Y$. $X \neq Q$ because X is not on U and Q is. By lemma 3.3.2 $U(Q,Y)$ d-connects Q and Y given $Z \setminus \{Q,Y\} = Z$. Also, $D(Q,X)$ d-connects Q and X given $Z \setminus \{Q,X\} = Z$. $D(Q,X)$ is out of Q , and Q is not in Z . By lemma 3.3.1, there is a path U' that d-connects X and Y given Z that is into X . If Y is not on D , then all of the edges containing Y in U' are in $U(Q,Y)$, and hence by lemma 3.3.1 U' is into Y if and only if U is. \therefore

In a directed acyclic graph G , $\text{ND}(Y)$ is the set of all vertices that do not have a descendant in Y

Lemma 3.3.4: If $P(V)$ satisfies the Markov condition for directed acyclic graph G over V , S is a subset of V , and $\text{ND}(Y)$ is included in S , then

$$\sum_{S} \left(\prod_{V \in V} P(V | \text{Parents}(V)) \right) = \sum_{S \setminus \text{ND}(Y)} \left(\prod_{V \in V \setminus \text{ND}(Y)} P(V | \text{Parents}(V)) \right)$$

Proof. S can be partitioned into $S \setminus ND(Y)$ and $S \cap ND(Y) = ND(Y)$. If V is in $V \setminus ND(Y)$ then no variable occurring in the term $P(V|Parents(V))$ occurs in $ND(Y)$; hence for each V in $V \setminus ND(Y)$, $P(V|Parents(V))$ can be removed from the scope of the summation over the values of variables in $ND(Y)$.

$$(1) \quad \sum_{S \setminus ND(Y)}^{\rightarrow} \left(\prod_{V \in S \setminus ND(Y)} P(V|Parents(V)) \times \left(\sum_{ND(Y)}^{\rightarrow} \left(\prod_{V \in ND(Y)} P(V|Parents(V)) \right) \right) \right)$$

We will now show that

$$\sum_{ND(Y)}^{\rightarrow} \left(\prod_{V \in ND(Y)} P(V|Parents(V)) \right) = 1$$

unless for some value of $S \setminus ND(Y)$ the set of values of $ND(Y)$ such that $P(V|Parents(V))$ is defined for each V in $ND(Y)$ is empty, in which case on the l.h.s of (1) no term containing that value of $S \setminus ND(Y)$ appears in the sum, and on the r.h.s.of (1) every term in the scope of the summation over $S \setminus ND(Y)$ that contains that value of $S \setminus ND(Y)$ is zero.

Let $P(W|Parents(W))$ be a term in the factorization such that W does not occur in any other term, i.e. W is not the parent of any other variable. If $ND(Y)$ is not empty W is in $ND(Y)$.

$$\begin{aligned} & \sum_{ND(Y)}^{\rightarrow} \left(\prod_{V \in ND(Y)} P(V|Parents(V)) \right) = \\ & \sum_{ND(Y) \setminus \{W\}}^{\rightarrow} \left(\prod_{V \in ND(Y) \setminus \{W\}} P(V|Parents(V)) \right) \times \left(\sum_W^{\rightarrow} P(W|Parents(W)) \right) \end{aligned}$$

The latter expression can now be written as

$$\sum_{ND(Y) \setminus \{W\}}^{\rightarrow} \left(\prod_{V \in ND(Y) \setminus \{W\}} P(V|Parents(V)) \right)$$

because $\sum_w P(W|\text{Parents}(W))$ is equal to one. Now some element in $\text{ND}(\mathbf{Y}) \setminus W$ is not a parent of any other member of $\text{ND}(\mathbf{Y}) \setminus \{W\}$, and the process can be repeated until each element is removed from $\text{ND}(\mathbf{Y})$. \therefore

In a directed acyclic graph G , if $\mathbf{Y} \cap \mathbf{Z} = \emptyset$, then V is in $\text{IV}(\mathbf{Y}, \mathbf{Z})$ (informative variables for \mathbf{Y} given \mathbf{Z}) if and only if V is d-connected to \mathbf{Y} given \mathbf{Z} , and V is not in $\text{ND}(\mathbf{YZ})$. (This entails that V is not in $\mathbf{Y} \cup \mathbf{Z}$ by definition of d-connection.) In a directed acyclic graph G , if $\mathbf{Y} \cap \mathbf{Z} = \emptyset$, W is in $\text{IP}(\mathbf{Y}, \mathbf{Z})$ (W has a parent that is an informative variable for \mathbf{Y} given \mathbf{Z}) if and only if W is a member of \mathbf{Z} , and W has a parent in $\text{IV}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}$.

Lemma 3.3.5: If P satisfies the Markov condition for directed acyclic graph G over \mathbf{V} , then

$$P(\mathbf{Y}|\mathbf{Z}) = \frac{\sum_{\substack{\rightarrow \\ \text{IV}(\mathbf{Y}, \mathbf{Z})}} \prod_{W \in \text{IV}(\mathbf{Y}, \mathbf{Z}) \cup \text{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}} P(W|\text{Parents}(W))}{\sum_{\substack{\rightarrow \\ \mathbf{V} \setminus \mathbf{Z} \cup \mathbf{Y}}} \prod_{W \in \text{IV}(\mathbf{Y}, \mathbf{Z}) \cup \text{IP}(\mathbf{Y}, \mathbf{Z}) \cup \mathbf{Y}} P(W|\text{Parents}(W))}$$

for all values of \mathbf{V} for which the conditional distributions in the factorization are defined, and for which $P(\mathbf{z}) \neq 0$.

Proof. Let $\mathbf{V}' = \mathbf{V} \setminus \text{ND}(\mathbf{YZ})$, i.e. the subset of \mathbf{V} with descendants in \mathbf{YZ} . It follows from the definition of conditional probability that

$$P(\mathbf{Y}|\mathbf{Z}) = \frac{P(\mathbf{YZ})}{P(\mathbf{Z})} = \frac{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{YZ}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))}{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{Z}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))}$$

By lemma 3.3.4,

$$\frac{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{YZ}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))}{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{Z}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))} = \frac{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{YZ}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))}{\sum_{\substack{\rightarrow \\ \mathbf{V}' \setminus \mathbf{Z}}} \prod_{W \in \mathbf{V}'} P(W|\text{Parents}(W))}$$

First we will show that we can factor the numerator and the denominator into a product of two sums. The second term in both the numerator and the denominator is the same, so it cancels. In the case of the denominator, we show that

$$\sum_{V' \setminus Z}^{\rightarrow} \prod_{W \in V'} P(W | \text{Parents}(W)) = \\ \sum_{\text{IV}(Y, Z) \cup Y}^{\rightarrow} \prod_{W \in \text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y} P(W | \text{Parents}(W)) \times \sum_{V' \setminus (\text{IV}(Y, Z) \cup YZ)}^{\rightarrow} \prod_{W \in V' \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)} P(W | \text{Parents}(W))$$

by demonstrating that if W is in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$, then neither W nor any parent of W occurs in the scope of the summation over $V' \setminus (\text{IV}(Y, Z) \cup YZ)$, and also that if W is in $V' \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$ then neither W nor any parent of W is in the scope of the summation over $\text{IV}(Y, Z) \cup Y$.

First we demonstrate that if W is in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$ then W is not in $V' \setminus (\text{IV}(Y, Z) \cup YZ)$. If W is in $\text{IV}(Y, Z) \cup Y$ then trivially it is not in $V' \setminus (\text{IV}(Y, Z) \cup YZ)$. If W is in $\text{IP}(Y, Z)$ then W is in Z , so W is not in $V' \setminus (\text{IV}(Y, Z) \cup YZ)$.

Now we will demonstrate if W is in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$ then no parent of W is in $V' \setminus (\text{IV}(Y, Z) \cup YZ)$. Suppose first that W is in $\text{IV}(Y, Z)$ and T is a parent of W . If T is in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$ this reduces to the previous case. Assume then that T is not in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$. We will show that T is in YZ . T is not d-connected to Y given Z . However, W , a child of T , is d-connected to Y given Z by some path U . If T is on U then T is d-connected to Y given Z , contrary to our assumption, unless T is in YZ . If T is not on U , and U is not into W , then the concatenation of the edge between T and W with U d-connects T and Y given Z , contrary to our assumption, unless T is in YZ . If T is not on U , but U is into W , then because W is in $\text{IV}(Y, Z)$ it has a descendant in YZ . If W has a descendant in Z , then W is a collider on the concatenation of the edge between T and W with U , and has a descendant in Z ; hence T is d-connected to Y given Z , contrary to our assumption, unless T is in YZ . If W does not have a descendant in Z , then there is a directed path D from W to Y that does not contain any member of Z . The concatenation of the edge from T to W and D d-connects T and Y given Z , contrary to our assumption, unless T is in YZ . In any case, T is in YZ , and not in $V' \setminus (\text{IV}(Y, Z) \cup YZ)$.

Suppose next that W is in $\text{IP}(Y, Z)$ and T is a parent of W . It follows that some parent R of W is in $\text{IV}(Y, Z)$ or in Y , and W is in Z . If T is in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$ this reduces to the

previous case. Assume then that T is not in $\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y$. If R is in Y , then T is d-connected to Y given Z by the concatenation of the edge from R to W and the edge from W to T , contrary to our assumption, unless T is in YZ . Hence T is in YZ , and not in $V \setminus (\text{IV}(Y, Z) \cup YZ)$. Assume then that R is in $\text{IV}(Y, Z)$. R is d-connected to Y given Z by some path U . If T is on U then T is d-connected to Y given Z unless T is in YZ . If W is on U , but T is not, then W is a collider on U , because W is in Z . W is also a collider on the concatenation of the edge from T to W with the subpath of U from W to Y ; hence this path d-connects T and Y given Z unless T is in YZ . If neither T nor W is on U , then the concatenation of the edge between T and W , the edge between W and R , and U , is a path on which W is a collider and R is not a collider (because R is a parent of W); hence this path d-connects T and Y given Z , unless W is in YZ . By hypothesis, T is not d-connected to Y given Z because T is not in $\text{IV}(Y, Z)$; it follows that T is in YZ . Hence T is not in $V \setminus (\text{IV}(Y, Z) \cup YZ)$.

Suppose finally that W is in Y and T is a parent of W . It follows that T is d-connected to Y given Z unless T is in YZ . By hypothesis, T is not d-connected to Y given Z because T is not in $\text{IV}(Y, Z)$ so T is in YZ . Hence T is not in $V \setminus (\text{IV}(Y, Z) \cup YZ)$.

Now we will demonstrate by contraposition that if W is in $V \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$ then neither W nor any parent of W is in the scope of the summation over $\text{IV}(Y, Z) \cup Y$. Suppose W or some parent T of W is in $\text{IV}(Y, Z) \cup Y$. If W is in $\text{IV}(Y, Z) \cup Y$ it follows trivially that W is not in $V \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$. Suppose T is in $\text{IV}(Y, Z) \cup Y$ but W is not. We will show that W is in YZ . If T is in Y , then W is d-connected to Y given Z , contrary to our assumption, unless T is in YZ . If T is in $\text{IV}(Y, Z)$ it follows that there is a path U d-connecting T and Y given Z . If W is on U , then W is d-connected to Y given Z , contrary to our hypothesis, unless W is in YZ . If W is not on U , then the concatenation of the edge between W and T with U d-connects W and Y given Z (because T is not a collider and not in Z), contrary to our hypothesis, unless W is in YZ . It follows that W is in YZ . If W is in Z , then W is in $\text{IP}(Y, Z)$, and hence not in $V \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$. If W is in Y , then W is not in $V \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$. Hence by contraposition, if W is in $V \setminus (\text{IV}(Y, Z) \cup \text{IP}(Y, Z) \cup Y)$ then neither W nor any parent of W is in the scope of the summation over $\text{IV}(Y, Z) \cup Y$.

The proof for the numerator is essentially the same. Hence,

$$\begin{aligned}
& \frac{\sum_{V' \setminus YZ}^{\rightarrow} \prod_{W \in V'} P(W | \text{Parents}(W))}{\sum_{V' \setminus Z}^{\rightarrow} \prod_{W \in V'} P(W | \text{Parents}(W))} = \\
& \frac{\sum_{IV(Y,Z)}^{\rightarrow} \prod_{W \in IV(Y,Z) \cup IP(Y,Z) \cup Y} P(W | \text{Parents}(W))}{\sum_{IV(Y,Z) \cup Y}^{\rightarrow} \prod_{W \in IV(Y,Z) \cup IP(Y,Z) \cup Y} P(W | \text{Parents}(W))} \times \\
& \frac{\sum_{V' \setminus (IV(Y,Z) \cup YZ)}^{\rightarrow} \prod_{W \in V' \setminus (IV(Y,Z) \cup IP(Y,Z) \cup Y)} P(W | \text{Parents}(W))}{\sum_{V' \setminus (IV(Y,Z) \cup YZ)}^{\rightarrow} \prod_{W \in V' \setminus (IV(Y,Z) \cup IP(Y,Z) \cup Y)} P(W | \text{Parents}(W))} = \\
& \frac{\sum_{IV(Y,Z)}^{\rightarrow} \prod_{W \in IV(Y,Z) \cup IP(Y,Z) \cup Y} P(W | \text{Parents}(W))}{\sum_{IV(Y,Z) \cup Y}^{\rightarrow} \prod_{W \in IV(Y,Z) \cup IP(Y,Z) \cup Y} P(W | \text{Parents}(W))}
\end{aligned}$$

⋮

Lemma 3.3.6: In a directed acyclic graph G , if V is d-connected to Y given Z , and X is d-separated from Y given Z , then V is d-connected to Y given XZ .

Proof. Suppose X is d-separated from Y given Z . If V is d-separated from Y given XZ , but d-connected to Y given Z , then there is a path U that d-connects V and some Y in Y given Z , but not given XZ . It follows that some non-collider X on U is in X . Hence $U(X,Y)$ d-connects X and Y given Z . \therefore

Lemma 3.3.7: In a directed acyclic graph G , if V is d-connected to Y given XZ , and X is d-separated from Y given Z , then V is d-connected to Y given Z .

Proof. Suppose X is d-separated from Y given Z . If V is d-separated from Y given Z , but d-connected to Y given XZ , then there is a path U that d-connects V and Y given XZ , but not given Z . Some vertex on U is a collider with a descendant in X , but not in Z . Let C be the vertex on U closest to Y that is the source of a directed path to some X in X that contains no member of Z . C is d-connected to Y given Z . If X is on U then $U(X,Y)$ d-connects X and Y

given \mathbf{Z} . If X is not on U , then there is a directed path from C to X that does not contain any member of \mathbf{Z} , and hence X is d-connected to Y given \mathbf{Z} , contrary to our assumption. \therefore

Lemma 3.3.8: In a directed acyclic graph G , if X is d-separated from Y given \mathbf{Z} , and P satisfies the Markov condition for G , then X is independent of Y given \mathbf{Z} .

Proof. We will show if X is d-separated from Y given \mathbf{Z} that $P(Y|XZ) = P(Y|Z)$ by showing that $IV(Y,XZ) = IV(Y,Z)$ and $IP(Y,XZ) = IP(Y,Z)$ and applying lemma 3.3.5.

Suppose that V is in $IV(Y,Z)$. V is d-connected to Y given \mathbf{Z} and has a descendant in YZ . Hence V has a descendant in XYZ . It follows by lemma 3.3.6 that V is d-connected to Y given XZ . Hence V is in $IV(Y,XZ)$.

Suppose then that V is in $IV(Y,XZ)$; we will show that V is also in $IV(Y,Z)$. Because V is in $IV(Y,XZ)$, V is not in XYZ , V has a descendant in XYZ and is d-connected to Y given XZ . Because V is not in XYZ it is not in XZ . By lemma 3.3.7 V is d-connected to Y given \mathbf{Z} . If V has a member X of X as a descendant, but no member of YZ as a descendant then there is a directed path from V to X that contains no member of Y or Z . It follows by lemma 3.3.3 that X is d-connected to Y given \mathbf{Z} , contrary to our hypothesis. Hence V has a member of YZ as a descendant, and is in $IV(Y,Z)$.

Suppose that V is in $IP(Y,Z)$. If V has a parent in Y , then V is in $IP(Y,XZ)$. If V has a parent T in $IV(Y,Z)$ then T is in $IV(Y,XZ)$ because $IV(Y,Z) = IV(Y,XZ)$. Hence V is in $IP(Y,XZ)$.

Suppose that V is in $IP(Y,XZ)$. Because V is in $IP(Y,XZ)$ V is in XZ and has a parent in $IV(Y,XZ) \cup Y$. We have already shown that $IV(Y,XZ) \cup Y = IV(Y,Z) \cup Y$. We will now show that V is not in X . If V is in X and has a member of Y as a parent, then X is d-connected to Y given \mathbf{Z} , contrary to our hypothesis. If V is in X and has some W in $IV(Y,XZ)$ as a parent, then W is in $IV(Y,Z)$. It follows that X is d-connected to Y given \mathbf{Z} , contrary to our hypothesis. Hence V is not in X , and $IP(Y,XZ) = IP(Y,Z)$.

By lemma 3.3.5, $P(Y|XZ) = P(Y|Z)$, and hence X is independent of Y given \mathbf{Z} . \therefore

Lemma 3.3.9: In a directed acyclic graph G , if X is not a descendant of Y , and X and Y are not adjacent, then X and Y are d-separated by $Parents(Y)$.

Proof. (A slight variant of this is stated in Pearl (1989)). Suppose on the contrary that some undirected path U d-connects X and Y given $Parents(X)$. If U is into Y then it contains some

member of $\text{Parents}(Y)$ not equal to X as a non-collider. Hence it does not d-connect X and Y given $\text{Parents}(Y)$, contrary to our assumption. If U is out of Y , then because X is not a descendant of Y , U contains a collider. Let C be the collider on U closest to Y . If U d-connects X and Y given $\text{Parents}(Y)$ then C has a descendant in $\text{Parents}(Y)$. But then C is an ancestor of Y , and Y is an ancestor of C , so G is cyclic, contrary to our assumption. Hence no undirected path between X and Y d-connects X and Y given $\text{Parents}(Y)$. \therefore

Theorem 3.3: $P(V)$ is faithful to directed acyclic graph G with vertex set V if and only if for all disjoint sets of vertices X , Y , and Z , X and Y are independent conditional on Z if and only if X and Y are d-separated given Z .

Proof. \Rightarrow Suppose that P is faithful to G . It follows that P satisfies the Markov condition for G . By lemma 3.3.8 if X and Y are d-separated given Z then X and Y are independent conditional on Z . By lemma 3.5.8 (proved below) there is a distribution P' that satisfies the Markov condition for G such that if X and Y are not d-separated given Z then X and Y are not independent conditional on Z . It follows that if X and Y are not d-separated given Z then the Markov condition does not entail that X and Y independent conditional on Z .

\Leftarrow Suppose that X and Y are independent conditional on Z in P if and only if X and Y are d-separated given Z . It follows from lemma 3.3.9 that that P satisfies the Markov condition for G because $\text{Parents}(V)$ d-separates V from $V \setminus (\text{Descendants}(V) \cup \text{Parents}(V))$. Hence all of the conditional independence relations entailed by the Markov condition are true of P . If the independence of X and Y conditional on Z is not entailed by the Markov condition for G then by lemma 3.5.8 X and Y are not d-separated in G , and X and Y are not independent conditional on Z . It follows that P is faithful to G . \therefore

13.5 Theorem 3.4

Theorem 3.4: If $P(V)$ is faithful to some directed acyclic graph, then $P(V)$ is faithful to directed acyclic graph G with vertex set V if and only if

- (i) for all vertices X, Y of G , X and Y are adjacent if and only if X and Y are dependent conditional on every set of vertices of G that does not include X or Y ; and
- (ii) for all vertices X, Y, Z such that X is adjacent to Y and Y is adjacent to Z and X and Z are not adjacent, $X \rightarrow Y \leftarrow Z$ is a subgraph of G if and only if X, Z are dependent conditional on every set containing Y but not X or Z .

Proof. The theorem follows from a theorem first proved in Verma and Pearl (1990b). \therefore

13.6 Theorem 3.5

Theorem 3.5: Let S be an LCT with directed acyclic graph G over the set of non-error variables V . Then for any two non-error vertices A, B in V and any subset H of $V \setminus \{A, B\}$, G linearly implies that $\rho_{AB,H} = 0$ if and only if A, B are d-separated given H .

The **distributed form** of an expression or equation E is the result of carrying out every multiplication, but no additions, subtractions, or divisions in E . If there are no divisions in an equation then its distributed form is a sum of terms. For example, the distributed form of the equation $u = (a + b)(c + d)v$ is $u = acv + adv + bcv + bdv$. In an LCF or LCT T , if an expression is equal to ce , where c is a non-zero constant, and e is a product of equation coefficients raised to positive integral powers, then e is the **equation coefficient factor**(e.c2.f.) of ce , and c is the **constant factor** (c.f.) of ce .

An acyclic directed graph G over V is an **I-map** of probability distribution $P(V)$ iff for every X, Y , and Z that are disjoint sets of random variables in V , if X is d-separated from Y given Z in G then X is independent of Y given Z in $P(V)$. An acyclic graph G is a **minimal I-map** of probability distribution P iff G is an I-map of P , and no proper subgraph of G is an I-map of P . An acyclic graph G over V is a **D-map** of probability distribution $P(V)$ iff for every X, Y , and Z that are disjoint sets of random variables in V , if X is not d-separated from Y given Z in G then X is not independent of Y given Z in $P(V)$. However, when minimal I-map, I-map, or D-map is applied to the graph in an LCT or LCF, the quantifiers in the definitions apply only to sets of *non-error* variables.

A **trek** $T(I,J)$ between two distinct vertices I and J is an unordered pair of acyclic directed paths from some vertex K to I and J respectively that intersect only at K . The source of the paths in the trek is called the **source** of the trek. I and J are called the **termini of the trek**. Given a trek $T(I,J)$ between I and J , $I(T(I,J))$ will denote the path in $T(I,J)$ from the source of $T(I,J)$ to I and $J(T(I,J))$ will denote the path in $T(I,J)$ from the source of $T(I,J)$ to J . One of the paths in a trek may be an empty path. However, since the termini of a trek are distinct, only one path in a trek can be empty. $T(I,J)$ is the set of all treks between I and J . $T(I,J)$ will represent a trek in $T(I,J)$. $S(T(I,J))$ represents the source of the trek $T(I,J)$.

The proofs of the following two lemmas are trivial.

Lemma 3.5.1: In a directed acyclic graph G , every undirected path $V = \langle V_1, V_2, \dots, V_{n-1}, V_n \rangle$ without colliders contains a vertex V_k such that $\langle V_k, \dots, V_1 \rangle$ and $\langle V_k, \dots, V_n \rangle$ are directed subpaths of V that intersect only at V_k .

Hence, corresponding to each undirected path $V = \langle V_1, V_2, \dots, V_{n-1}, V_n \rangle$ without colliders is a trek $T = (\langle V_k, \dots, V_1 \rangle, \langle V_k, \dots, V_n \rangle)$. When V is a directed path, one of the paths is empty; for example, $V_k = V_1$.

Lemma 3.5.2: In a directed acyclic graph G , for every trek $(\langle V_1, \dots, V_n \rangle, \langle V_1, \dots, V_m \rangle)$, the concatenation of $\langle V_n, \dots, V_1 \rangle$ with $\langle V_1, \dots, V_m \rangle$ is an undirected path from V_n to V_m without colliders.

We will say that a directed acyclic graph has error variables if every vertex of indegree not equal to 0 has an edge into it from a vertex of indegree 0 and outdegree 1. If each independent random variable in an LCT S is normally distributed, then the joint distribution of the set of all random variables in the LCT is multi-variate normal. We will say the random variables in such an LCT have a linear multi-variate normal distribution. The next series of lemmas demonstrate that every directed acyclic graph with error variables is faithful to some LCT S in which the joint distribution Q of the random variables in S is linear multi-variate normal.

Lemma 3.5.3: If S is an acyclic multi-variate normal LCT with directed acyclic graph G' and distribution P , V is the set of non-error terms in S , G is the subgraph of G' over V , and the exogenous variables are jointly independent, then G is a minimal I-map of $P(V)$.

Proof. Let V be the set of non-error terms in S , and G be the subgraph of G' over V . First we will show that if A and B are distinct variables in V , and B is not a descendant of A or a parent of A in G , then A is independent of B given $\text{Parents}(G, A)$. ε_A is normally distributed and uncorrelated with any of the parents of A or B . B is not a linear function of $\text{Parents}(G, A)$ because the distribution is positive. Hence, if we write A as a linear function of $\text{Parents}(G, A)$, B , and ε_A , this is a regression model of A . The coefficient of B in such an equation is zero. The coefficient of B in such a linear equation for A is zero if and only if A and B are independent conditional on $\text{Parents}(G, A)$. (See Whittaker 1990.). Hence B is independent of A given $\text{Parents}(G, A)$. Because the joint distribution is normal, it follows that A is independent of the set of its non-parental non-descendants given its parents. Hence G is an I-map of $P(V)$.

We will now show that $P(V)$ satisfies the Minimality Condition for G . Suppose, on the contrary, that G is not a minimal I-map of $P(V)$. It follows that some some subgraph of G is an I-map of $P(V)$. Let G_{Sub} be a subgraph of G that is an I-map of $P(V)$, and in which the only difference between G and G_{Sub} is that X is a parent of Y in G , but not in G_{Sub} . Because $\text{Parents}(G_{Sub}, Y) \cup \{X\} = \text{Parents}(G, Y)$, when Y is written as a linear function of $\text{Parents}(G_{Sub}, Y), X$, and ε_Y , the coefficient of X is not zero. But because X is not a parent of Y in G_{Sub} , and not a descendant of Y in G_{Sub} , it follows that X and Y are d-separated given $\text{Parents}(G_{Sub}, Y)$. Because G_{Sub} is an I-map of $P(V)$, X and Y are independent given $\text{Parents}(G_{Sub}, Y)$. But this entails that the coefficient of X in the linear equation for Y in terms of $\text{Parents}(G, Y)$ and ε_Y is zero, which is a contradiction. \therefore

Lemma 3.5.4: If a polynomial equation Q in real variables $\langle X_1, \dots, X_n \rangle$ is not an identity, then for every solution a of Q , and for every $\varepsilon > 0$ there is a non-solution b of Q such that $|b - a| < \varepsilon$.

Proof. The proof is by induction on the number n of variables in Q .

Base case: If $n = 1$, then there are only a finite number of solutions of Q . It follows that for every solution a of Q , and for every $\varepsilon > 0$ there is a non-solution b of Q such that $|b - a| < \varepsilon$.

Induction case: Suppose that Q is a polynomial equation in $\langle X_1, \dots, X_n \rangle$, Q is not an identity, and the lemma is true for $n-1$. Take an arbitrary solution $\langle a_1, \dots, a_n \rangle$ of Q . Transform Q into a polynomial equation Q' in X_n by fixing the variables $\langle X_1, \dots, X_{n-1} \rangle$ at the value $\langle a_1, \dots, a_{n-1} \rangle$. There are two cases.

In the first case, Q' is not an identity. Hence, by the induction hypothesis, there is a non-solution of Q' whose distance from a_n is $< \varepsilon$. Let a'_n be this non-solution of Q' . Then $a' = \langle a_1, \dots, a_{n-1}, a'_n \rangle$ is a non-solution of Q , and $|a - a'| < \varepsilon$.

In the second case, Q' is an identity. Rewrite Q so that it is of the form

$$\sum_m Q_m X_n^m$$

where each Q_m is a polynomial in at most X_1, \dots, X_{n-1} .

For each m , the equation $Q_m = 0$ is a polynomial equation in less than n variables. If Q' is an identity, then when terms of the same power of X_n are added together, the coefficient of each

power of X_n is zero. This implies that $\langle a_1, \dots, a_{n-1} \rangle$ is a solution to $Q_m = 0$ for each m . If, for each m , $Q_m = 0$ is an identity, then so is Q ; hence for some m , $Q_m = 0$ is not an identity. For this value of m , by the induction hypothesis, there is a non-solution $\langle a'_1, \dots, a'_{n-1} \rangle$ to $Q_m = 0$ that is less than distance ε from $\langle a_1, \dots, a_{n-1} \rangle$. If $\langle a'_1, \dots, a'_{n-1} \rangle$ is substituted for $\langle X_1, \dots, X_{n-1} \rangle$ in Q , the resulting polynomial equation in X_n is not an identity. This reduces to the first case. \therefore

Lemma 3.5.5: If G' is a subgraph of G , and there is some LCT S' with directed acyclic graph G and distribution P' such that $\rho_{IJ,Z} \neq 0$ in P' , then there is some LCT S containing G and distribution P such that $\rho_{IJ,Z} \neq 0$ in P .

Proof. By lemma 3.2.1 in S' $\rho_{IJ,Z} = 0$ is equivalent to a polynomial equation in the linear coefficients and variances of independent variables in S' . Since there is some LCT S' containing G' such that $\rho_{IJ,Z} \neq 0$ in S' , the polynomial equation is not an identity.

Let S be an LCT with directed acyclic graph G such that for all variables J, I , if the coefficient c' of J in the equation for I in S' is not equal to zero, then the coefficient of J in the equation for I in S is equal to c' . In S , $\rho_{IJ,Z} = 0$ is equivalent to a polynomial equation E in the linear coefficients and variances of independent variables in S . When labels of the edges in G but not in G' are set to zero, the polynomial in E equals the polynomial in E' . No label of an edge in G but not in G' occurs in E' . Hence when the labels of the edges in G but not in G' are set to non-zero values, the polynomial in E contains all of the terms that are in E' and possibly some extra terms. Let us say that two terms in a polynomial equation are **like terms** if they contain the same variables raised to the same powers. Each of the terms that are in E but not E' contain some linear coefficient that does not appear in any term in E' ; hence each of the additional terms in E is not like any term in E' .

If E were an identity, then the sum of the coefficients of like terms in E would be equal to zero. Since E' is not an identity, there are like terms in E' such that the sum of their coefficients is not zero. These same like terms appear in E . Furthermore, since the only additional terms in E that are not in E' are not like any term in E' , it follows that if the sum of the coefficients of like terms in E' is not zero, then the sum of the coefficients of the same like terms in E is not identically zero. Hence E is not identically zero, and there is some LCT S containing G such that $\rho_{IJ,Z} \neq 0$ in S . \therefore

The next lemma states that given a set Z of partial correlations and a directed acyclic graph G , if it is possible to construct a set S of LCTs with directed acyclic graph G such that each Z in Z

fails to vanish for some one of the LCTs in S , then it is possible to construct a single LCT with directed acyclic graph G such that all of the Z in \mathbf{Z} fail to vanish.

Lemma 3.5.6: Given a set of partial correlations \mathbf{Z} and a directed acyclic graph G , if for all Z in \mathbf{Z} there exists an LCT S' with directed acyclic graph G and distribution P' such that $Z \neq 0$ in P' , then there exists a single LCT S with directed acyclic graph G and distribution P such that for all Z in \mathbf{Z} , $Z \neq 0$ in P .

Proof. The proof is by induction on the cardinality of \mathbf{Z} .

Base Case: If the only member of \mathbf{Z} is Z , then by assumption there is an LCT S containing G such that $Z = 0$.

Induction Case: Suppose that the lemma is true for each set of cardinality $n-1$, \mathbf{Z} is of cardinality n , and for each Z_i in Z , there is an LCT S' with directed acyclic graph G and distribution P' such that $Z_i \neq 0$ in P' . By the induction hypothesis, there is an LCT S with directed acyclic graph G and distribution P such that $Z_i \neq 0$, $i \leq 1 \leq n-1$. Let V be a set of values for the linear coefficients and variances of independent variables such that $Z_i \neq 0$, $i \leq 1 \leq n-1$. The valuation V either makes Z_n equal to zero or it doesn't. If it doesn't, then the proof is done. If it does, we will show how to perturb V by a small amount to make $Z_n \neq 0$, while keeping each $Z_i \neq 0$, $i \leq 1 \leq n-1$.

By lemma 3.2.1, each of the partial correlations in Z_i in \mathbf{Z} is equivalent to a polynomial Q_i in the linear coefficients and the variances of independent variables in G . Suppose that the smallest non-zero value for any of the Q_i under the valuation V is δ . By lemma 3.5.4, for arbitrarily small ε there is a non-solution V' to $Z_n = 0$ within distance ε of V . Choose an ε small enough so that the largest possible change in any of the Q_i is less than δ . For the valuation V' then $Z_i \neq 0$, $i \leq 1 \leq n$. ∴

Recall that if a graph with error variables is a D-map of some distribution P , then we consider only dependencies among the non-error variables.

Lemma 3.5.7: For every directed acyclic graph G with error variables, there is an LCT S with directed acyclic graph G and joint linear multi-variate normal distribution Q , such that G is a D-map of Q .

Proof. In order to show that G is a D-map of Q , we must show that for all disjoint sets of variables \mathbf{X} , \mathbf{Y} , and \mathbf{Z} , if \mathbf{X} and \mathbf{Y} are not d-separated in G , then \mathbf{X} is not independent of \mathbf{Y} given \mathbf{Z} in Q . In a linear multi-variate normal distribution, if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint sets of

variables, then $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$ iff $X \perp\!\!\!\perp Y | Z$ for each X in \mathbf{X} and Y in \mathbf{Y} ; similarly if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint sets of variables then \mathbf{X} and \mathbf{Y} are d-separated given \mathbf{Z} iff for all X in \mathbf{X} and Y in \mathbf{Y} , X and Y are d-separated given \mathbf{Z} . Hence, we need consider only dependency statements of the form X and Y are not independent given \mathbf{Z} , where X and Y are individual variables. Also in a linear multi-variate normal distribution, $\rho_{XY|\mathbf{Z}} = 0$ iff $X \perp\!\!\!\perp Y | \mathbf{Z}$. So it suffices to prove that there is an LCT S with directed acyclic graph G and distribution P such that for each X , Y , and Z in G such that X and Y are not d-separated given \mathbf{Z} in G , $\rho_{XY|\mathbf{Z}} \neq 0$ in P . The proof is by induction. We assume that in all of the LCTs constructed, the independent random variables are normally distributed.

Base Case. If \mathbf{Z} is empty, then by lemma 3.5.1, X and Y are not d-separated given \mathbf{Z} iff there is a trek connecting them. Form a subgraph G' and a sub-LCT S' with directed acyclic graph G' and distribution P' , such that there is exactly one trek between X and Y . It was proved in Glymour et al. (1987) that in this case the covariance between X and Y is equal to the product of the labels of the edges in the trek (the linear coefficients) times the variance of the source of the trek. If each of these quantities is non-zero, so is the covariance, and also the correlation in P' . By lemma 3.5.5 if ρ_{XY} is not identically zero in S' it is also not identically zero in some LCT S with directed acyclic graph G . By lemma 3.5.6 there exists a LCT containing G in which for all X and Y , if X and Y are not d-separated by the empty set then the correlation between X and Y is not zero.

Induction Case. Suppose that there is an LCT S with directed acyclic graph G and distribution P such that for each X , Y , and for each \mathbf{A} of cardinality less than n that does not contain X or Y , such that X and Y are not d-separated given \mathbf{A} in G , $\rho_{XY|\mathbf{A}} \neq 0$ in P . Let \mathbf{Z} be of cardinality n . Suppose that X and Y are not d-separated by \mathbf{Z} in G . It follows that there is an undirected path U between X and Y such that every vertex without a collider is not in \mathbf{Z} , and every vertex V_i on U that is a collider is the source of a directed path U_i from V_i to a variable in \mathbf{Z} . Form a subgraph G' , such that G' contains only the undirected path U , one directed path U_i from each collider V_i on U , the vertices in those paths, and the vertices in \mathbf{Z} . Shorten each U_i so that it contains only one variable in \mathbf{Z} . Finally, if two variables V_n and V_m that are colliders on U are the sources of directed paths U_n and U_m that intersect, let F be the first point of intersection of U_n and U_m . Replace the subpath of U from V_n to V_m by the concatenation of the subpaths of $U_n(V_n, F)$ and $U_m(F, V_m)$, and replace U_n and U_m by $U_n(F, Z)$, where Z is in \mathbf{Z} . The new path has one fewer collider than the old path. Repeat this process until none of the U_i intersect each other or there are no colliders on U . There are two cases.

In the first case, U contains no vertices with a collider, and hence no vertices in \mathbf{Z} . By lemma 3.5.1 there is a trek between X and Y that contains no vertices in \mathbf{Z} . Let R be an arbitrary vertex in \mathbf{Z} , and $\mathbf{W} = \mathbf{Z} \setminus \{R\}$. There is a trek between X and Y that contains no vertices in \mathbf{W} . It follows that \mathbf{W} does not d-separate X and Y , so by the induction hypothesis, there is an LCT with directed acyclic graph G' and distribution P' such that $\rho_{XY,W} \neq 0$. It follows from lemma 3.5.3 that in P' that $\rho_{XR,W} = 0$ and $\rho_{YR,W} = 0$ because by construction there are no undirected paths from X to R or Y to R . By the recursion formula for partial correlation, $\rho_{XY,W} = 0$ iff $\rho_{XY,W} = \rho_{XR,W} \times \rho_{YR,W}$. But $\rho_{XY,W}$ is non-zero in P' , and $\rho_{XR,W} \times \rho_{YR,W}$ is zero in P' . Hence $\rho_{XY,Z} \neq 0$ in P' . By lemma 3.5.5, there is some LCT S'' with directed acyclic graph G and distribution P'' such that $\rho_{XY,Z} \neq 0$ in P'' .

In the second case, U contains vertices with colliders, but every vertex that is not a collider is not in \mathbf{Z} . See figure 1.

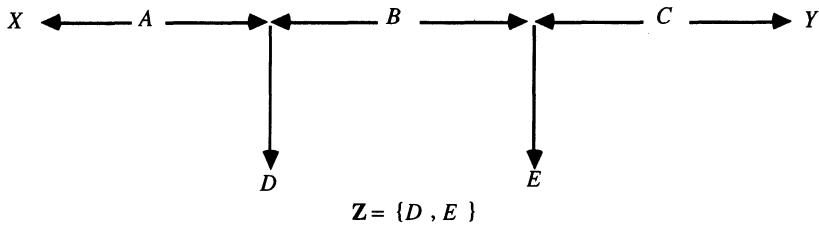


Figure 1

Let E be the vertex that is the sink of the directed path from the collider closest to Y on U , and $\mathbf{W} = \mathbf{Z} \setminus \{E\}$. Since by construction there is a trek between Y and E that does not contain any variables in \mathbf{W} , Y and E are not d-separated by \mathbf{W} . There is also an undirected path from X to E such that every vertex that is not a collider is not in \mathbf{W} , and every vertex that does contain a collider has a descendant in \mathbf{W} . Hence X and E are not d-separated by \mathbf{W} . By the induction hypothesis, there is an LCT S' with directed acyclic graph G' and distribution P' such that $\rho_{XE,W} \neq 0$, and $\rho_{YE,W} \neq 0$ in P' .

On the other hand, since path U was constructed so that each vertex that is a collider has only one descendant in \mathbf{Z} , and \mathbf{W} does not contain E , X and Y are d-separated by \mathbf{W} . Hence by lemma 3.5.3 $\rho_{XY,W} = 0$ in P' .

$\rho_{XY.W} = 0$ iff $\rho_{XY.W} = \rho_{XE.W} \times \rho_{YE.W}$. Since $\rho_{XY.W} = 0$, while $\rho_{XE.W} \times \rho_{YE.W} \neq 0$, $\rho_{XY.Z} \neq 0$ in P' . By lemma 3.5.5, there is an LCT S'' with directed acyclic graph G and distribution P'' such that $\rho_{XY.Z} \neq 0$ in P'' .

Since for each triple X, Y, Z such that X and Y are not d-separated given Z in G there is an LCT S' with directed acyclic graph G and distribution P' such that $\rho_{XY.Z} \neq 0$ in P' , by lemma 3.5.6 there is an LCT S'' with directed acyclic graph G and distribution P'' such that for each triple X, Y, Z for which X and Y are not d-separated given Z in G , $\rho_{XY.Z} \neq 0$ in P'' . Because the LCTs constructed in lemmas 3.5.5 and 3.5.6 don't change the normality of the independent variables, the joint distribution of the random variables in S is linear multi-variate normal. Hence there is an LCT S such that Q is a linear multi-variate normal distribution and G is a D-map of Q . \therefore

Lemma 3.5.8: For every directed acyclic graph G with error variables, there is an LCT S containing G with a linear multivariate normal distribution Q such that G is faithful to Q .

Proof. This follows immediately from lemmas 3.5.7 and 3.5.3. \therefore

The next theorem states that the d-separability relations between sets of non-error variables can be determined from a subgraph that does not include error terms.

Lemma 3.5.9: In an acyclic LCT S with directed acyclic graph G , let G' be the subgraph of G over the non-error variables. Given three disjoint sets X , Y , and Z of non-error variables, X is d-separated from Y given Z in G iff X is d-separated from Y given Z in G' .

Proof. If an error variable occurs on an undirected path, then that error variable is either the source or the sink of the undirected path. Hence, error variables do not occur on any undirected path between non-error variables. It follows that the undirected paths in G and G' between non-error variables are exactly the same. The lemma then follows from the definition of d-separability. \therefore

A directed acyclic graph G **linearly implies** $\rho_{AB.H} = 0$ if and only if $\rho_{AB.H} = 0$ in all distributions linearly represented by G . (We assume all partial correlations exist for the distribution.) Kiiveri and Speed (1982) explicitly notes the connection between the Markov Condition and zero partial correlations.

Lemma 3.5.10: In an LCT S with directed acyclic graph G over the set of non-error variables V and the distribution $P(V)$, if Y d-separates X and Z , then S linearly implies that $\rho_{XZ,Y} = 0$.

Proof. Suppose Y d-separates X and Z in G . The values of the partial correlations in $P(V)$ are completely determined by the values of the linear coefficients and the variances of the independent variables. Consider a multi-variate normal distribution $P'(V)$ in the LCT with the same linear coefficients and the same variances of independent variables as S , but in which the independent variables are normally distributed and jointly independent. By lemma 3.5.3, G is an I-map of $P'(V)$, and because Y d-separates X and Z , $X \perp\!\!\!\perp Z|Y$ in $P'(V)$. Because $P'(V)$ is a multi-variate normal distribution, $X \perp\!\!\!\perp Z|Y$ if and only $\rho_{XZ,Y} = 0$. It follows that $\rho_{XZ,Y} = 0$ in $P'(V)$, and hence $\rho_{XZ,Y} = 0$ in $P(V)$. \therefore

Theorem 3.5: Let S be an LCT with directed acyclic graph G over the set of non-error variables V . Then for any two non-error vertices A, B in V and any subset H of $V \setminus \{A, B\}$, G linearly implies that $\rho_{AB,H} = 0$ if and only if A, B are d-separated given H .

Proof. The if clause follows from Lemma 3.5.10.

The only if clause follows from lemma 3.5.7. By lemma 3.5.7 there is an LCT S such that Q , the joint distribution of the random variables is linear multi-variate normal, and G is a D-map of Q . In S , if A and B are not d-separated given H , then A and B are not independent given H , and $\rho_{AB,H} \neq 0$. Hence if A and B are not d-separated given H , G does not linearly imply that $\rho_{AB,H} = 0$. \therefore

Corollary 3.5.1: In an LCT $S = \langle G, (\Omega, f, P), EQ, L \rangle$ in which the exogenous variables are jointly independent, if X and Z are distinct non-error variables, and Y is a set of non-error variables not including X and Z , if $\rho_{XZ,Y}$ is linearly implied to vanish then $X, Z \perp\!\!\!\perp Y$.

Corollary 3.5.2: In an LCT $S = \langle G, (\Omega, f, P), EQ, L \rangle$, if P is faithful to G , X and Z are distinct non-error variables, and Y is a set of non-error variables not including X and Z , G linearly implies that $\rho_{XZ,Y} = 0$ if and only if $X \perp\!\!\!\perp Z|Y$.

13.7 Theorem 3.6 (Manipulation Theorem)

Theorem 3.6 (Manipulation Theorem): Given directed acyclic graph G_{Comb} over vertex set $\mathbf{V} \cup \mathbf{W}$ and distribution $P(\mathbf{V} \cup \mathbf{W})$ that satisfies the Markov condition for G_{Comb} , if changing the value of \mathbf{W} from w_1 to w_2 is a manipulation of G_{Comb} with respect to \mathbf{V} , G_{Unman} is the unmanipulated graph, G_{Man} is the manipulated graph, and

$$P_{Unman}(\mathbf{W})(\mathbf{V}) = \prod_{X \in \mathbf{V}} P_{Unman}(\mathbf{W})(X | \text{Parents}(G_{Unman}, X))$$

for all values of \mathbf{V} for which the conditional distributions are defined, then

$$\begin{aligned} P_{Man}(\mathbf{W})(\mathbf{V}) = & \\ \prod_{X \in \text{Manipulated}(\mathbf{W})} & P_{Man}(\mathbf{W})(X | \text{Parents}(G_{Man}, X)) \times \\ \prod_{X \in \mathbf{V} \setminus \text{Manipulated}(\mathbf{W})} & P_{Unman}(\mathbf{W})(X | \text{Parents}(G_{Unman}, X)) \end{aligned}$$

for all values of \mathbf{V} for which each of the conditional distributions is defined.

If G is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, and $\mathbf{V} \cap \mathbf{W} = \emptyset$, then \mathbf{W} is **exogenous with respect to \mathbf{V}** in G if and only if there is no directed edge from any member of \mathbf{V} to any member of \mathbf{W} . If G_{Comb} is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, and $P(\mathbf{V} \cup \mathbf{W})$ satisfies the Markov condition for G_{Comb} , then changing the value of \mathbf{W} from w_1 to w_2 is a **manipulation** of G_{Comb} with respect to \mathbf{V} if and only if \mathbf{W} is exogenous with respect to \mathbf{V} , and $P(\mathbf{V} | \mathbf{W} = w_1) \neq P(\mathbf{V} | \mathbf{W} = w_2)$.

We define $P_{Unman}(\mathbf{W})(\mathbf{V}) = P(\mathbf{V} | \mathbf{W} = w_1)$, and $P_{Man}(\mathbf{W})(\mathbf{V}) = P(\mathbf{V} | \mathbf{W} = w_2)$, and similarly for various marginal and conditional distributions formed from $P(\mathbf{V})$.

We refer to G_{Comb} as the **combined graph**, and the subgraph of G_{Comb} over \mathbf{V} as the **unmanipulated graph** G_{Unman} .

V is in **Manipulated(\mathbf{W})** (that is, V is a variable directly influenced by one of the manipulation variables) if and only if V is in **Children(\mathbf{W})** \cap \mathbf{V} ; we will also say that the variables in

Manipulated(W) have been **directly manipulated**. We will refer to the variables in **W** as **policy variables**.

The **manipulated graph**, G_{Man} is a subgraph of G_{Unman} for which $P_{Man(W)}(V)$ satisfies the Markov Condition and which differs from G_{Unman} in at most the parents of members of **Manipulated(W)**.

Lemmas 3.6.1 and 3.6.2 show that distributions satisfying the antecedent of Theorem 3.6 exist.

In a directed acyclic graph G over V , X is in **Non-Descendants**(G, Y) if and only if X is in V and there is no directed path from any member of Y to X in G .

Lemma 3.6.1: Given directed acyclic graph G_{Comb} over vertex set $V \cup W$ and distribution $P(V \cup W)$ that satisfies the Markov condition for G , if changing the value of W from w_1 to w_2 is a manipulation of G_{Comb} with respect to V , and G_{Unman} is the unmanipulated graph, then $P_{Unman(W)}(V)$ satisfies the Markov Condition for G_{Unman} .

Proof. $P_{Unman(W)}(V)$ satisfies the Markov Condition for G_{Unman} if for each vertex V in V , V is independent of **Non-Descendants**(G_{Unman}, V)\Parents(G_{Unman}, V) conditional on Parents(G_{Unman}, V) $\cup W$. Suppose that on the contrary that for some V in V , V is dependent on **Non-Descendants**(G_{Unman}, V)\Parents(G_{Unman}, V) conditional on Parents(G_{Unman}, V) $\cup W$. It follows that there is some path U in G_{Comb} that d-connects V and some member X in **Non-Descendants**(G_{Unman}, V) given Parents(G_{Unman}, V) $\cup W$. Every member of W that occurs on U is a collider on U because U d-connects X and V given Parents(G_{Unman}, V) $\cup W$. Because W is exogenous to V , U contains no member of W . It follows that no collider on U has a descendant in W . Hence U d-connects V and X given Parents(G_{Unman}, V) in G_{Comb} . The path corresponding to U in G_{Unman} also d-connects V and X given Parents(G_{Unman}, V). But this contradicts lemma 3.3.9. ∴

Lemma 3.6.2: Given directed acyclic graph G_{Comb} over vertex set $V \cup W$ and distribution $P(V \cup W)$ that satisfies the Markov condition for G_{Comb} , if changing the value of W from w_1 to w_2 is a manipulation of G_{Comb} with respect to V , and G_{Unman} is the unmanipulated graph, then $P_{Man(W)}(V)$ satisfies the Markov Condition for some subgraph of G_{Unman} .

Proof. The proof that $P_{Man(W)}(V)$ satisfies the Markov Condition for G_{Unman} is essentially the same as that of lemma 3.6.1. Because G_{Unman} is an (improper) subgraph of itself, $P_{Man(W)}(V)$ satisfies the Markov Condition for some subgraph of G_{Unman} .

Theorem 3.6 (Manipulation Theorem): Given directed acyclic graph G_{Comb} over vertex set $V \cup W$ and distribution $P(V \cup W)$ that satisfies the Markov condition for G_{Comb} , if changing the value of W from w_1 to w_2 is a manipulation of G_{Comb} with respect to V , G_{Unman} is the unmanipulated graph, G_{Man} is the manipulated graph, and

$$P_{Unman(W)}(V) = \prod_{X \in V} P_{Unman(W)}(X | \text{Parents}(G_{Unman}, X))$$

for all values of V for which the conditional distributions are defined, then

$$\begin{aligned} P_{Man(W)}(V) = \\ \prod_{X \in \text{Manipulated}(W)} P_{Man(W)}(X | \text{Parents}(G_{Man}, X)) \times \\ \prod_{X \in V \setminus \text{Manipulated}(W)} P_{Unman(W)}(X | \text{Parents}(G_{Unman}, X)) \end{aligned}$$

for all values of V for which each of the conditional distributions is defined.

Proof. By assumption, $P_{Man(W)}(V)$ satisfies the Markov Condition for G_{Man} . Hence

$$\begin{aligned} P_{Man(W)} = \prod_{X \in V} P(X | \text{Parents}(G_{Man}, X)) = \\ \prod_{X \in \text{Manipulated}(W)} P(X | \text{Parents}(G_{Man}, X)) \times \prod_{X \in V \setminus \text{Manipulated}(W)} P(X | \text{Parents}(G_{Man}, X)) \end{aligned}$$

for all values of V for which the conditional distributions exist. No member of W is a descendant of any variable in V in G_{Comb} , so for each V in $V \setminus \text{Manipulated}(W)$, W is d-separated from V given $\text{Parents}(G_{Comb}, V)$ in G_{Comb} . For any member X of $V \setminus \text{Manipulated}(W)$, $\text{Parents}(G_{Comb}, X) = \text{Parents}(G_{Unman}, X) = \text{Parents}(G_{Man}, X)$. It follows that $P(V | \text{Parents}(G_{Man}, X), W = w_2) = P(V | \text{Parents}(G_{Man}, X)) = P(V | \text{Parents}(G_{Man}, X), W = w_1) = P(V | \text{Parents}(G_{Unman}, X), W = w_1)$. Hence

$$\begin{aligned} P_{Man(W)}(V) = \\ \prod_{X \in \text{Manipulated}(W)} P_{Man(W)}(X | \text{Parents}(G_{Man}, X)) \times \prod_{X \in V \setminus \text{Manipulated}(W)} P_{Unman(W)}(X | \text{Parents}(G_{Unman}, X)) \end{aligned}$$

for all values of V for which the conditional distributions are defined. \therefore

13.8 Theorem 3.7

Theorem 3.7: If G is a directed acyclic graph over V , X , Y , and Z are disjoint subsets of V , and $P(V)$ satisfies the Markov condition for G and the deterministic relations in **Deterministic(V)** then if X and Y are D-separated given Z and **Deterministic(V)**, X and Y are independent given Z in P .

We will say that a set of variables Z **determines** the set of variables A , when every variable in A is a deterministic function of the variables in Z , and not every variable in A is a deterministic function of any proper subset of Z . Suppose G is a directed acyclic graph over V , and **Deterministic(V)** is a set of ordered tuples of variables in V , where for each tuple D in **Deterministic(V)**, if D is $\langle V_1, \dots, V_n \rangle$ then V_n is a deterministic function of V_1, \dots, V_{n-1} and is not a deterministic function of any subset of V_1, \dots, V_{n-1} ; we also say $\{V_1, \dots, V_{n-1}\}$ **determines** V_n . For a given **Deterministic(V)**, if Z is included in V , then **Det(Z)** is the set of variables determined by any subset of Z . Note that Z is included in **Det(Z)**.

If G is a directed acyclic graph over V , and Z is included in V , then G' is in **Mod(G)** relative to **Deterministic(V)** and Z if and only if for each V in V

- (i) if there exists a set of vertices included in Z that are non-descendants of V in G and that determine V , then **Parents**(G', V) = X , where X is some set of vertices included in Z that are non-descendants of V in G and that determine V ;
- (ii) if there is no set X of vertices included in Z that are non-descendants of V in G and that determine V , then **Parents**(G', V) = **Parents**(G, V).

If G is a directed acyclic graph with vertex set V , Z is a set of vertices not containing X or Y , and $X \neq Y$, then X and Y are **D-separated** given Z and **Deterministic(V)** if and only if there is no undirected path U in G between X and Y such that each collider on U has a descendant in Z , and no other vertex on U is in **Det(Z)**; otherwise if $X \neq Y$ and X and Y are not in Z , then X and Y are **D-connected** given Z and **Deterministic(V)**. Similarly, if X , Y , and Z are disjoint sets of variables, and X and Y are non-empty, then X and Y are D-separated given Z and **Deterministic(V)** if and only if each pair $\langle X, Y \rangle$ in the Cartesian product of X and Y are **D-separated** given Z and **Deterministic(V)**; otherwise if X , Y , and Z are disjoint, and X and Y are non-empty, then X and Y are **D-connected** given Z and **Deterministic(V)**.

If G is a directed acyclic graph over V , Z is a subset of V that does not contain X or Y , and $X \neq Y$, then X and Y are **det-separated** given Z and **Deterministic(V)** if and only if either X and Y are d-separated given $Z \cup \text{Det}(Z)$ in some **Mod**(G) relative to **Deterministic(V)** and Z , or X or Y is in **Det(Z)**; otherwise if $X \neq Y$ and X and Y are not in Z , then X and Y are **det-connected** given Z and **Deterministic(V)**. If X, Y and Z are disjoint sets of variables in V , and X and Y are non-empty, then X and Y are **det-separated** given Z if and only if every member X of X and every member Y of Y are det-separated given Z ; otherwise if X, Y and Z are disjoint sets of variables in V , and X and Y are non-empty, then X and Y are **det-connected** given Z and **Deterministic(V)**.

Lemma 3.7.1: Let G be a directed acyclic graph with vertex set V , Ord an ordering of variables in V such that if A is before B in Ord then A is not a descendant of B in G , **Predecessors**(Ord, V) the set of all vertices before V in Ord , and $P(V)$ a distribution over V . $P(V)$ satisfies the Minimality and Markov Conditions for G if and only if for each V in V , V is independent of **Predecessors**(Ord, V)**Parents**(G, V) given **Parents**(G, V) and for no proper subset $X(V)$ of **Parents**(G, V), V is independent of **Predecessors**(Ord, V) $X(V)$ given $X(V)$.

Proof. See Pearl(1988). \therefore

Lemma 3.7.2: If G is a directed acyclic graph over V , and X, Y , and Z are disjoint subsets of V , and $P(V)$ satisfies the Markov condition for G and the deterministic relations in **Deterministic(V)**, then if X and Y are det-separated given Z and **Deterministic(V)**, X and Y are independent given Z in P .

Proof. First we will prove that $P(V)$ satisfies the Markov condition for each directed acyclic graph G' in **Mod**(G). First form an acceptable ordering Ord of the variables in V for G . Let **Predecessors**(Ord, V) be the variables that precede V in Ord . From lemma 3.7.1 it follows that if G' is a directed acyclic graph in which for each V in V , V is independent of **Predecessors**(V)**Parents**(V) given **Parents**(V), then G' is an I-map of $P(V)$. If X is a subset of **Parents**(V) that determines V , it follows that V is independent of **Predecessors**($V \setminus X$) given X . Hence if in G' **Parents**(V) = X , G' is still an I-map of $P(V)$.

If either X or Y is included in **Det(Z)**, it follows that X and Y are independent given $Z \cup \text{Det}(Z)$. Suppose then that neither X nor Y is included in **Det(Z)**. By definition of det-separability, $X \setminus \text{Det}(Z)$ and $Y \setminus \text{Det}(Z)$ are d-separated given $Z \cup \text{Det}(Z)$. Hence

$$P((X \cup Y) \setminus \text{Det}(Z) | Z \cup \text{Det}(Z)) = P(X \setminus \text{Det}(Z) | Z \cup \text{Det}(Z))P(Y \setminus \text{Det}(Z) | Z \cup \text{Det}(Z))$$

It now follows that X is independent of Y given Z because

$$\begin{aligned}
 P(X \cup Y | Z) &= P(X \cup Y | Z \cup \text{Det}(Z)) = P((X \cup Y) \setminus \text{Det}(Z) | Z \cup \text{Det}(Z)) = \\
 P(X \setminus \text{Det}(Z) | Z \cup \text{Det}(Z))P(Y \setminus \text{Det}(Z) | Z \cup \text{Det}(Z)) = \\
 P(X | Z \cup \text{Det}(Z))P(Y | Z \cup \text{Det}(Z)) &= P(X | Z)P(Y | Z)
 \end{aligned}$$

∴

Theorem 3.7: If G is a directed acyclic graph over V , X , Y , and Z are disjoint subsets of V , and $P(V)$ satisfies the Markov condition for G and the deterministic relations in $\text{Deterministic}(G)$ then if X and Y are D-separated given Z and $\text{Deterministic}(V)$, X and Y are independent given Z in P .

Proof. We will prove that if X and Y are det-connected given Z and $\text{Deterministic}(V)$, then X and Y are D-connected given Z and $\text{Deterministic}(V)$. It follows then that if X and Y are D-separated given Z and $\text{Deterministic}(V)$, then X and Y are det-separated given Z and $\text{Deterministic}(V)$, and by lemma 3.7.1, X and Y are independent given Z in P .

Suppose some X in X is det-connected to some Y in Y given Z and $\text{Deterministic}(V)$. It follows by definition that X and Y are not in Z and not in $\text{Det}(Z)$. Because X and Y are det-connected given Z there is an undirected path U' that d-connects X and Y given Z in some graph G' in $\text{Mod}(G)$.

First, we will show that the path U corresponding to U' exists in G ; then we will show that U D-connects X and Y given Z and $\text{Deterministic}(V)$ in G .

No member of $\text{Det}(Z)$ is a non-collider on U' because U' d-connects X and Y given $Z \cup \text{Det}(Z)$. Hence for each non-collider A on U' , $\text{Parents}(G', A)$ equals $\text{Parents}(G, A)$. It follows that if there is an edge into A in G' , there is a corresponding edge into A in G .

Suppose then that A is a collider on U' . If there is an edge into A in G' that does not exist in G , then every parent of A is in Z . It follows that either the endpoints of U' are in Z , or some non-collider on U' is in Z . But then U' does not d-connect X and Y given $Z \cup \text{Det}(Z)$ in G' . Hence if there is an edge into A on U' , then the corresponding edge exists in G .

It follows that the path U in G corresponding to U' in G' exists.

The endpoints of U are not in $Z \cup \text{Det}(Z)$, because they are equal to the endpoints of U' , which are not in $Z \cup \text{Det}(Z)$.

No non-collider on U is in $\mathbf{Z} \cup \mathbf{Det}(\mathbf{Z})$, because each non-collider on U is a non-collider on U' , and no non-collider on U' is in $\mathbf{Z} \cup \mathbf{Det}(\mathbf{Z})$.

Finally suppose that A is a collider on U' . It follows that A has a descendant in $\mathbf{Z} \cup \mathbf{Det}(\mathbf{Z})$ in G' . There are two cases.

If A has a descendant in \mathbf{Z} in G' , then it has a descendant in \mathbf{Z} in G . Suppose that A has a descendant X in \mathbf{Z} in G , and let $D(A,X)$ be a directed path from A to X in G' . Let Z be the member of \mathbf{Z} closest to A on $D(A,X)$. Every edge that is in G' but not in G is out of a member of \mathbf{Z} . $D(A,Z)$ has no edges out of a member of \mathbf{Z} . Hence every edge in $D(A,Z)$ exists in G , and A has a descendant in \mathbf{Z} in G .

Suppose A does not have a descendant in \mathbf{Z} in G' . It follows that there is a directed path $D(A,X)$ from A to a member X of $\mathbf{Det}(\mathbf{Z}) \setminus \mathbf{Z}$ in G' . If A itself is in $\mathbf{Det}(\mathbf{Z})$ then it has parents not in \mathbf{Z} , because U' d-connects X and Y given $\mathbf{Z} \cup \mathbf{Det}(\mathbf{Z})$. Because G' is in $\mathbf{Mod}(G)$, it follows from the fact that A has a parent not in \mathbf{Z} that A has a descendant in \mathbf{Z} in G . If A is not in $\mathbf{Det}(\mathbf{Z})$ then $D(A,X)$ is not an empty path, and it does not contain any member of \mathbf{Z} . Hence X has a parent that is not in \mathbf{Z} . Because G' is in $\mathbf{Mod}(G)$, it follows from the fact that X has a parent not in \mathbf{Z} that X has a descendant in \mathbf{Z} in G . $D(A,X)$ exists in G because every edge in G' but not in G is out of a member of \mathbf{Z} , and $D(A,X)$ contains no member of \mathbf{Z} . Hence A has a descendant in \mathbf{Z} in G .

It follows that U D-connects X and Y given \mathbf{Z} and **Deterministic(V)** in G .

13.9 Theorem 4.1

Theorem 4.1: Two directed acyclic graphs G_1, G_2 , are strongly statistically indistinguishable if and only if (i) they have the same vertex set V , (ii) vertices V_1 and V_2 are adjacent in G_1 if and only if they are adjacent in G_2 , and (iii) for every triple V_1, V_2, V_3 in V , the graph $V_1 \rightarrow V_2 \leftarrow V_3$ is a subgraph of G_1 if and only if it is a subgraph of G_2 .

Proof. \Leftarrow Suppose two directed acyclic graphs G_1 and G_2 contain the same vertices, the same adjacencies and the same colliders, and G_1 is a minimal I-map of P . By Theorem 3.4 the same distributions are faithful to G_1 and G_2 so they have the same d-separability relations, and hence G_2 is also an I-map of P .

G_2 is also minimal. Every subgraph of G_1 has the same d-separability relations as does the corresponding subgraph of G_2 because removing corresponding vertices and adjacencies from both graphs leaves subgraphs that contain the same vertices, adjacencies and colliders. Hence, if a subgraph of G_2 is an I-map of P , then the corresponding subgraph of G_1 is an I-map of P . But by supposition, no proper subgraph of G_1 is an I-map of P . Hence no proper subgraph of G_2 is an I-map of P . By definition, G_2 is a minimal I-map of P . It follows that G_1 and G_2 are s.s.i.

⇒ Now consider the case where G_1 and G_2 differ either in their sets of vertices, their adjacencies, or their colliders. We will show that there exists a distribution P such that G_1 is a minimal I-map of P , while G_2 is not. By definition, it follows that G_1 and G_2 are not s.s.i.

Case 1. Suppose first that G_1 and G_2 differ in their sets of vertices. By definition they are not s.s.i.

Case 2. Suppose that G_1 and G_2 differ in their adjacencies. Suppose without loss of generality that G_1 contains an adjacency not in G_2 . Then there is a pair of vertices X and Y such that X and Y are d-separated given a subset S in G_2 , while X and Y are not d-separated given S in G_1 . There is a distribution P faithful to G_1 . G_1 is also a minimal I-map of P . In G_1 , X and Y are dependent conditional on S . But because X and Y are d-separated given a subset S in G_2 , G_2 is not an I-map of P . Hence G_1 and G_2 are not s.s.i.

Case 3. Suppose that G_1 and G_2 differ in their unshielded colliders but not in any adjacencies. Let Y be an unshielded collider on the path $\langle X, Y, Z \rangle$ in G_1 , but not in G_2 . Let P be a distribution faithful to G_1 . It follows that G_1 is a minimal I-map of P . In G_2 , X and Z are d-separated given a set S containing Y , while in G_1 X and Z are not d-separated given S . Since G_1 is faithful to P , X and Z are dependent conditional on S . Hence G_2 is not a minimal I-map of P , and G_1 and G_2 are not s.s.i.

Case 4. Finally, suppose that G_1 and G_2 differ in their shielded colliders but not in any adjacencies or unshielded colliders. Let Y be a shielded collider on the path $\langle X, Y, Z \rangle$ in G_1 , but not in G_2 . Suppose G_2' is the subgraph of G_2 with the edge between X and Z removed. G_2' is faithful to some distribution P . G_2 is not a minimal I-map of P (because it contains a subgraph which is an I-map of P). We will now show that G_1 is a minimal I-map of P .

First, G_1 is an I-map of P . G_1 is f.i. to G_2 . G_2 is a proper supergraph of G_2' , and so the d-separation relations true of G_2 are included in the d-separation relations true of G_2' ; hence the d-separation relations true of G_1 are included in the d-separation relations true of G_2' . It follows that G_1 is an I-map of P .

G_1 is also minimal. If G_1' is a subgraph obtained by deleting from G_1 any edge other than the $X - Z$ edge, by Case 2, the subgraph is not an I-map of P . If G_1' is a subgraph obtained by deleting from G_1 just the $X - Z$ edge, then G_1' contains an unshielded collider at Y that does not occur in G_2' . By Case 3, G_1' is not an I-map of P .

Because G_1 is a minimal I-map of P , and G_2 is not, G_1 and G_2 are not s.s.i. \therefore

13.10 Theorem 4.2

Theorem 4.2: Two directed acyclic graphs G and H are faithfully indistinguishable if and only if (i) they have the same vertex set, (ii) any two vertices are adjacent in G if and only if they are adjacent in H , and (iii) any three vertices, X, Y, Z , such that X is adjacent to Y and Y is adjacent to Z but X is not adjacent to Z in G or H , are oriented as $X \rightarrow Y \leftarrow Z$ in G if and only if they are so oriented in H .

Proof. This was proved in Verma and Pearl (1990b).

It also follows directly from Theorem 3.4. \therefore

13.11 Theorem 4.3

Theorem 4.3: Two directed acyclic graphs are faithfully indistinguishable if and only if some distribution faithful to one is faithful to the other and conversely; i.e. they are f.i. if and only if they are w.f.i.

Proof. Suppose G_1 and G_2 are f.i. By lemma 3.5.8 there is some distribution P faithful to G_1 . Hence P is faithful to G_2 , and G_1 and G_2 are w.f.i.

Suppose that G_1 and G_2 are w.f.i. Then there is some distribution P faithful to G_1 and G_2 . It follows that G_1 and G_2 have the same d-separation relations, so any distribution faithful to G_1 is also faithful to G_2 and vice-versa. \therefore

13.12 Theorem 4.4

Theorem 4.4: If probability distribution P satisfies the Markov Condition for directed acyclic graphs G and H , and P is faithful to H , then for all vertices X, Y , if X, Y are adjacent in H they are adjacent in G .

Proof. If P is faithful to H then X is adjacent to Y in H only if X, Y are dependent conditional on every set of vertices not containing X or Y . Suppose then that P satisfies the Markov condition for G but, contrary to the claim, X and Y are not adjacent in G . Then X is not a parent of Y and Y is not a parent of X . Either X is not a descendant of Y or Y is not a descendant of X ; suppose without loss of generality that X is not a descendant of Y . Then by the Markov Condition, X and Y are independent in P conditional on the set of all parents of Y , which is a contradiction. \therefore

13.13 Theorem 4.5

Theorem 4.5: If probability distribution P satisfies the Markov and Minimality Conditions for directed acyclic graphs G , and P is faithful to graph H , then (i) for all X, Y, Z such that $X \rightarrow Y \leftarrow Z$ is in H and X is not adjacent to Z in H , either $X \rightarrow Y \leftarrow Z$ in G or X, Z are adjacent in G and (ii) for every triple X, Y, Z of vertices such that $X \rightarrow Y \leftarrow Z$ is in G and X is not adjacent to Z in G , if X is adjacent to Y in H and Y is adjacent to Z in H then $X \rightarrow Y \leftarrow Z$ in H .

Proof. (i) Suppose that P satisfies the Markov and Minimality Conditions for directed acyclic graphs G , and P is faithful to graph H . Suppose $X \rightarrow Y \leftarrow Z$ is in H and X is not adjacent to Z in H . By Theorem 4.4, X is adjacent to Y and Y is adjacent to Z in G . Suppose Y is not a collider on $\langle X, Y, Z \rangle$ in G and X and Z are not adjacent in G . Then by the Markov Condition X and Z are independent conditional on some set containing Y ; but since H is faithful, this is impossible.

(ii) Suppose Y is an unshielded collider on the path $\langle X, Y, Z \rangle$ in G . Then X and Z are d-separated in G given some set of vertices, and hence d-separated given $\text{Parents}(G, X)$ or $\text{Parents}(G, Z)$. It follows that X and Z are independent given $\text{Parents}(G, X)$ or $\text{Parents}(G, Z)$ in P . Y is not a parent of X or Z in G ; hence in P , X and Z are independent given some set not containing Y . But if X, Y and Z are adjacent in H and Y is not a collider on $\langle X, Y, Z \rangle$, then there is a trek between X and Z containing only X, Y , and Z ; hence in H , X and Z are not d-separated given any set of variables not containing Y . Because P is faithful to H , X and Z are not independent given any set of variables containing Y . This is a contradiction. \therefore

Corollary 4.1: If probability distribution P satisfies the Markov condition for directed acyclic graph G and P is faithful to directed acyclic graph H and G and H agree on an ordering of the variables (as, for example, by time) such that $X -> Y$ only if $X < Y$ in the order, then H is a subgraph of G .

Proof. An immediate consequence of Theorem 4.4.

13.14 Theorem 4.6

Theorem 4.6: No two distinct s.s.i. directed acyclic graphs with the same vertex set are rigidly statistically indistinguishable.

Proof. Suppose G_1 and G_2 are distinct s.s.i. directed acyclic graphs with vertex set V . Because they are s.s.i. they have the same adjacencies; hence if they are distinct graphs there is some edge $A -> B$ in G_1 and $B -> A$ in G_2 . Let U_1 and U_2 be variables not in V . Embed G_1 and G_2 in H_1 and H_2 respectively by adding edges from U_1 to A and U_2 to B . Then H_1 and H_2 are not s.s.i. because they have different colliders. \therefore

13.15 Theorem 5.1

Theorem 5.1: If the input to the PC, SGS, PC-1, PC-2, PC* or IG algorithms is data faithful to directed acyclic graph G , the output is a pattern that represents G .

In a graph G , let V be in **Undirected**(X, Y) if and only if V lies on some undirected path between X and Y .

Lemma 5.1.1: In a directed acyclic graph G , if X is not a descendant of Y , and Y and X are not adjacent in G , then X is d-separated from Y given **Parents**(Y) \cap **Undirected**(X, Y).

Proof. Suppose on the contrary that some undirected path U d-connects X and Y given **Parents**(X) \cap **Undirected**(X, Y). If U is into Y then it contains some member of **Parents**(Y) \cap **Undirected**(X, Y) not equal to X as a non-collider. Hence it does not d-connect X and Y given **Parents**(Y) \cap **Undirected**(X, Y), contrary to our assumption. If U is out of Y , then because X is not a descendant of Y , U contains a collider in **Undirected**(X, Y). Let C be the collider on U closest to Y . If U d-connects X and Y given **Parents**(Y) \cap **Undirected**(X, Y) then C has a descendant in **Parents**(Y) \cap **Undirected**(X, Y). But then C is an ancestor of Y , and Y is an ancestor of C , so G is cyclic, contrary to our assumption. Hence no undirected path between X and Y d-connects X and Y given **Parents**(Y) \cap **Undirected**(X, Y). \therefore

Lemma 5.1.2: In a directed acyclic graph G , if X is adjacent to Y , and Y is adjacent to Z , and X is not adjacent to Z , then the edges are oriented as $X \rightarrow Y \leftarrow Z$ if and only for every subset S of V , X is d-connected to Z given $\{Y\} \cup S \setminus \{X, Z\}$.

Proof. This follows from Theorem 3.4. \therefore

Lemma 5.1.3 was suggested in Pearl(1990a).

Lemma 5.1.3: In a directed acyclic graph G , if X is adjacent to Y , and Y is adjacent to Z , and X is not adjacent to Z , then either Y is in every set of variables that d-separates X and Z , or it is in no set of variables that d-separates X and Z .

Proof. Assume that in G , X, Z are not adjacent but X is adjacent to Y and Y is adjacent to Z . Since X, Z are not adjacent, they are d-separated given some subset $S \setminus \{X, Z\}$. In G , the $X - Y$ and $Y - Z$ edges collide at Y if and only if there is no set S containing Y and not X or Z such that X, Z are d-separated given S . If the $X - Y$ and $Y - Z$ edges do not collide at Y , then there is an undirected path U between X and Z that contains no colliders (including Y). Any set $S \setminus \{X, Z\}$ that does not contain Y will fail to d-separate X and Z because of this path. \therefore

Theorem 5.1: If the input to the PC, SGS, PC-1, PC-2, PC* or IG algorithms is data faithful to directed acyclic graph G , the output is a pattern that represents G .

Proof. The correctness of the SGS algorithm is evident from Theorem 3.4 since the procedure simply verifies the conditions for faithfulness given in that theorem.

Let G' be the output of one any of the algorithms except SGS. Suppose that X and Y are not adjacent in G' . None of the algorithms removes an edge between X and Y unless X and Y are d-separated given some subset of $V \setminus \{X, Y\}$. If X and Y are d-separated given some subset of $V \setminus \{X, Y\}$, then they are not adjacent in G . Hence if X and Y are not adjacent in G' , X and Y are not adjacent in G .

Suppose X and Y are adjacent in the output G' of any of the algorithms except PC*. It follows that in G X and Y are not d-separated given any subset of the adjacencies of X or any of the adjacencies of Y in G' . From what we have just proved, the adjacencies of X in G' are a superset of **Parents**(G, X) and the adjacencies of Y in G' are a superset of **Parents**(G, Y). Hence X and Y are not d-separated given **Parents**(X, G) or **Parents**(Y, G) in G . It follows from lemma 3.5.9 that X and Y are adjacent in G .

Suppose X and Y are adjacent in the output G' of PC*. **Undirected**(X, Y) in G' is a superset of **Undirected**(X, Y) in G . This, together with lemmas 3.5.9 and 5.1.1 entails that X and Y are adjacent in G .

We will show by induction on the number of applications of orientation rules in the repeat loop of the algorithm that the orientations are correct in the output G' .

Base Case: Suppose that $X \rightarrow Y$ is oriented by the rule that if X is adjacent to Y , and Y is adjacent to Z , and X is not adjacent to Z , then the edges are oriented as $X \rightarrow Y \leftarrow Z$ if and only Y is not in **Sepset**(X, Z). This is a correct orientation by lemmas 5.1.2 and 5.1.3.

Induction Case: Suppose that the orientations of G' after n applications of orientation rules are correct. Suppose first that $X \rightarrow Y$ is oriented because there is a directed path from X to Y in G' . It follows from the induction hypothesis that there is a directed path from X to Y in G , and hence $X \rightarrow Y$ in G because G is acyclic. Suppose next that $X \rightarrow Y$ is oriented because there is an edge $Z \rightarrow X$ and the edge between X and Y in G' has no arrowhead at X . It follows that Y is in **Sepset**(X, Z), and hence Y is not a collider on the path $\langle X, Y, Z \rangle$ in G . Also by the induction hypothesis $Z \rightarrow X$ in G , and hence $X \rightarrow Y$ in G . ∴

13.16 Theorem 6.1

Theorem 6.1: (Verma and Pearl): If \mathbf{V} is a set of vertices, \mathbf{O} is a subset of \mathbf{V} containing A and B , and G is a directed acyclic graph over \mathbf{V} (or an inducing path graph over \mathbf{O}) then A and B are not d-separated by any subset of $\mathbf{O} \setminus \{A, B\}$ if and only if there is an inducing path over the subset \mathbf{O} between A and B .

(Theorem 6.1 was first stated and proved in Verma and Pearl 1990 for directed acyclic graphs, but that paper did not include the parts of the lemmas relating the existence of an inducing path that is into (or out of) its endpoints to the existence of d-connecting paths that are into (or out of) their endpoints.)

If G is a directed acyclic graph over a set of variables \mathbf{V} , \mathbf{O} is a subset of \mathbf{V} containing A and B , and $A \neq B$, then an undirected path U between A and B is an **inducing path relative to \mathbf{O}** if and only if every member of \mathbf{O} on U except for the endpoints is a collider on U , and every collider on U is an ancestor of either A or B . We will sometimes refer to members of \mathbf{O} as **observed variables**. In a graph G , an edge between A and B is **into A** if and only if the mark at the A end of edge is an " $>$ ". If an undirected path U between A and B contains an edge into A we will say that U is **into A** . In a graph G , an edge between A and B is **out of A** if and only if the mark at the A endpoint is the empty mark. If an undirected path U between A and B contains an edge out of A we will say that U is **out of A** .

Lemma 6.1.1: If \mathbf{V} is a set of vertices, \mathbf{O} is a subset of \mathbf{V} , G is a directed acyclic graph over \mathbf{V} (or an inducing path graph over \mathbf{O}) if there is an inducing path relative to \mathbf{O} between A and B that is out of A and into B , then for any subset \mathbf{Z} of $\mathbf{O} \setminus \{A, B\}$ there is an undirected path C that d-connects A and B given \mathbf{Z} that is out of A and into B .

Proof. Let U be an inducing path over \mathbf{O} between A and B that is out of A and into B . Every observed vertex on U except for the endpoints is a collider, and every collider is an ancestor of either A or B .

If every collider on U has a descendant in \mathbf{Z} , then let $C = U$. C d-connects A and B given \mathbf{Z} because every collider has a descendant in \mathbf{Z} , and no non-collider is in \mathbf{Z} . C is out of A and into B .

Suppose that not every collider on U has a descendant in \mathbf{Z} . Let R be the collider on U closest to A that does not have a descendant in \mathbf{Z} , and W be the collider on U closest to A . $R \neq A$ and $R \neq B$ because A and B are not colliders on U .

Suppose first that $R = W$. There is a directed path from R to B that does not contain A , because otherwise there is a cycle in G . R is not in \mathbf{Z} because R has no descendant in \mathbf{Z} . B is not on $U(A,R)$. By lemma 3.3.2, $U(A,R)$ d-connects A and R given \mathbf{Z} , and is out of A . By lemma 3.3.3 there is a d-connecting path C between A and B given \mathbf{Z} that is out of A and into B .

Suppose then that $R \neq W$. Because U is out of A , W is a descendant of A . W has a descendant in \mathbf{Z} by definition of R . It follows that every collider on U that is an ancestor of A has a descendant in \mathbf{Z} . Hence R is an ancestor of B , and not of A . B is not on $U(A,R)$. By lemma 3.3.2, $U(A,R)$ d-connects A and R given \mathbf{Z} and is out of A . By hypothesis, there is a directed path D from R to B that does not contain A or any member of \mathbf{Z} . By lemma 3.3.3, there is a path that d-connects A and B given \mathbf{Z} that is out of A and into B . ∴

Lemma 6.1.2: If V is a set of vertices, O is a subset of V , G is a directed acyclic graph over V (or an inducing path graph over O), there is an inducing path U over O between A and B that is into A and into B then for every subset Z of $O \setminus \{A,B\}$ there is an undirected path C that d-connects A and B given Z that is into A and into B .

Proof. If every collider on U has a descendant in Z , then U is a d-connecting path between A and B given Z that is into A and into B . Suppose then that there is a collider that does not have a descendant in Z . Let W be the collider on U closest to A that does not have a descendant in Z . Suppose that W the source of a directed path D to B that does not contain A . B is not on $U(A,W)$. By lemma 3.3.2, $U(A,W)$ is a path that d-connects A and W given Z , and is into A . By lemma 3.3.3, there is an undirected path C that d-connects A and B given Z and is into A and into B . Similarly, if the first collider W on U after B that does not have a descendant in Z is the source of a directed path D to A that does not contain B , then by lemma 3.3.3, A and B are d-connected given Z by an undirected path into A and into B .

Suppose then that the collider W on U closest to A that does not have a descendant in Z is not the source of a directed path to B that does not contain A , and that the collider R on U closest to B that does not have a descendant in Z is not the source of a directed path to A that does not contain B . It follows that there is a directed path D from W to A that does not contain B or any member of Z , and there is a directed path D' from R to B that does not contain A or any member of Z . By lemma 3.3.2, $U(W,R)$ d-connects R and W given $Z = Z \setminus \{R,W\}$. D' d-connects R and B given $Z = Z \setminus \{R,B\}$ and D d-connects W and A given $Z = Z \setminus \{A,W\}$. By

lemma 3.3.1 there is an undirected path that d-connects A and B given \mathbf{Z} that is into A and into B . \therefore

In a graph G , Let $\mathbf{A}(A,B)$ be the union of the ancestors of A or B .

Lemma 6.1.3: If \mathbf{V} is a set of vertices, \mathbf{O} is a subset of \mathbf{V} , G is a directed acyclic graph over \mathbf{V} (or an inducing path graph over \mathbf{O}) and an undirected path U in G d-connects A and B given $(\mathbf{A}(A,B) \cap \mathbf{O}) \setminus \{A,B\}$ then U is an inducing path between A and B over \mathbf{O} .

Proof. If there is a path U that d-connects A and B given $(\mathbf{A}(A,B) \cap \mathbf{O}) \setminus \{A,B\}$ then every collider on U is an ancestor of a member of $(\mathbf{A}(A,B) \cap \mathbf{O}) \setminus \{A,B\}$, and hence an ancestor of A or B . Every vertex on U is an ancestor of either A or B or a collider on U , and hence every vertex on U is an ancestor of A or B . If U d-connects A and B given $(\mathbf{A}(A,B) \cap \mathbf{O}) \setminus \{A,B\}$, then every member of $(\mathbf{A}(A,B) \cap \mathbf{O}) \setminus \{A,B\}$ that is on U , except for the endpoints, is a collider. Since every vertex on U is in $\mathbf{A}(A,B)$, every member of \mathbf{O} that is on U , except for the endpoints, is a collider. Hence U is an inducing path between A and B over \mathbf{O} . \therefore

The following pair of lemmas state some basic properties of inducing paths.

Lemma 6.1.4: If G is a directed acyclic graph over \mathbf{V} , \mathbf{O} is a subset of \mathbf{V} that contains A and B , and G contains an inducing path over \mathbf{O} between A and B that is out of A , then there is a directed path from A to B in G .

Proof. Let U be an inducing path between A and B relative to \mathbf{O} that is out of A . If U does not contain a collider, then U is a directed path from A to B . If U does contain a collider, let C be the first collider after A . By definition of inducing path, there is a directed path from C to B or C to A . There is no path from C to A because there is no cycle in G ; hence there is a directed path from C to B . Because U is out of A , and C is the first collider after A , there is a directed path from A to C . Hence there is a directed path from A to B . \therefore

Lemma 6.1.5: If \mathbf{V} is a set of vertices, \mathbf{O} is a subset of \mathbf{V} , G is a directed acyclic graph over \mathbf{V} (or an inducing path graph over \mathbf{O}) that contains an inducing path relative to \mathbf{O} between A and B that is out of A , then every inducing path relative to \mathbf{O} between A and B is into B .

Proof. By lemma 6.1.4, if there is an inducing path out of A , and an inducing path out of B , there is a cycle in G . \therefore

Theorem 6.1: (Verma and Pearl): If \mathbf{V} is a set of vertices, \mathbf{O} is a subset of \mathbf{V} containing A and B , G is a directed acyclic graph over \mathbf{V} (or an inducing path graph over \mathbf{O}) A and B are not

d-separated by any subset of $\mathbf{O} \setminus \{A, B\}$ if and only if there is an inducing path over the subset \mathbf{O} between A and B .

Proof. This follows from lemmas 6.1.1, 6.1.2, 6.1.3, and 6.1.5. ∴

13.17 Theorem 6.2

Theorem 6.2: In an inducing path graph G' over \mathbf{O} , if A is not an ancestor of B , and A and B are not adjacent then A and B are d-separated given $\mathbf{D}\text{-SEP}(A, B)$.

If G' is an inducing path graph over \mathbf{O} and $A \neq B$, let $V \in \mathbf{D}\text{-SEP}(A, B)$ if and only if $A \neq V$ and there is an undirected path U between A and V such that every vertex on U is an ancestor of A or B , and (except for the endpoints) is a collider on U .

Lemma 6.2.1: If G' is the inducing path graph for G over \mathbf{O} and there is a directed path from A to B in G' , then there is a directed path from A to B in G .

Proof. Suppose there is a directed path D from A to B in G' . Let X and Y be any two vertices adjacent on the directed path and that occur in that order. There is a directed edge from X to Y in G' . By the definition of inducing path graph, there is an inducing path between X and Y in G that is out of X . Hence by lemma 6.1.4, there is a directed path from X to Y in G .

In G , the concatenation of the directed paths between vertices that are adjacent on D contains a subpath that is a directed path from A to B . ∴

Lemma 6.2.2: If G' is the inducing path graph for G over \mathbf{O} , and there is a path U d-connecting A and B given \mathbf{Z} in G' then there is a path d-connecting A and B given \mathbf{Z} in G .

Proof. Suppose that U d-connects A and B in G' . If there are vertices R , S , and T on U such that R and S are adjacent on U , and S and T are adjacent on U , and S is in \mathbf{Z} , then S is a collider on U . By the definition of inducing path graph, in G there are inducing paths over \mathbf{O} between R and S , and S and T , such that each of them is into S . By lemmas 6.1.1 and 6.1.2, in G there is a d-connecting path given $\mathbf{Z} \setminus \{R, S\}$ between R and S , and a d-connecting path given $\mathbf{Z} \setminus \{S, T\}$ between S and T , such that each of them is into S .

If there are vertices R , S , and T on U such that R and S are adjacent on U , and S and T are adjacent on U , and S is a collider on U , then S has a descendant in \mathbf{Z} in G' . By the definition of

inducing path graph, in G there are inducing paths between R and S , and S and T , that are both into S . By lemmas 6.1.1 and 6.1.2, in G there is a d-connecting path given $\mathbf{Z}\{R,S\}$ between R and S , and a d-connecting path given $\mathbf{Z}\{S,T\}$ between S and T , and both are into S . If S has a descendant in \mathbf{Z} in G' then by lemma 6.2.1 it has a descendant in \mathbf{Z} in G .

By lemma 3.3.1, there is a path in G that d-connects A and B given \mathbf{Z} . \therefore

Lemma 6.2.3: If G' is the inducing path graph for directed acyclic graph G over \mathbf{O} and there is an inducing path U over \mathbf{O} between A and C in G' , then there is an edge between A and C in G' .

Proof. Suppose there is an inducing path over \mathbf{O} between A and C in G' . By lemmas 6.1.1 and 6.1.2, in G' there is an undirected path d-connecting A and C given $\mathbf{A}(A,C) \cap \mathbf{O}\setminus\{A,C\}$. Hence by lemma 6.2.2 there is an undirected path in G such that A and C are d-connected given $\mathbf{A}(A,C) \cap \mathbf{O}\setminus\{A,C\}$ in G . By lemma 6.1.3 there is an inducing path over \mathbf{O} between A and C in G . It follows by definition that there is an edge between A and C in G' . \therefore

Let a total order Ord of variables in an inducing path graph or directed acyclic graph G' be **acceptable** if and only if whenever $A \neq B$ and there is a directed path from A to B in G' , A precedes B in Ord . In a graph G , vertex X is **after** vertex Y if and only if there is a directed path from Y to X in G , and it is **before** vertex Y if and only if there is a directed path from X to Y in G . For inducing path graph G' and acceptable total ordering Ord , let **Predecessors**(Ord, V) equal the set of all variables that precede V (not including V) according to Ord . For inducing path graph G' and acceptable total ordering Ord , W is in $\mathbf{SP}(Ord, G', V)$ (separating predecessors of V in G' for ordering Ord) if and only if $W \neq V$ and there is an undirected path U between W and V such that each vertex on U except for V precedes V in Ord and every vertex on U except for the endpoints is a collider on U . Notice that by this definition each parent of V is in $\mathbf{SP}(Ord, G', V)$. For example in figure 2, if $Ord = \langle X, S, T, R, M, Z, Q, Y \rangle$, then $\mathbf{SP}(Ord, G', Y) = \{Q, T, S\}$ and if $Ord = \langle X, S, T, R, M, Z, Y, Q \rangle$ then $\mathbf{SP}(Ord, G', Y) = \emptyset$.

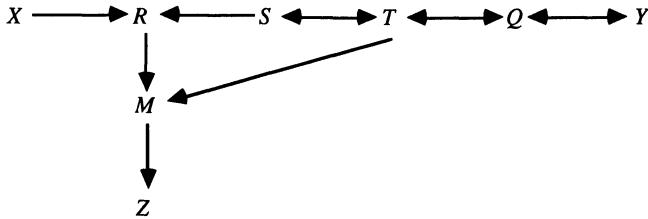


Figure 2

Lemma 6.2.4: If G' is an inducing path graph and Ord an acceptable total ordering then $\text{Predecessors}(Ord, X) \setminus \text{SP}(Ord, G', X)$ is d-separated from X given $\text{SP}(Ord, G', X)$.

Proof. Suppose on the contrary that there is a path U that d-connects some V in $\text{Predecessors}(Ord, X) \setminus \text{SP}(Ord, G', X)$ to X given $\text{SP}(Ord, G', X)$. There are three cases.

First suppose U has an edge into X that is not a double-headed arrow. (By a double-headed arrow we mean e.g. $A <> B$.) Then some parent R of X is on U , and is not a collider on U . R is in $\text{SP}(Ord, G', X)$ and hence is not equal to V . Because R is not a collider on U , U does not d-connect V to X given $\text{SP}(Ord, G', X)$, contrary to our assumption.

Next suppose U has an edge out of X . Since V is in $\text{Predecessors}(Ord, X) \setminus \text{SP}(Ord, G', X)$ it precedes X in Ord ; hence there is no directed path from X to V . It follows that U contains a collider. Let the first collider after X on U be R . R is a descendant of X , and the descendants of R are descendants of X . It follows that no descendant of R (including R itself) is in $\text{SP}(Ord, G', X)$, and hence U does not d-connect V and X , contrary to our assumption.

Suppose finally that U contains a double-arrow into X . Because U d-connects X and V given $\text{SP}(Ord, G', X)$, each collider along U has a descendant in $\text{SP}(Ord, G', X)$ and hence precedes X in Ord ; it follows that every ancestor of a collider on U precedes X in Ord . Let W be the vertex on U closest to X not in $\text{SP}(Ord, G', X)$, and R be the vertex adjacent to W on U and between W and X . If R is not a collider on U , then U does not d-connect V and X given $\text{SP}(Ord, G', X)$. If R is a collider on U , then $W \rightarrow R$ on U . W is either an ancestor of V or of a collider on U , in which case it precedes X , and is a member of $\text{SP}(Ord, G', X)$, contrary to our assumption. \therefore

Theorem 6.2: In an inducing path graph G' over \mathbf{O} , if A is not an ancestor of B , and A and B are not adjacent then A and B are d-separated given $\mathbf{D}\text{-SEP}(A, B)$.

Proof. Suppose that A and B are not adjacent, and A is not an ancestor of B . Let the total order Ord on the variables in G' be such that all ancestors of A and all ancestors of B except for A are prior to A , and all other vertices are after A . Then $\text{SP}(Ord, G', A) = \text{D-SEP}(A, B)$. Hence by lemma 6.2.4, if B is not in $\text{D-SEP}(A, B)$ then $\text{D-SEP}(A, B)$ d-separates A from B in G . B is in $\text{D-SEP}(A, B)$ if and only if there is a path from A to B in which each vertex except the endpoints is a collider on the path, and each vertex on the path is an ancestor of A or B . But then there is an inducing path between A and B , and by lemma 6.2.3 A and B are adjacent, contrary to our assumption. \therefore

13.18 Theorem 6.3

Theorem 6.3: If the input to the CI algorithm is data over \mathbf{O} that is faithful to G , the output is a partially oriented inducing path graph of G over \mathbf{O} .

In an inducing path graph G' , U is a **discriminating path** for B if and only if U is an undirected path between X and Y containing B , $B \neq X$, $B \neq Y$, and

- (i) if V and V' are adjacent on U , and V' is between V and B on U , then $V *-> V'$ on U ,
- (ii) if V is between X and B on U and V is a collider on U then $V -> Y$ in G' , else $V <-* Y$ in G' ,
- (iii) if V is between Y and B on U and V is a collider on U then $V -> X$ in G' , else $V <-* X$ in G' ,
- (iv) X and Y are not adjacent in G' .

B is a **definite non-collider** on undirected path U if and only if either B is an endpoint of U , or there exist vertices A and C such that U contains one of the subpaths $A <- B *-* C$, $A *-* B -> C$, or $A *-* \underline{B} *-* C$.

In a partially oriented inducing path graph π , U is a **definite discriminating path** for B if and only if U is an undirected path between X and Y containing B , $B \neq X$, $B \neq Y$, every vertex on U except for B and the endpoints is a collider or a definite non-collider on U , and

- (i) if V and V' are adjacent on U , and V' is between V and B on U , then $V *-> V'$ on U ,
- (ii) if V is between X and B on U and V is a collider on U then $V -> Y$ in π , else $V <-* Y$ in π ,

- (iii) if V is between Y and B on U and V is a collider on U then $V \rightarrow X$ in π , else $V <-* X$ in π ,
(iv) X and Y are not adjacent in π .

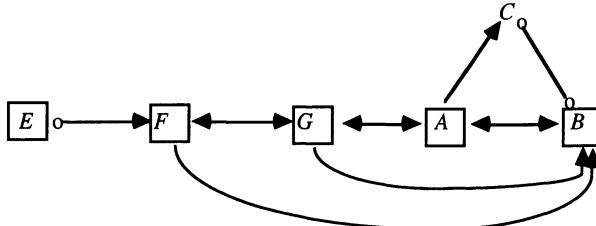


Figure 3: $\langle E, F, G, A, C, B \rangle$ is a definite discriminating path for C

Lemma 6.3.1: If G' is an inducing path graph, U is a discriminating path for B between X and Y , and X and Y are d-separated given S , then for every vertex V on U not equal to B , V is in S if and only if V is a collider on U .

Proof. First we will prove for each vertex V on U between X and B that V is in S if and only if V is a collider on U . The proof is by induction on the number of vertices between X and V on U .

Base Case: Let A be the first vertex on U after X . If $A = B$, then trivially for every vertex V between X and A , V is in S if and only if V is a collider on U . Suppose then that $A \neq B$. If A is a collider on U then there is an edge from A to Y . A is not a collider on the concatenation of $U(X,A)$ and the edge between A and Y , and hence that path d-connects X and Y given S unless A is in S . If A is not a collider on U then there is an edge between Y and A that is into A . By definition of discriminating path, the edge between X and A is into A . Hence A is a collider on the concatenation of $U(X,A)$ and the edge between A and Y . Hence that path d-connects X and Y given S unless A is not in S .

Induction Case: Suppose that if there are n or fewer vertices between X and V on U , then V is in S if and only if V is a collider on U . If there are only n vertices between X and B then we are done. Otherwise let A be the vertex such that there are $n+1$ vertices between X and A on U . Except for the endpoints, if V is on $U(X,A)$ then V is a collider on U if and only if U is in S . If A is a collider on U , then there is a directed edge from A to Y . A is not a collider on the concatenation of $U(X,A)$ and the edge from A to Y , so that path d-connects X and Y given S unless A is in S . If A is not a collider on U , then there is an edge between A and Y that is into

A. Hence A is a collider on the concatenation of $U(X,A)$ and the edge from A to Y , so that path d -connects X and Y given S unless A is not in S .

Similarly, if V is between Y and B , V is in S if and only if V is a collider on U . \therefore

Lemma 6.3.2: If G' is an inducing path graph, U is a discriminating path for B between X and Y , and X and Y are d -separated given S , then B is in S if and only if B is not a collider on U .

Proof. By lemma 6.3.1, for every vertex V on U not equal to B , V is a collider on U if and only if V is in S . If B is a collider and in S , then U d -connects X and Y given S , contrary to our assumption. If B is not a collider and not in S , then U d -connects X and Y given S , contrary to our assumption. Hence B is in S if and only if B is not a collider on U . \therefore

Theorem 6.3: If the input to the CI algorithm is data over \mathbf{O} that is faithful to G , the output is a partially oriented inducing path graph of G over \mathbf{O} .

Proof. The proof is by induction on the number of applications of orientation rules in the repeat loop of the Causal Inference Algorithm. Let G' be the inducing path graph of G . Let the object constructed by the algorithm after the n^{th} iteration of the repeat loop be π_n .

Base Case: Suppose that the only orientation rule that has been applied is that if $A *-* B *-* C$ in F , but A and C are not adjacent in F , $A *-* B *-* C$ is oriented as $A *-> B <-* C$ if B is not a member of $\text{Sepset}(A,C)$ and as $A *-* \underline{B} *-* C$ if B is a member of $\text{Sepset}(A,C)$. Suppose $A *-> B <-* C$ in π_0 , but not in G' . It follows that B is not a member of $\text{Sepset}(A,C)$, and either B is a parent of A or a parent of C in G' . If B is a parent of either A or C in G' , then there is an undirected path between A and C that does not collide at B , and except for the endpoints contains only B . For any subset S , if that path in G' does not d -connect A and C given S , then S contains B . It follows that $\text{Sepset}(A,C)$ contains B , which is a contradiction.

Suppose that $A *-* \underline{B} *-* C$ in π_0 , but the edges between A and B , and B and C collide at B in G' . It follows that $\text{Sepset}(A,C)$ does contain B but every set that d -separates A and C in G' does not contain B . Hence $\text{Sepset}(A,C)$ does not contain B , which is a contradiction.

Induction Case: Suppose π_n is a partially oriented inducing path graph of G . We will now show that π_{n+1} is a partially oriented inducing path graph of G .

Case 1: There is a directed path from A to B and an edge $A *-* B$ in π_n , so $A *-* B$ is oriented as $A *-> B$. By the induction hypothesis if there is an edge $R -> S$ in π_n , then there is an edge

$R \rightarrow S$ in G' . It follows that if there is a directed path from A to B in π_n , then there is a directed path from A to B in G' . Because G' is acyclic, $A \xrightarrow{*} B$ in G' .

Case 2: If B is a collider along $\langle A, B, C \rangle$ in π_n , B is adjacent to D , and A and C are not d-connected given D , then orient $B \xrightarrow{*} D$ as $B \leftarrow * D$. By the induction hypothesis, B is a collider along $\langle A, B, C \rangle$ and D is adjacent to B in G' . If in G' A and C are not d-connected given D by $\langle A, B, C \rangle$ then B has no descendant in $\{D\}$. Hence $D \xrightarrow{*} B$ in G' .

Case 3: If U is a definite discriminating path between A and B for M in π_n , and P and R are adjacent to M on U , and $P-M-R$ is a triangle, then

if M is in $\text{Sepset}(A, B)$ then mark M as a non-collider on subpath $P \xrightarrow{*} M \xrightarrow{*} R$
 else orient $P \xrightarrow{*} M \xrightarrow{*} R$ as $P \xrightarrow{*} M \leftarrow * R$.

By the induction hypothesis, if U is a definite discriminating path for M in π_n , then it is a discriminating path for M in G' . By lemma 6.3.2, in G' , if U is a discriminating path for M , then M is a collider on $\langle P, Q, R \rangle$ if and only if M is not in $\text{Sepset}(A, B)$.

Case 4: If $P \xrightarrow{*} M \xrightarrow{*} R$ then the orientation is changed to $P \xrightarrow{*} M \rightarrow R$. By the induction hypothesis, if $P \xrightarrow{*} M \xrightarrow{*} R$ in π_n , then in G' the edge from P to M is into M , but M is not a collider on $P \xrightarrow{*} M \xrightarrow{*} R$. It follows that $P \xrightarrow{*} M \rightarrow R$ in G' . ∴

13.19 Theorem 6.4

Theorem 6.4: If the input to the FCI algorithm is data over \mathbf{O} that is faithful to G , the output is a partially oriented inducing path graph of G over \mathbf{O} .

If $A \neq B$ in partially oriented inducing path graph π , V is in $\text{Possible-D-Sep}(A, B)$ in π if and only if $V \neq A$, and there is an undirected path U between A and V in π such that for every subpath $\langle X, Y, Z \rangle$ of U either Y is a collider on the subpath, or Y is not a definite non-collider on U , and X , Y , and Z form a triangle in π .

Lemma 6.4.1: If G' is the inducing path graph of directed acyclic graph G over \mathbf{O} , and F' is the partially oriented graph constructed in step C) of Fast Causal Inference Algorithm for G

over \mathbf{O} , A and B are in \mathbf{O} , and A is not an ancestor of B in G' , then every vertex in $\mathbf{D}\text{-SEP}(A,B)$ in G' is in $\mathbf{Possible-D-SEP}(A,B)$ in F .

Proof. Suppose that A is not an ancestor of B . If V is in $\mathbf{D}\text{-SEP}(A,B)$ in G' , then there is an undirected path U from A to V in which every vertex except the endpoints is a collider. It follows that in G' for every subpath $\langle X,Y,Z \rangle$ of U , Y is a collider on the subpath. Hence in π , Y is either a collider, or X , Y , and Z form a triangle in π and Y is not a definite non-collider. \therefore

Theorem 6.4: If the input to the FCI algorithm is data over \mathbf{O} that is faithful to G , the output is a partially oriented inducing path graph of G over \mathbf{O} .

Proof. This follows immediately from Theorem 6.3 and lemma 6.4.1. \therefore

13.20 Theorem 6.5

Theorem 6.5: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is a directed path U from A to B in π , then there is a directed path from A to B in G .

Lemma 6.5.1: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and $A \rightarrow B$ in π , then there is a directed path from A to B in G .

Proof. Let G' be the inducing path graph of G . If $A \rightarrow B$ in π , then $A \rightarrow B$ in G' . If $A \rightarrow B$ in G' , then in G there is an inducing path from A to B that is not into A . Hence by lemma 6.1.4 there is a directed path from A to B in G . \therefore

Theorem 6.5: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is a directed path U from A to B in π , then there is a directed path from A to B in G .

Proof. By lemma 6.5.1, for each edge between R and S in U there is a directed path from R to S in G . The concatenation of the directed paths in G contains a subpath that is a directed path from A to B in G . \therefore

13.21 Theorem 6.6

Theorem 6.6: If π is the CI partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is no semi-directed path from A to B in π , then there is no directed path from A to B in G .

Lemma 6.6.1: Suppose that G is a directed acyclic graph, and in G there is a sequence of vertices M starting with A and ending with C , and a set of paths F such that for every pair of vertices I and J adjacent in M there is exactly one inducing path W over \mathbf{O} between I and J in F . Suppose further that if $J \neq C$ then W is into J , and if $I \neq A$ then W is into I , and I and J are ancestors of either A or C . Then in G there is an inducing path T over \mathbf{O} between A and C such that if the path in F between A and its successor in M is into A then U is into A , and if the path in F between C and its predecessor in M is into C then U is into C .

Proof. Suppose that in G there is a sequence M of vertices in \mathbf{O} starting with A and ending with C , and a set of paths F such that for every pair of vertices I and J adjacent in M there is exactly one inducing path W over \mathbf{O} between I and J , and if $J \neq C$ then W is into J , and if $I \neq A$ then W is into I , and I and J are ancestors of either A or C . Let T' be the concatenation of the paths in F . T' may not be an acyclic undirected path because it might contain undirected cycles. Let T be an acyclic undirected subpath of T' between A and C . We will now show that except for the endpoints, every vertex in \mathbf{O} on T is a collider, and every collider on T is an ancestor of A or C .

If V is a vertex in \mathbf{O} that is on T but that is not equal to A or C , every edge on every path in F is into V . Hence, every edge on T that contains V is into V because the edges on T are a subset of the edges on inducing paths in F .

Let R and S be the endpoints of W . We will now show that every vertex on W is either an ancestor of A or an ancestor of C . By hypothesis, R is an ancestor of either A or C , and S is an ancestor of either A or C . Because W is an inducing path over \mathbf{O} , every collider on W is an ancestor of either R or S , and hence an ancestor of either A or C . Every non-collider on W is either an ancestor of R or S , or an ancestor of a collider on W . Hence every vertex on W is an ancestor of either A or C . It follows that every collider on T is an ancestor of A or C , because the vertices on T are a subset of the vertices on paths in F .

By definition, T is an inducing path between A and C over \mathbf{O} . Suppose the path in F between A and its successor is into A . If the edge on T with endpoint A is on path in F on which A is an

endpoint, then T is into A because by hypothesis that inducing path is into A . If the edge on T with endpoint A is on an inducing path over O in which A is not an endpoint of the path, then T is into A because A is in O , and hence a collider on every inducing path for which it is not an endpoint. Similarly, T is into C if in F the path between C and its predecessor is into A . \therefore

In an inducing path or directed acyclic graph G that contains an undirected path U between X and Y , the the edge between V and W is **substitutable** for $U(V,W)$ in U if and only if V and W are on U , V is between X and W on U , G contains an edge between V and W , V is a collider on the concatenation of $U(X,V)$ and the edge between V and W if and only if it is a collider on U , and W is a collider on the concatenation of $U(Y,W)$ and the edge between V and W if and only if it is a collider on U .

Lemma 6.6.2: If G' is an inducing path graph for directed acyclic graph G over O , C is a descendant of B in G , and U is an undirected path in G' between X and R containing subpath $A * \rightarrow B \leftrightarrow C$ where A is between X and B , then in G' there is a vertex E on U between X and A inclusive such that the edge between E and C is substitutable for $U(E,C)$ in U . Furthermore the concatenation of $U(X,E)$ and the edge between E and C is into C , and if U is into X , then the concatenation of $U(X,E)$ and the edge between E and C is into X .

Proof. Suppose G' is an inducing path graph for directed acyclic graph G over O , C is a descendant of B in G , and U is an undirected path in G' between X and R containing subpath $A * \rightarrow B \leftrightarrow C$ where A is between X and B . If E and F are on U , we will say that F is the successor of E on U if and only if there is an edge between E and F on U and E is between X and F or $E = X$. Let Y be the successor of X on U .

First we consider the case where there is no vertex V on U between X and A inclusive such that the edge from V to C is substitutable for $U(V,C)$ in U , but each vertex on U between Y and A inclusive is adjacent to C in G' . We will show that there is a directed path from Y to B .

Suppose that $U(Y,B)$ is not a directed path from Y to B . Let E be the vertex on U closest to B such that $U(E,B)$ is not a directed path from E to B . Let F be the successor of E on U . F is an ancestor of B in G' , not a collider on U , and by assumption F is adjacent to C . The edge between C and F is not out of C and into F , because G' is acyclic. Hence it is into C . If $F = B$, then $A \leftrightarrow B \leftrightarrow C$ in G' . It follows that in G there is an inducing path between A and C that is into A and C , and hence $A \leftrightarrow C$ in G' , and the edge between A and C is substitutable for the subpath of U between A and C . Suppose then that $F \neq B$. $U(F,B)$ is a directed path from F to B in G' . Because the edge between F and C is not substitutable for $U(F,C)$ in U it follows that F is a collider on the concatenation of $U(X,F)$ with the edge between F and C . Hence the

edge between F and C is into F and into C , and the edge between E and F on U is into F . It follows that the edge between E and F is also into E because E is not an ancestor of B , and F is. Hence G' contains the path $E \leftrightarrow F \leftrightarrow C$. Because F is an ancestor of B in G' , it is an ancestor of B in G . Since F is an ancestor of B in G , and B is an ancestor of C in G , F is an ancestor of B in G . It follows by lemma 6.6.1 that there is an inducing path between E and C in G relative to \mathbf{O} that is into E and into C . But then in G' the edge between E and C is substitutable for $U(E,C)$ in U , which is a contradiction.

We have shown that $U(Y,B)$ is a directed path from Y to B . It follows that Y is an ancestor of B in G , and because B is an ancestor of C in G , Y is an ancestor of C in G . We have shown that the edge between Y and its successor on U is out of Y . Hence Y is not a collider on U . By assumption there is an edge between Y and C in G' . If the edge between Y and C is not substitutable for $U(Y,C)$ in U , then the edge between Y and C is into Y , and because G' is acyclic (i.e. there is no directed cycle in G'), the edge between Y and C is also into C . Because the edge between Y and C is not substitutable for $U(Y,C)$ in U , and the edge between Y and C is into Y , it follows that the edge between X and Y is into Y . Hence G' contains the path $X * \rightarrow Y \leftrightarrow C$, and Y is an ancestor of C in G . It follows that there is an inducing path between X and C in G relative to \mathbf{O} that is into C , and if U is into X , also into X . But then the edge between X and C is substitutable for $U(X,C)$ in U , which is a contradiction.

Next we consider the case where there is no vertex V on U between X and A inclusive such that the edge from V to C is substitutable for $U(V,C)$ in U , but some vertex on U between Y and A inclusive is not adjacent to C . Let E be the vertex on U closest to C and between X and C that is not adjacent to C , and let F be the successor of E on U . $E \neq A$, because by lemma 6.6.1 there is an inducing path between A and C in G , and hence A is adjacent to C in G' . From the previous case, it follows that either there is an edge between V on $U(E,C)$ and C that is substitutable for $U(V,C)$ in $U(E,C)$ or F is an ancestor of B in G' . Suppose first that there is an edge between V on $U(E,C)$ and C that is substitutable for $U(V,C)$ in $U(E,C)$. E is not adjacent to C , so $V \neq E$, and V lies on $U(F,C)$. If the edge between V and C is substitutable for $U(V,C)$ in $U(E,C)$, then it is also substitutable for $U(V,C)$ in U , which is a contradiction. Hence F is an ancestor of B in G' . By the definition of E , F is adjacent to C in G' . The edge between F and C is not out of C and into F , because G' is acyclic. The edge between F and C is not out of F and into C because the edge between F and C is not substitutable for $U(F,C)$ in $U(E,C)$, and $U(F,B)$ is a directed path from F to B . Hence the edge between F and C is into F and C . If the edge $E \leftarrow F$ is on U , then the $F \leftrightarrow C$ edge is substitutable for $U(F,C)$ in U . If $E * \rightarrow F$ in G' then G' contains the path $E * \rightarrow F \leftrightarrow C$, and F is an ancestor of C in G' and hence in G ; it follows that there is an

inducing path between E and C relative to \mathbf{O} in G , and E is adjacent to C in G' . This is a contradiction.

It follows that for some vertex E on U between X and A inclusive there is an edge from E to C that is substitutable for $U(E,C)$ in U and is into C . If $E = X$ then there is an inducing path between X and C that contains the edge on U with X as endpoint. If $E \neq X$ then there is some vertex $E \neq X$ on U such that there is an edge between E and C that is substitutable for $U(E,C)$ in U . In the first case, the inducing path is into X if U is into X and hence the edge between C and X is into X . In the second case the path consisting of the concatenation of $U(X,V)$ and the edge between V and C contains the edge on U with X as endpoint, and hence is into X if U is.

∴

Lemma 6.6.4: If π is the CI partially oriented inducing path graph of graph G over \mathbf{O} , and $A *-> B$ in π , then every inducing path in G between A and B is into B .

Proof. We will prove that each orientation rule in the Causal Inference Algorithm is such that if the rule orients the edge between A and B as $A *-> B$, then every inducing path between A and B over \mathbf{O} in G is into B . Let G' be the inducing path graph of G .

Case 1: By lemma 6.5.1 any of the rules that orients the edge between A and B as $A -> B$ in π entails that there is a directed path from A to B in G . If there is an inducing path over \mathbf{O} between A and B in G that is out of B , and there is a directed path from B to A in G . But G is not cyclic, so there is no inducing path between A and B in G that is not into B .

Case 2: Suppose the edge between A and B is oriented as $A *-> B$ in order to avoid a cycle in π because there is a directed path from A to B in π . By Theorem 6.5 there is a directed path from A to B in G . If there is an inducing path over \mathbf{O} between A and B in G that is out of B , then there is a directed path from B to A in G . But G is not cyclic, so there is no inducing path over \mathbf{O} between A and B in G that is out of B .

Case 3: Suppose that the edge between A and B is oriented as $A *-> B$ because there is a vertex C such that A and B are adjacent in π , B and C are adjacent in π , A and C are not adjacent in π , and B is not in $\text{Sepset}(A,C)$. It follows that $A *-> B <-* C$ in G' . By the construction of G' it follows that in G there is an inducing path over \mathbf{O} between A and B into B , and an inducing path over \mathbf{O} between B and C into B . Suppose contrary to the theorem that there is another inducing path over \mathbf{O} between A and B in G that is out of B . By lemma 6.1.4, A is a descendant of B in G . By lemma 6.6.1 there is an inducing path over \mathbf{O} between A and C . But

if there is an inducing path over \mathbf{O} between A and C in G , then A and C are adjacent in π , contrary to our assumption.

Case 4: Suppose that the edge between A and B is oriented as $A *-> B$ because B is a collider along $\langle C, B, D \rangle$ in π , B is adjacent to A , and C and D are not d-connected given A . Suppose, contrary to the theorem, that in G there is an inducing path over \mathbf{O} between A and B that is out of B . It follows that A is a descendant of B in G . Because there is an edge between C and B that is into B in π , there is an edge between C and B that is into B in G' . The edge between C and B in G' d-connects C and B given A and is into B . By lemmas 6.1.1 and 6.1.2 there is a path in G that d-connects C and B given A that is into B . Similarly, there is a path in G that d-connects D and B given A that is into B . By lemma 3.3.1, C and D are d-connected given A in G . This is a contradiction.

Case 5: Suppose the edge between A and B in π is oriented as $A *-> B$ because in π U is a definite discriminating path for B between X and Y , B is in a triangle on U , and B is not in $\text{Sepset}(X, Y)$. Let A and C be the vertices adjacent to B on U . If U is a definite discriminating path for B in π , then by the induction hypothesis, the corresponding path U' in G' is a discriminating path for B . In G' , X and Y are d-separated given $\text{Sepset}(X, Y)$ because by definition of definite discriminating path they are not adjacent. If X and Y are d-separated given $\text{Sepset}(X, Y)$ in G' , then by lemma 6.3.1 every collider on U' except for B is in $\text{Sepset}(X, Y)$, and every non-collider on U' is not in $\text{Sepset}(X, Y)$.

Suppose that there is an inducing path over \mathbf{O} between B and A in G that is out of B . It follows that there is a directed path from B to A in G and that $A <-> B$ in G' . By definition of discriminating path it follows that A is a collider on U' or $A = X$. By lemma 6.3.1 A is in $\text{Sepset}(X, Y)$. Hence B is a collider on U' in G' , and B has a descendant in $\text{Sepset}(X, Y)$ in G .

If some vertex Z on U is in $\text{Sepset}(X, Y)$ then Z is a collider on U . Let R and T be the vertices on U' that are adjacent to Z on U' . By the definition of inducing path graph, in G there are inducing paths over \mathbf{O} between R and Z , and Z and T , such that each of them is into Z . By lemmas 6.1.1 and 6.1.2, in G there is a d-connecting path given $\mathbf{S}\backslash\{R, Z\}$ between R and Z , and a d-connecting path given $\mathbf{S}\backslash\{Z, T\}$ between Z and T , such that each of them is into Z .

If there are vertices R , Z , and T on U' such that R and Z are adjacent on U , and Z and T are adjacent on U' , and Z is a collider on U' , then either Z is in $\text{Sepset}(X, Y)$ (if $Z \neq B$), or Z has a descendant in $\text{Sepset}(X, Y)$ in G (if $Z = B$). In either case Z has a descendant in $\text{Sepset}(X, Y)$

in G . By the definition of inducing path graph, in G there are inducing paths over \mathbf{O} between R and Z , and Z and T , that are both into Z . By lemmas 6.1.1 and 6.1.2, in G there is a d-connecting path given $\text{Sepset}(X,Y) \setminus \{R,Z\}$ between R and Z , and a d-connecting path given $\text{Sepset}(X,Y) \setminus \{Z,T\}$ between Z and T , that are both into Z . By lemma 3.3.1, there is a path in G that d-connects X and Y given $\text{Sepset}(X,Y)$. But this contradicts the assumption that X and Y are d-separated given $\text{Sepset}(X,Y)$. Hence there is no inducing path in G that is out of B . \therefore

A semi-directed path from A to B in partially oriented inducing path graph π is an undirected path U from A to B in which no edge contains an arrowhead pointing towards A , that is, there is no arrowhead at A on U , and if X and Y are adjacent on the path, and X is between A and Y on the path, then there is no arrowhead at the X end of the edge between X and Y .

Theorem 6.6: If π is the CI partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and there is no semi-directed path from A to B in π , then there is no directed path from A to B in G .

Proof. Suppose there is a directed path P from A to B in G . Let P' in π be the sequence of vertices in \mathbf{O} along P in the order in which they occur. P' is an undirected path in π because for each pair of vertices X and Y adjacent in P' for which X is between A and Y or $X = A$ there is an inducing path over \mathbf{O} in G that is out of X . P' is a semi-directed path from X to Y in π because by lemma 6.6.4, there is no arrowhead into X on P' . \therefore

13.22 Theorem 6.7

Theorem 6.7: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , A and B are adjacent in π , and there is no undirected path between A and B in π except for the edge between A and B , then in G there is a trek between A and B that contains no variables in \mathbf{O} other than A or B .

Proof. Suppose that every trek between A and B in G contains some member of \mathbf{O} other than A or B . Because there is an edge between A and B in π , there is an inducing path between A and B in G . Hence, A and B are d-connected given the empty set in G , and there is a trek T between A and B . Let U be the sequence of observed vertices on T . Each subpath of T between

variables adjacent in U is an inducing path relative to \mathbf{O} . Hence U is an undirected path in π that contains a member of \mathbf{O} other than A or B . ∴

13.23 Theorem 6.8

Theorem 6.8: If π is the CI partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and every semi-directed path from A to B contains some member of \mathbf{C} in π , then every directed path from A to B in G contains some member of \mathbf{C} .

Proof. Suppose that U is a directed path in G from A to B that does not contain a member of \mathbf{C} . Let the sequence of observed variables on U in G be U' . Let X and Y be two adjacent vertices in U' , where X is between A and Y . $U(X,Y)$ is a directed subpath of U that contains no observed variables except for the endpoints. Hence $U(X,Y)$ is an inducing path between X and Y given \mathbf{O} that is out of X . It follows that there is an edge between X and Y in π , and by lemma 6.6.4 the edge between X and Y is not into X . Hence U' is a semi-directed path from A to B in π that does not contain any member of \mathbf{C} . ∴

13.24 Theorem 6.9

Theorem 6.9: If π is a partially oriented inducing path graph of directed acyclic graph G over \mathbf{O} , and $A \leftrightarrow B$ in π , then there is a latent common cause of A and B in G .

Proof. By Theorem 6.6, every inducing path over \mathbf{O} in G between A and B is into B and into A . By lemma 6.1.2, there is in G a d-connecting path U between A and B given the empty set that is into A and into B in G . Because U d-connects A and B given the empty set in G it contains no colliders, and hence no members of \mathbf{O} except A and B . Because U contains an edge into A and an edge into B , U is not a single edge between A and B . Hence there is some vertex C not in \mathbf{O} on U that is a common cause of A and B . ∴

13.25 Theorem 6.10 (Tetrad Representation Theorem)

Tetrad Representation Theorem 6.10: In an acyclic LCF G , there exists an $LJ(T(I,J), T(K,L), T(I,L), T(J,K))$ choke point or an $IK(T(I,J), T(K,L), T(I,L), T(J,K))$ choke point iff G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

In a graph G , the **length** of a path equals the number of vertices in the path minus one. In a graph G , a path U of length n is an **initial segment** of path V of length m iff $m \geq n$, and for $1 \leq i \leq n+1$, the i^{th} vertex of V equals the i^{th} vertex of U . In a graph G , path U of length n is a **final segment** of path V of length m , iff $m \geq n$, and for $1 \leq i \leq n+1$, the i^{th} vertex of U equals the $(m-n+i)^{\text{th}}$ vertex of V . A path U of length n is a **proper initial segment** of path V of length m iff U is an initial segment of V and $U \neq V$. A path U of length n is a **proper final segment** of path V of length m iff U is a final segment of V and $U \neq V$.

The proofs of the following lemma is obvious.

Lemma 6.10.1: In a directed graph G , if $R(U,I)$ is an acyclic path, and X is a vertex on $R(U,I)$, then there is a unique initial segment of $R(U,I)$ from U to X .

Because the proofs refer to many different paths, we will usually designate a directed path by $R(X,Y)$ where X and Y are the endpoints of the path. When there is a path $R(U,I)$ in a proof, and a vertex X on $R(U,I)$, $R(U,X)$ will refer to the unique initial segment of $R(U,I)$ from U to I , and $R(X,I)$ will refer to the unique final segment of $R(U,I)$ from X to I .

In a directed acyclic graph G , the **last point of intersection** of directed path $R(U,I)$ with directed path $R(V,J)$ is the last vertex on $R(U,I)$ that is also on $R(V,J)$. Note that if G is a directed acyclic graph, the last point of intersection of directed path $R(U,I)$ with directed path $R(V,J)$ equals the last point of intersection of $R(V,J)$ with $R(U,I)$; this is not true of directed cyclic paths.

Lemma 6.10.2: If G is a directed acyclic graph, for all variables Y and Z in G , if $Y \neq Z$ and R and R' are two intersecting directed paths with sinks Y and Z respectively then there is a trek between Y and Z that consists of subpaths of R and R' .

Proof. Since R and R' intersect, they have a last point of intersection X . Let the source of the trek to be constructed be X . $R(X,Y)$ and $R(X,Z)$ do not intersect anywhere except at X . Since $Y \neq Z$, one of $R(X,Y)$ and $R(X,Z)$ is not empty. Hence $\{R(X,Y), R(X,Z)\}$ is a trek. \therefore

In a directed acyclic graph, directed paths $R(U,I)$ and $R(U,J)$ **contain trek** T iff $I(T(I,J))$ is a final segment of $R(U,I)$ and $J(T(I,J))$ is a final segment of $R(U,J)$.

Lemma 6.10.3: In a directed acyclic graph, if $R(U,I)$ and $R(U,J)$ are directed paths that contain both $T(I,J)$ and $T'(I,J)$, then $T(I,J) = T'(I,J)$.

Proof. In a directed acyclic graph, there is a unique last point of intersection of $R(U,I)$ and $R(U,J)$, and unique final segments of R and R' whose source is the last point of intersection of $R(U,I)$ and $R(U,J)$. \therefore

If G is a directed acyclic graph, let Let \mathbf{P}_{XY} be the set of all directed paths in G from X to Y . In an LCF S , the **path form of a product of covariances** $\gamma_{IJ}\gamma_{KL}$ is the distributed form of

$$\left(\sum_{U \in \mathbf{U}_{IJ}} \left(\sum_{R \in \mathbf{P}_{UI}} \sum_{R' \in \mathbf{P}_{UJ}} L(R)L(R')\sigma_U^2 \right) \right) \left(\sum_{V \in \mathbf{U}_{KL}} \left(\sum_{R'' \in \mathbf{P}_{VK}} \sum_{R''' \in \mathbf{P}_{VL}} L(R'')L(R''')\sigma_V^2 \right) \right)$$

$\gamma_{IJ}\gamma_{KL} - \gamma_{IL}\gamma_{JK}$ is in **path form** iff both terms are in path form.

Henceforth, we will assume that all variances, covariances, products of covariances, and tetrad differences are in path form unless otherwise stated.

We will adopt the following terminology. Suppose that m is a term in the path form of a product of covariances $\gamma_{IJ}\gamma_{KL}$. By definition, m is of the form $L(R(U,I))L(R(U,J))L(R(V,K))L(R(V,L))\sigma_U^2\sigma_V^2$. Let the paths associated with m be the ordered quadruple $\langle R(U,I), R(U,J), R(V,K), R(V,L) \rangle$. There is a one-to-one correspondence between terms in the path form of a product of covariances, and such ordered quadruples. We will consider terms m and m' to be distinct terms if their associated paths are different (i.e. the terms may contain the same number of occurrences of the same edge labels, but in different orders.) Note that under this criterion of identity of terms, no term appears twice in the path form of a product of covariances or tetrad difference. Henceforth when we consider sets of terms appearing in some expression, we will do so under the assumption that each term occurs at most once in the expression (although distinct terms that have identically equal values may

occur in the expression). We will say that a term m contains a path or trek X if its associated quadruple contains X .

Lemma 6.10.4: A tetrad difference $\gamma_{JJKL} - \gamma_{LYJK}$ is not linearly implied to vanish by an LCF S if there is a term m in the path form of γ_{JJKL} such that every term m' in the path form of γ_{LYJK} contains an edge not in m .

Proof. Suppose that there is a term m in the path form of γ_{JJKL} such that every term m' in the path form of γ_{LYJK} contains an edge not in m . Set every variable not in m to be zero. Then γ_{LYJK} is zero since every term in γ_{LYJK} contains a variable not in m . Set every variable in m to be positive. Then every non-zero term in the path form of γ_{JJKL} is positive, since the e.c2.f of each non-zero term is positive, and the c.f. of each non-zero term is positive. γ_{JJKL} is not zero since every term in it is either 0 or positive, and some are positive. Hence the tetrad difference is not linearly implied to vanish. ∴

Lemma 6.10.5: In an LCF S , if the paths in a term m in the path form of a tetrad difference have different sources than the paths in a term m' , then m contains some variable not in m' .

Proof. Each of the sources of the paths in m and m' are independent random variables, and it is not the case that all of the paths in m or m' are empty. Let $\{I,J\}$ be the sources of the paths in m , and $\{K,Z\}$ be the sources of the paths in m' and suppose that $\{I,J\} \neq \{K,Z\}$. Suppose w.l.g. that $I \neq K$. Since I , K , and Z are independent I does not occur on any paths with source K or Z . m contains at least one edge X out of I . Since I does not occur on any path with source K or Z , X does not occur on any path with source K or Z . Hence m contains a variable (the label of X) that does not occur in m' . ∴

In an LCF F , $e(S)$ is equal to S if S is an independent variable, and it is equal to the error variable into S if S is not an independent variable.

Lemma 6.10.6: In an LCF S , if there exist $T(I,J) \in \mathbf{T}(I,J)$ and $T(K,L) \in \mathbf{T}(K,L)$ such that $I(T(I,J)) \cap K(T(K,L)) = \emptyset$, $J(T(I,J)) \cap L(T(K,L)) = \emptyset$, and $I(T(I,J)) \cap L(T(K,L)) = \emptyset$, then there exists a term m in γ_{JJKL} such that every term m' in γ_{LYJK} contains an edge not in m .

Proof. Let S be the source of $T(I,J)$ and S' be the source of $T(K,L)$. (Note that since $I(T(I,J))$ does not intersect $L(T(K,L))$, the source of $T(I,J)$ does not equal the source of $T(K,L)$, and hence $e(S)$ does not equal $e(S')$. See figure 4.

Let $m = L(R(e(S),I))L(R(e(S),J))L(R(e(S'),K))L(R(e(S'),L))$. m is the coefficient of a term in γ_{JJKL} (the full term also contains a factor equal to the product of the variances of the sources of paths in m .)

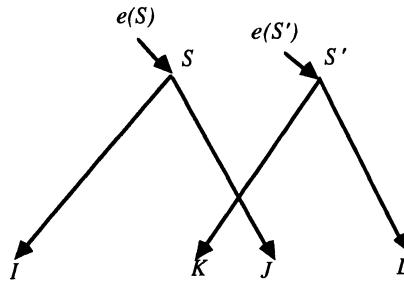


Figure 4

Suppose there is a term m' in $\gamma_{JL}\gamma_{JK}$ whose associated paths contain only edges in m . m' contains the product of the labels of edges in a trek $T(I,L)$. Let the source of $T(I,L)$ be S'' . If $S'' \neq S$ and $S'' \neq S'$, then $e(S'') \neq e(S)$ and $e(S'') \neq e(S')$. Since $e(S'')$ is an independent variable, and the only independent variables in m are $e(S)$ and $e(S')$, if $e(S'') \neq e(S)$ and $e(S'') \neq e(S')$, then $T(I,L)$ contains an edge label not in m . Suppose then w.l.g. that $S'' = S$. There is a path $R(S,L)$ containing edge labels only in m . Since $J(T(I,J)) \cap L(T(K,L)) = \emptyset$, and $I(T(I,J)) \cap L(T(K,L)) = \emptyset$, the only path in m that contains L is $L(T(K,L))$. Hence $R(S,L)$ intersects $L(T(K,L))$ at some vertex. The only two paths in m with source S are $I(T(I,J))$ and $J(T(I,J))$, and neither of them intersects $L(T(K,L))$. Hence one of them intersects some other paths that in turn intersects $L(T(K,L))$. The only other path in m that intersects $L(T(K,L))$ is $K(T(K,L))$. So $R(S,L)$ intersects $K(T(K,L))$. Since the last point of intersection of $L(T(K,L))$ and $K(T(K,L))$ is S' , $R(S,L)$ intersects $K(T(K,L))$ at or before S' . But the only paths with source S in m are $J(T(I,J))$ and $I(T(I,J))$, and neither of them intersects $K(T(K,L))$ at or before S' . Hence, there is no path from S to L containing only edge labels in m . Similarly it can be shown that there is no path from S' to I containing only edge labels in m . Hence m' contains an edge label not in m . \therefore

Lemma 6.10.7: In an LCF S , if there exists a $T(I,J) \in \mathbf{T}(I,J)$ and $T(K,L) \in \mathbf{T}(K,L)$ such that $I(T(I,J)) \cap K(T(K,L)) = \emptyset$, and $L(T(K,L)) \cap J(T(I,J)) = \emptyset$, or there exists a $T(I,L) \in \mathbf{T}(I,L)$ and $T(J,K) \in \mathbf{T}(J,K)$ such that $I(T(I,L)) \cap K(T(J,K)) = \emptyset$, and $L(T(I,L)) \cap J(T(J,K)) = \emptyset$, then S does not linearly imply that $\gamma_{JL}\gamma_{KI} - \gamma_{LJ}\gamma_{JK}$ vanishes.

Proof. Suppose w.l.g. that $I(T(I,J)) \cap K(T(K,L)) = \emptyset$, and $L(T(K,L)) \cap J(T(I,J)) = \emptyset$. There are four cases: either (i) $I(T(I,J)) \cap L(T(K,L)) = \emptyset$ and $J(T(I,J)) \cap K(T(K,L)) = \emptyset$, or (ii) $I(T(I,J)) \cap L(T(K,L)) = \emptyset$ and $J(T(I,J)) \cap K(T(K,L)) \neq \emptyset$, or (iii) $I(T(I,J)) \cap L(T(K,L))$

$\neq \emptyset$ and $J(T(I,J)) \cap K(T(K,L)) = \emptyset$, or (iv) $I(T(I,J)) \cap L(T(K,L)) \neq \emptyset$ and $J(T(I,J)) \cap K(T(K,L)) \neq \emptyset$.

In the first three cases, by lemma 6.10.6 there exists a term m in $\gamma_{IJ}\gamma_{KL}$ such that every m' in $\gamma_{IL}\gamma_{JK}$ contains an edge label not in m .

In the fourth case, let X be the last point of intersection of $I(T(I,J))$ and $L(T(K,L))$, and Y be the last point of intersection of $J(T(I,J))$ and $K(T(K,L))$. X is not the source of either trek, since otherwise $I(T(I,J)) \cap K(T(K,L)) \neq \emptyset$ or $J(T(I,J)) \cap L(T(K,L)) \neq \emptyset$. Similarly, Y is not the source of either trek. $\{R(X,I), R(X,L)\}$ is a trek $T(I,L)$ between I and L , by lemma 6.10.2. Similarly, $\{R(Y,J), R(Y,K)\}$ form a trek $T(J,K)$. (See figure 5.)

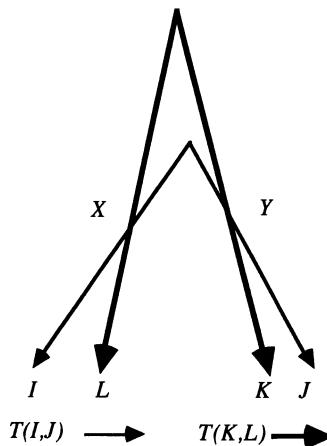


Figure 5

Now we will show that $T(I,L) \cap T(J,K) = \emptyset$. $I(T(I,L)) \cap J(T(J,K)) = \emptyset$ since $I(T(I,L))$ is a proper subpath of $I(T(I,J))$ and $J(T(J,K))$ is a proper subpath of $J(T(I,J))$, and the last point of intersection of $I(T(I,J))$ and $J(T(J,K))$ is the source of $T(I,L)$. $I(T(I,L)) \cap K(T(J,K)) = \emptyset$, since $I(T(I,L))$ is a subpath of $I(T(I,J))$ and $K(T(J,K))$ is a subpath of $K(T(K,L))$, and $I(T(I,J)) \cap K(T(K,L)) = \emptyset$ by hypothesis. For similar reasons, $L(T(I,L)) \cap J(T(J,K)) = \emptyset$, and $L(T(I,L)) \cap K(T(J,K)) = \emptyset$. It follows from lemma 6.10.7 there exists a term m in $\gamma_{IL}\gamma_{JK}$ such that every m' in $\gamma_{IJ}\gamma_{KL}$ contains an edge label not in m .

Since there exists a term m in $\gamma_{IL}\gamma_{JK}$ such that every m' in $\gamma_{IJ}\gamma_{KL}$ contains an edge not in m , by lemma 6.10.4 $\gamma_{IJ}\gamma_{KL} - \gamma_{IL}\gamma_{JK}$ is not linearly implied. \therefore

A vanishing tetrad difference is a constraint upon the covariances of four pairs of variables: $\langle I,J \rangle$, $\langle K,L \rangle$, $\langle I,L \rangle$ and $\langle J,K \rangle$. Roughly speaking, a choke point for such a foursome of variable pairs is a point where all of the treks between I and J intersect all of the treks between K and L , and all of the treks between I and L intersect all of the treks between J and K . (A more precise definition is given later.) In this section, we will prove that in an LCF G , the existence of such a choke point is a necessary condition for the corresponding tetrad difference to vanish in distributions perfectly represented by G . We will prove this by showing that the existence of a choke point in G is equivalent to a condition that has already been proved to be a necessary condition for S to linearly imply a vanishing tetrad difference; namely, the trek intersection condition described in lemma 6.10.7. Unfortunately, this proof is long and tedious because there are many different ways in which a choke point can fail to exist, depending upon which treks are assumed to intersect and which treks are assumed not to intersect. In each case we show that the non-existence of a choke point implies the violation of the necessary condition described in lemma 6.10.7.

Two strategies are employed in the proofs. The first is to show that the assumptions about which treks intersect and don't intersect lead to contradictions. The second is to show that it is possible to construct a pair of treks $T'(I,J)$ and $T'(K,L)$ such that $I(T'(I,J))$ and $K(T'(K,L))$ don't intersect, and $J(T'(I,J))$ and $L(T'(K,L))$ don't intersect, or to construct a pair of treks $T'(I,L)$ and $T'(J,K)$ such that $I(T'(I,L))$ and $K(T'(J,K))$ don't intersect, and $J(T'(J,K))$ and $L(T'(I,L))$ don't intersect. In either case, by lemma 6.10.7, it follows that $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK}$ is not linearly implied by G .

In general, when constructing a trek $T(I,J)$ we will speak as if it suffices to show how to construct a pair of (acyclic) directed paths R and R' from a common source S to sinks I and J respectively, without showing that the pair of directed paths constructed do not intersect. This is because even if R and R' do not form a trek because they intersect each other at some vertex other than S , we have shown in lemma 6.10.2 that directed subpaths of R and R' do form a trek, and the existence of the directed subpaths of R and R' is enough for our purposes. We are generally interested in showing that particular pairs of trek branches fail to intersect. If R_1 and R_2 fail to intersect, then directed subpaths of R_1 and R_2 also fail to intersect. Hence, if the goal is to show that trek branches T and T' fail to intersect, it suffices to show that R_1 and R_2 fail to intersect, even if T and T' are actually equal to directed subpaths of R_1 and R_2 respectively.

Let S be a set of vertices, and $\mathbf{R}_K(S)$ be the set of all directed paths with sink K and a source in S . Let $R(S,I)$ be a directed path from S in S to I . Let X_n be the n^{th} vertex on $R(S,I)$ such that some directed path in $\mathbf{R}_K(S)$ intersects it. Let the set of sources of directed paths in $\mathbf{R}_K(S)$ whose first point of intersection with $R(S,I)$ is X_n be S_n . Let the last vertex in $R(S,I)$ that is the first intersection of some directed path in $\mathbf{R}_K(S)$ with $R(S,I)$ be X_{\max} . Note that X_{\max} is not necessarily the last point of intersection of some directed path in $\mathbf{R}_K(S)$ with $R(S,I)$; it is merely the last of the first points of intersection. See figure 6.

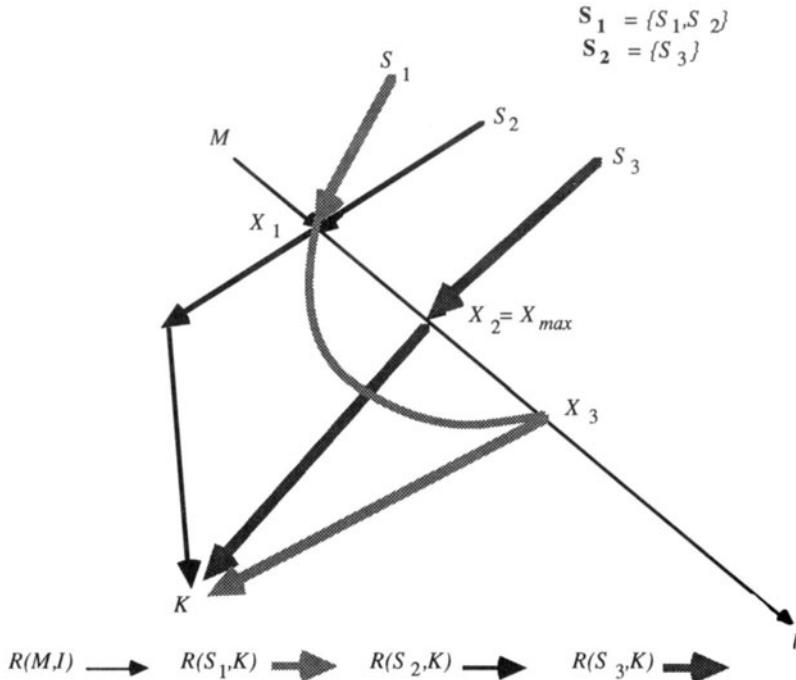


Figure 6

Lemma 6.10.8: In a directed acyclic graph G , if $R(M,I)$ is a directed path, and $\mathbf{R}_K(S)$ is the set of all directed paths to K from a given set of sources S , and there does not exist a vertex Z such that all of the directed paths in $\mathbf{R}_K(S)$ intersect $R(M,I)$ at Z , then there is a pair of directed paths, R and R' , with the following properties: M is the source of R , R' has a source in S ,

either R has sink I and R' has sink K or R has sink K and R' has sink I , and R does not intersect R' .

Proof. If there is a path R' in $\mathbf{R}_K(S)$ that does not intersect $R(M,I)$ the proof is done. Assume then that every path in $\mathbf{R}_K(S)$ intersects $R(M,I)$. Let S'' be the source of a path in S_{\max} (the set of all sources of paths in $\mathbf{R}_K(S)$ whose first intersection with $R(M,I)$ is X_{\max} .) The proof is by induction on the number of distinct vertices in which the paths in $\mathbf{R}_K(S)$ intersect $R(M,I)$.

Base Case: Suppose the antecedent in the statement of the lemma is true. The paths in $\mathbf{R}_K(S)$ intersect $R(M,I)$ in two distinct vertices. There is a path $R(S',K)$ that does not intersect $R(M,I)$ at X_2 ($= X_{\max}$), since otherwise all paths in $\mathbf{R}_K(S)$ would intersect X_2 , contrary to our hypothesis. In addition, $R(S',K)$ does not intersect $R(M,I)$ at any vertex prior to X_1 , since otherwise the paths in $\mathbf{R}_K(S)$ would intersect $R(M,I)$ at more than two distinct vertices, contrary to our hypothesis. Similarly, there is a path $R(S'',K)$ that intersects $R(M,I)$ only at X_2 .

Let $R(X_1,K)$ be a final segment of $R(S',K)$ and $R(S'',X_2)$ an initial segment of $R(S'',K)$. There are two cases.

1. $R(X_1,K)$ does not intersect $R(S'',X_2)$. See figure 7. Let $R(M,X_1)$ be an initial segment of $R(M,I)$, $R(X_2,I)$ be a final segment of $R(M,I)$, $R = R(M,X_1) \& R(X_1,K)$ and $R' = R(S'',X_2) \& R(X_2,I)$. R and R' do not intersect for the following reasons.

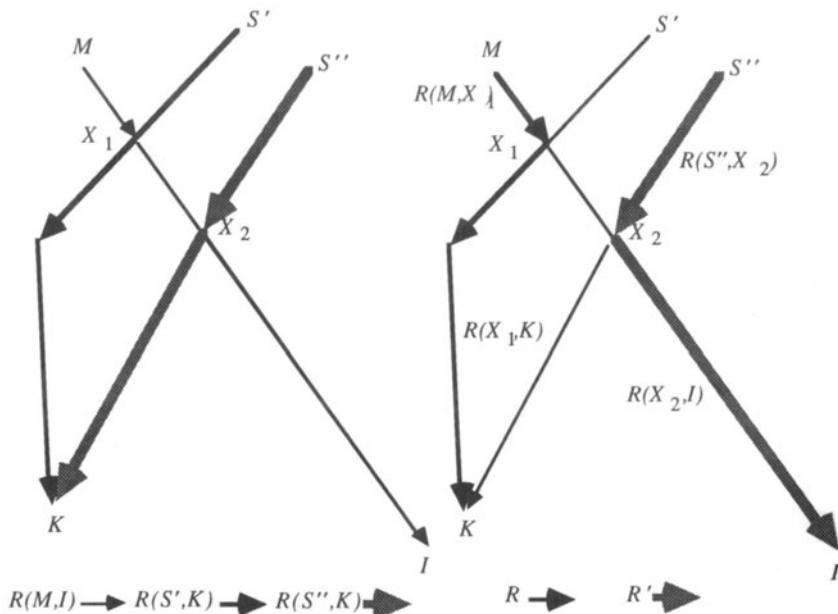


Figure 7

$R(M,X_1)$ does not intersect $R(S'',X_2)$. $R(S'',X_2)$ is a subpath of $R(S'',K)$, which, by hypothesis intersects $R(M,I)$ only at X_2 . Since X_2 occurs after X_1 on $R(M,I)$, X_2 does not occur on $R(M,X_1)$. $R(M,X_1)$ does not intersect $R(X_2,I)$. $R(M,X_1)$ and $R(X_2,I)$ are both subpaths of $R(M,I)$. G is acyclic, and by hypothesis X_1 occurs before X_2 . $R(X_1,K)$ does not intersect $R(S'',X_2)$ by hypothesis. $R(X_1,K)$ does not intersect $R(X_2,I)$. $R(X_1,K)$ is a subpath of $R(S',K)$ and $R(X_2,I)$ is a subpath of $R(M,I)$; by hypothesis $R(S',K)$ intersects $R(M,I)$ only at X_1 , which does not occur on $R(X_2,I)$.

2. $R(X_1,K)$ does intersect $R(S'',X_2)$ at Y . See figure 8. Let $R(S'',Y)$ be an initial segment of $R(S'',K)$, $R(Y,K)$ be a final segment of $R(S',K)$, $R = R(M,I)$ and $R' = R(S'',Y) \& R(Y,K)$. R and R' do not intersect for the following reasons.

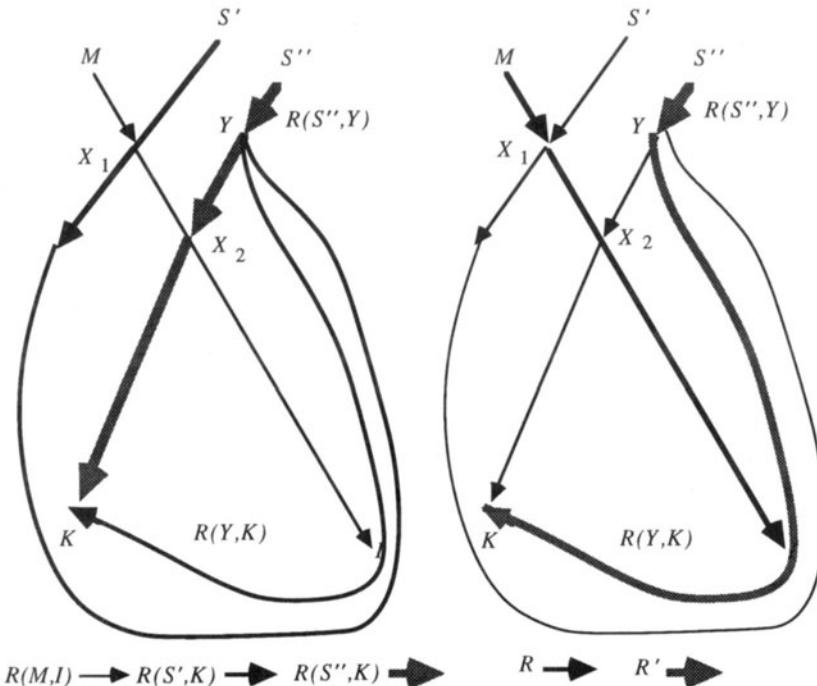


Figure 8

First we will show that $R(M,I)$ does not intersect $R(S'',Y)$. $Y \neq X_2$ since $R(X_1,K)$ intersect $R(M,I)$ only at X_1 . Also, G is acyclic, Y is prior to X_2 on $R(S'',K)$, and X_2 is the first point of intersection of $R(S'',K)$ with $R(M,I)$. Next we will show that $R(M,I)$ does not intersect $R(Y,K)$. Y is on $R(S'',K)$ which does not contain X_1 ; hence Y is not equal to X_1 . It follows that $R(Y,K)$ does not contain X_1 , since Y occurs after X_1 on $R(S',K)$, and $R(S',K)$. By hypothesis $R'(M,K)$ intersects $R(M,I)$ only at X_1 , so that $R(Y,K)$ does not intersect $R(M,I)$ at all.

Induction Case: Assume that the antecedent is true, and that the theorem is true for all $m < n$. If there is a path in $\mathbf{R}_K(S)$ that does not intersect $R(M,I)$, the proof is done. Suppose then that every path in $\mathbf{R}_K(S)$ intersects $R(M,I)$ and that the set of paths in $\mathbf{R}_K(S)$ intersects $R(M,I)$ at exactly n distinct vertices. Let $R(X_{max},I)$ be a final segment of $R(M,I)$. Since not every path in $\mathbf{R}_K(S)$ intersects $R(M,I)$ at X_{max} , there is a point of intersection prior to X_{max} on $R(M,I)$. Hence the number of distinct points of intersection of the paths in $\mathbf{R}_K(S)$ with $R(X_{max},I)$ is

less than n . By the induction hypothesis, there is a path R_1 with source X_{max} and a path R_1' with a source S' in the sources of $\mathbf{R}_K(S)$, such that one of R_1 and R_1' has a sink I , the other has sink K , and R_1 and R_1' do not intersect. Suppose w.l.g. that R_1 has sink I and R_1' has sink K . Since R_1' does not contain X_{max} , its first point of intersection with $R(M,I)$ is some vertex X_r , which occurs on $R(M,I)$ before X_{max} (by definition of X_{max} .) Let $R_1'(X_r,K)$ be a final segment of R_1' , $R(S'',K)$ be a path in $\mathbf{R}_K(S)$ whose first point of intersection with $R(M,I)$ is X_{max} , and $R(S'',X_{max})$ an initial segment of $R(S'',K)$. There are two cases.

1. Assume that $R(X,K)$ does not intersect $R(S'',X_{max})$. Let $R = R(M,X_r) \& R_1'(X_r,K)$ and $R' = R(S'',X_{max}) \& R_1'$. R and R' do not intersect for reasons analogous to those in case 1 of the base case (with X_r substituted for X_1 , and X_{max} substituted for X_2 ; see figure 9.)
2. Assume that $R_1'(X_r,K)$ does intersect $R(S'',X_{max})$, and the last point of intersection is Y . $Y \neq X_{max}$ because it lies on $R_1'(X_r,K)$ and $R_1'(X_r,K)$ does not contain X_{max} . Let $R_1'(Y,K)$ be a final segment of $R_1'(X_r,K)$. There are two cases.
 - a. Assume that $R_1'(Y,K)$ intersects $R(M,X_{max})$ and the first point of intersection is Z . Let $R(S'',Y)$ be an initial segment of $R(S'',X_{max})$, $R(Y,Z)$ an initial segment of $R_1'(Y,K)$, and $R(M,Z)$ an initial segment of $R(M,I)$. $Z \neq X_{max}$ because $R_1'(Y,K)$ does not intersect X_{max} . (See figure 9).

We will now prove Z is not after X_{max} . Consider the path $R(S'',Y) \& R(Y,Z)$. $R(S'',Y)$ does not intersect $R(M,I)$ because Y occurs before X_{max} , $R(S'',Y)$ is an initial segment of $R(S'',K)$ and the first point of intersection of $R(M,I)$ and $R(S'',K)$ is X_{max} . The first point of intersection of $R(Y,Z)$ and $R(M,I)$ is Z , since $R(Y,Z)$ is an initial segment of $R_1'(Y,K)$ and Z is the first point of intersection of $R_1'(Y,K)$ and $R(M,I)$. Hence the first point of intersection of $R(S'',Y) \& R(Y,Z)$ with $R(M,I)$ is Z . $R(S'',Y) \& R(Y,Z)$ is an initial segment of a path from S'' to K that is in $\mathbf{R}_K(S)$. It follows that there is a path in $\mathbf{R}_K(S)$ whose first point of intersection with $R(M,I)$ is Z . If Z is after X_{max} , then there is a path in $\mathbf{R}_K(S)$ whose first point of intersection with $R(M,I)$ is after X_{max} , contrary to the definition of X_{max} .

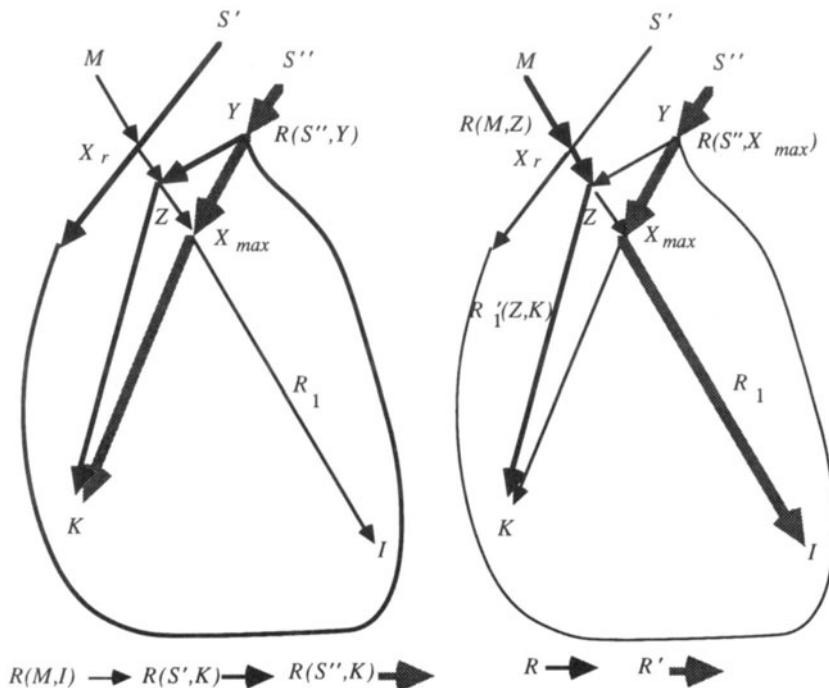


Figure 9

Let $R = R(M, Z) \& R_1'(Z, K)$ and $R' = R(S'', X_{max}) \& R_1$. $R(M, Z)$ does not intersect $R(S'', X_{max})$ since $R(S'', X_{max})$ is an initial segment of $R(S'', K)$ and $R(M, Z)$ is an initial segment of $R(M, I)$ and the first point of intersection of $R(M, I)$ and $R(S'', K)$ is X_{max} . $R(M, Z)$ does not intersect R_1 (which has source X_{max}) since Z occurs before X_{max} and the directed graph is acyclic. $R_1'(Z, K)$ does not intersect R_1 since $R_1'(Z, K)$ is a subpath of R_1' that does not intersect R_1 by construction. $R_1'(Z, K)$ does not intersect $R(S'', X_{max})$ since $R_1'(Z, K)$ is a final segment of $R_1'(X_r, K)$, Z is after Y , and Y is the last point of intersection of $R_1'(X_r, K)$ and $R(S'', X_{max})$.

b. Assume that $R_1'(Y, K)$ does not intersect $R(M, X_{max})$. (This is similar to part 2 of the Base case, with X_{max} substituted for $X2$. See figure 8.) Let $R' = R(S'', Y) \& R_1'(Y, K)$ and $R = R(M, X_{max}) \& R_1$. We have already shown that $R(S'', Y)$ does not intersect $R(M, I)$ and $R(M, X_{max})$ is an initial segment of $R(M, I)$. $R(S'', Y)$ does not intersect R_1 because Y is before

X_{max} , and the directed graph is acyclic. $R_1'(Y,K)$ does not intersect $R(M,X_{max})$ by hypothesis, and $R_1'(Y,K)$ does not intersect R_1 because it is a subpath of R_1' that does not intersect R_1 by construction. \therefore

In an directed acyclic graph G , if all $L(T(K,L))$ and all $J(T(I,J))$ intersect at a vertex Q , then Q is an $LJ(T(I,J),T(K,L))$ **choke point**. Similarly, if all $L(T(K,L))$ and all $J(T(I,J))$ intersect at a vertex Q , and all $L(T(I,L))$ and all $J(T(J,K))$ also intersect at Q , then Q is a $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ **choke point**.

Lemma 6.10.9: In a directed acyclic graph G , if there is no $LJ(T(I,J),T(K,L))$ choke point, then either there is a trek $T(K,L)$ such that there is no vertex V that occurs in the intersection of all $J(T(I,J))$ with $L(T(K,L))$, or there is a trek $T(I,J)$ such that there is no vertex V that occurs in the intersection of all $L(T(K,L))$ with $J(T(I,J))$.

Proof. Suppose that the lemma is false. Then, for each trek $T(K,L)$ there is a non-empty set of points $P(T(K,L))$ such that every point in $P(T(K,L))$ is in the intersection of all $J(T(I,J))$ with $L(T(K,L))$. Similarly, for each trek $T(I,J)$ there is a non-empty set of points $P(T(I,J))$ such that every point in $P(T(I,J))$ is in the intersection of all $L(T(K,L))$ with $J(T(I,J))$. Every $J(T(I,J))$ contains every vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ (since every $J(T(I,J))$ intersects each

$L(T(K,L))$ at some vertex in $P(T(K,L))$), and every vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ occurs on

some trek $L(T(K,L))$. Similarly, every $L(T(K,L))$ contains every vertex in $\bigcup_{T(I,J) \in T(I,J)} P(T(I,J))$.

Furthermore, for every vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ there is some $L(T(K,L))$ that does not

contain it (else all $J(T(I,J))$ and all $L(T(K,L))$ intersect at a single vertex), and some $L(T'(K,L))$ that does contain it. Similarly, for every vertex in $\bigcup_{T(I,J) \in T(I,J)} P(T(I,J))$ there is some

$J(T(I,J))$ that does not contain it and some $J(T'(I,J))$ that does contain it.

Since every vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ occurs on every $J(T(I,J))$, they can be ordered by the

order of their occurrence on some $J(T(I,J))$; similarly every vertex in $\bigcup_{T(I,J) \in T(I,J)} P(T(I,J))$ can be

ordered. By the antecedent of the lemma, there are at least two vertices in each of

$\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ and $\bigcup_{T(I,J) \in T(I,J)} P(T(I,J))$.

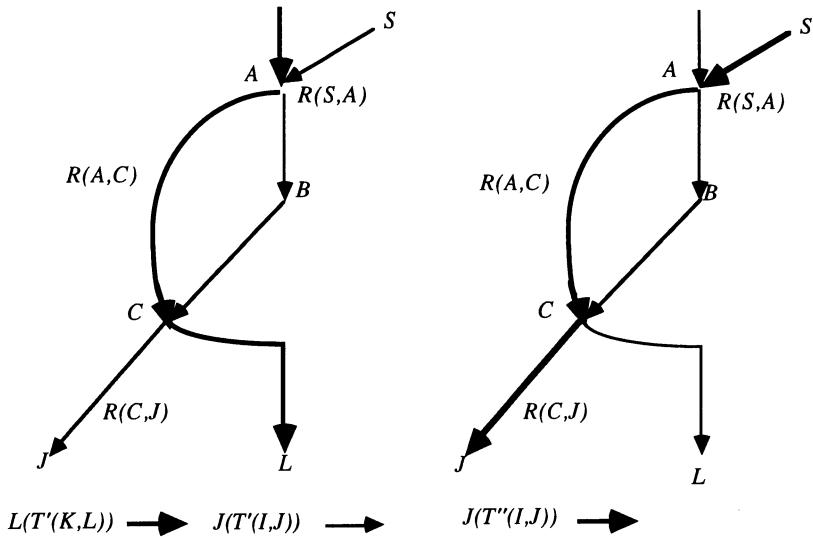


Figure 10

See figure 10. Let A be the first vertex in $\bigcup_{T(I,J) \in T(K,L)} P(T(I,J))$ and B be the first vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$. Suppose w.l.g. that A is before B . There exists an $L(T'(K,L))$ that contains A (since every $L(T(K,L))$ contains A), that does not contain B , but that does contain some vertex C ($\neq B$) in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$.

There is also a $J(T'(I,J))$ that contains A . Let S be the source of $T'(I,J)$, $R(S,A)$ an initial segment of $J(T'(I,J))$, $R(A,C)$ a segment of $L(T'(K,L))$, and $R(C,J)$ a final segment of $J(T'(I,J))$. Let $J(T''(I,J)) = R(S,A) \& R(A,C) \& R(C,J)$, and $I(T''(I,J)) = I(T'(I,J))$. $J(T''(I,J))$ does not contain B for the following reasons. $R(S,A)$ does not contain B because A occurs before B . $R(A,C)$ does not contain B because it is a segment of $L(T'(K,L))$ which does not contain B . $R(C,J)$ does not contain B because it is a segment of $J(T'(I,J))$, and since B is the first vertex in $\bigcup_{T(K,L) \in T(K,L)} P(T(K,L))$ it occurs before C on $J(T'(I,J))$.

$T(K,L) \in T(K,L)$

But this contradicts the fact that for every $T(I,J)$, $J(T(I,J))$ contains B . \therefore

Lemma 6.10.10. In a directed acyclic graph G , if there is no $IK(T(I,J),T(K,L))$ choke point, then either there is a trek $T'(K,L)$ such that there is no vertex V that occurs in the intersection of all $J(T(I,J))$ with $K(T'(K,L))$, or there is a trek $T'(I,J)$ such that there is no vertex V that occurs in the intersection of all $K(T(KL))$ with $I(T(I,J))$

Proof. The proof of lemma 6.10.10 is the same as that of lemma 6.10.9 with I, J, K, L permuted. \therefore

Lemma 6.10.11: In an acyclic LCF G , if there is a trek $T'(K,L)$ such that there is no vertex V that occurs in the intersection of all $J(T(I,J))$ with $L(T'(K,L))$, then either there are treks $T''(I,J)$ and $T''(K,L)$ such that $J(T''(I,J))$ does not intersect $L(T''(K,L))$ or $\rho_{IJ}\rho_{KI} - \rho_{IL}\rho_{JK}$ is not linearly implied by G .

Proof. Let S be the source of $T'(K,L)$, and S' be the set of sources of treks between I and J . By lemma 6.10.8 it is possible to construct a pair of paths R and R' , with sources S and S' (in S), and sinks J and L , such that R and R' do not intersect. There are two cases.

1. If R is a path from S to L , and R' is a path from S' to J , then the following treks can be formed from subpaths of R and R' . (See figure 11.) $J(T''(I,J)) = R'$, $I(T''(I,J)) = I(T'(I,J))$, $K(T''(K,L)) = K(T'(K,L))$, and $L(T''(K,L)) = R$. By construction R does not intersect R' ; hence $J(T''(I,J))$ does not intersect $L(T''(K,L))$.

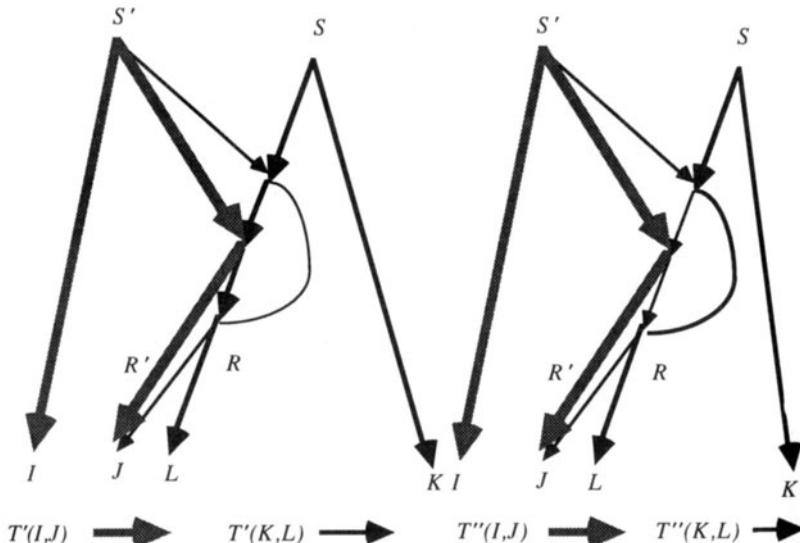


Figure 11

2. If R is a path from S to J , and R' is a path from S' to L , there are two cases.

a. $K(T'(K,L))$ intersects $I(T'(I,J))$, and the first vertex of intersection is Y . Let $R(S,Y)$ be an initial segment of $K(T'(K,L))$, $R(Y,K)$ a final segment of $K(T'(K,L))$, $R(S',Y)$ an initial segment of $I(T'(I,J))$, $R(Y,I)$ a final segment of $I(T'(I,J))$, $J(T''(I,J)) = R$, $I(T''(I,J)) = R(S,Y) \& R(Y,I)$, $K(T''(K,L)) = R(S',Y) \& R(Y,K)$, and $L(T''(K,L)) = R'$. (See figure 12.) By construction, $J(T''(I,J))$ and $L(T''(K,L))$ do not intersect.

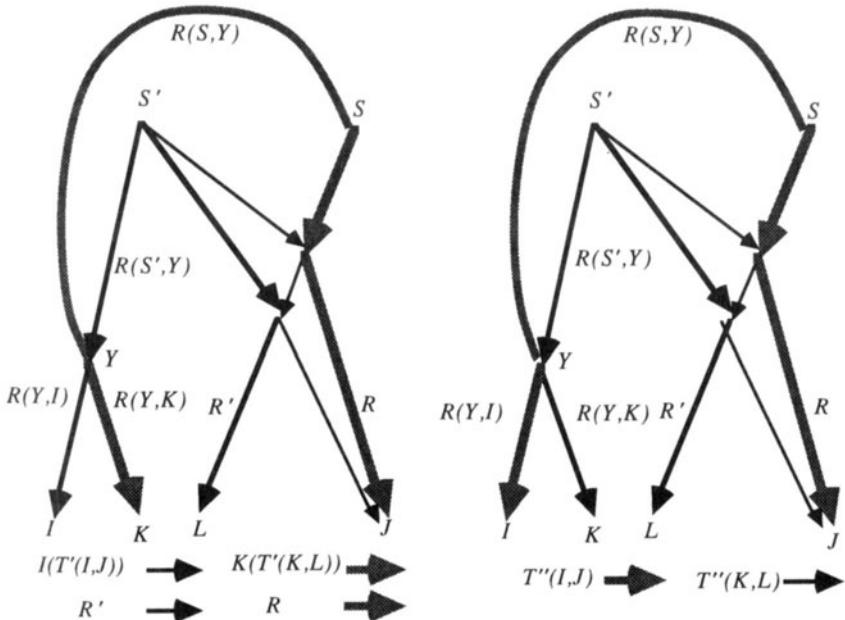


Figure 12

b. If $K(T'(K,L))$ does not intersect $I(T'(I,J))$, the following treks can be formed. (See figure 13.) $I(T'(I,L)) = I(T'(I,J))$, $L(T'(I,L)) = R'$, $J(T'(J,K)) = R$, and $K(T'(J,K)) = K(T'(K,L))$. By hypothesis, $K(T'(J,K))$ does not intersect $I(T'(I,L))$. By construction, $L(T'(I,L))$ does not intersect $J(T'(J,K))$. Hence by lemma 6.10.7, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK}$ is not linearly implied by G . \therefore

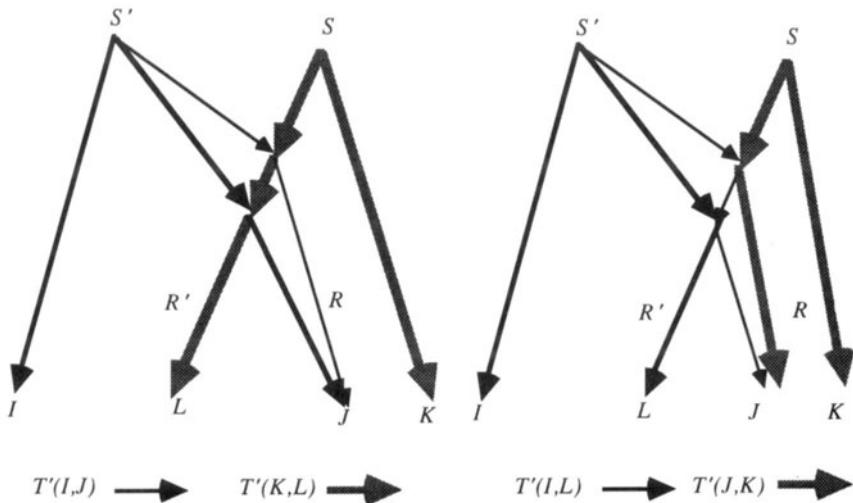


Figure 13

Lemma 6.10.12: In an acyclic LCF G , if there is a trek $T'(I,J)$ such that there is no vertex V' that occurs in the intersection of all $L(T(K,L))$ with $J(T'(I,J))$, then either there are treks $T''(I,J)$ and $T''(K,L)$ such that $J(T''(I,J))$ does not intersect $L(T''(K,L))$ or $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

Lemma 6.10.13: In an acyclic LCF G , if there is a trek $T'(I,J)$ such that there is no vertex V' that occurs in the intersection of all $K(T(K,L))$ with $I(T'(I,J))$, then either there are treks $T''(I,J)$ and $T''(K,L)$ such that $I(T''(I,J))$ does not intersect $K(T''(K,L))$ or $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

Lemma 6.10.14: In an acyclic LCF G , if there is a trek $T'(K,L)$ such that there is no vertex V' that occurs in the intersection of all $I(T(I,J))$ with $K(T'(K,L))$, then either there are treks $T''(I,J)$ and $T''(K,L)$ such that $I(T''(I,J))$ does not intersect $K(T''(K,L))$ or $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

The proofs of lemmas 6.10.12, 6.10.13, and 6.10.14 can all be obtained from the proof of lemma 6.10.11 by permuting I, J, K , and 1.

Lemma 6.10.15: In an acyclic LCF G , if there is no $LJ(T(I,J),T(K,L))$ choke point, and there is no $IK(T(I,J),T(K,L))$ choke vertex, then there exist treks $T'(I,J)$, $T'(K,L)$, $T''(I,J)$, and $T''(K,L)$ such that $I(T'(I,J))$ does not intersect $K(T'(K,L))$ and $J(T''(I,J))$ does not intersect $L(T''(K,L))$, or $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

Proof. This follows directly from lemmas 6.10.9 through 6.10.14. ∴

Lemma 6.10.16: In an acyclic LCF G , if there is no $LJ(T(I,J),T(K,L))$ choke point, and there is no $IK(T(I,J),T(K,L))$ choke point, then $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

Proof. Assume that there is no $LJ(T(I,J),T(K,L))$ choke point, and there is no $IK(T(I,J),T(K,L))$ choke point. By lemma 6.10.15 either $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G or there exist treks $T'(I,J)$, $T'(K,L)$, $T''(I,J)$, and $T''(K,L)$ such that $I(T'(I,J))$ does not intersect $K(T'(K,L))$ and $J(T''(I,J))$ does not intersect $L(T''(K,L))$. If $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G , the proof is done. Assume then that there exist treks $T'(I,J)$, $T'(K,L)$, $T''(I,J)$, and $T''(K,L)$ such that $I(T'(I,J))$ does not intersect $K(T'(K,L))$ and $J(T''(I,J))$ does not intersect $L(T''(K,L))$. There are three cases.

1. Suppose for all $T(I,J)$, $J(T(I,J))$ intersects $L(T'(K,L))$ at each vertex in a non-empty set of vertices \mathbf{P}' , and all $L(T(K,L))$ intersects $J(T'(I,J))$ at each vertex in a non-empty set of vertices \mathbf{P} . Hence, all $L(T(K,L))$ contain every vertex in \mathbf{P} and all $J(T(I,J))$ contain every vertex in \mathbf{P}' . Since there is no $LJ(T(I,J),T(K,L))$ choke point, there is no vertex Z such that for all $T(I,J)$ and all $T(K,L)$, Z occurs in the intersection of $L(T(I,J))$ and $J(T(I,J))$. Hence \mathbf{P} and \mathbf{P}' do not intersect.

Let A be the first vertex in \mathbf{P} , and B be the first vertex in \mathbf{P}' . Suppose w.l.g. that A occurs before B . Let $S'(I,J)$ be the source of $T'(I,J)$, $S'(K,L)$ the source of $T'(K,L)$ and $S''(I,J)$ the source of $T''(I,J)$, and $S''(K,L)$ the source of $T''(K,L)$. $L(T''(K,L))$ contains A (since all $L(T(K,L))$ contain A), and $J(T''(I,J))$ contains B , (since all $J(T(I,J))$ contain B .) There are two cases.

- a. Suppose $K(T''(K,L))$ does not intersect $I(T''(I,J))$. Then, since $K(T''(K,L))$ does not intersect $I(T''(I,J))$ and $J(T''(K,L))$ does not intersect $L(T''(K,L))$, by lemma 6.10.7, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .
- b. Suppose $K(T''(K,L))$ does intersect $I(T''(I,J))$ at a vertex X . (See figure 14.) Let $R(S''(I,J),X)$ be an initial segment of $I(T''(I,J))$, $R(X,K)$ a final segment of $L(T''(K,L))$. Let $R(S''(I,J),B)$ be an initial segment of $J(T''(I,J))$ and $R(B,L)$ be a final segment of $L(T(K,L))$.

Form the trek $K(T''(K,L)) = R(S''(I,J),X) \& R(X,K)$, and $L(T''(K,L)) = R(S''(I,J),B) \& R(B,L)$. $R(S''(I,J),B)$ does not contain A , since it is a subpath of $J(T''(I,J))$ which does not intersect $L(T''(K,L))$, which does contain A . $R(B,L)$ does not contain A , since A occurs before B . Hence $L(T''(K,L))$ does not contain A ; but this is a contradiction.

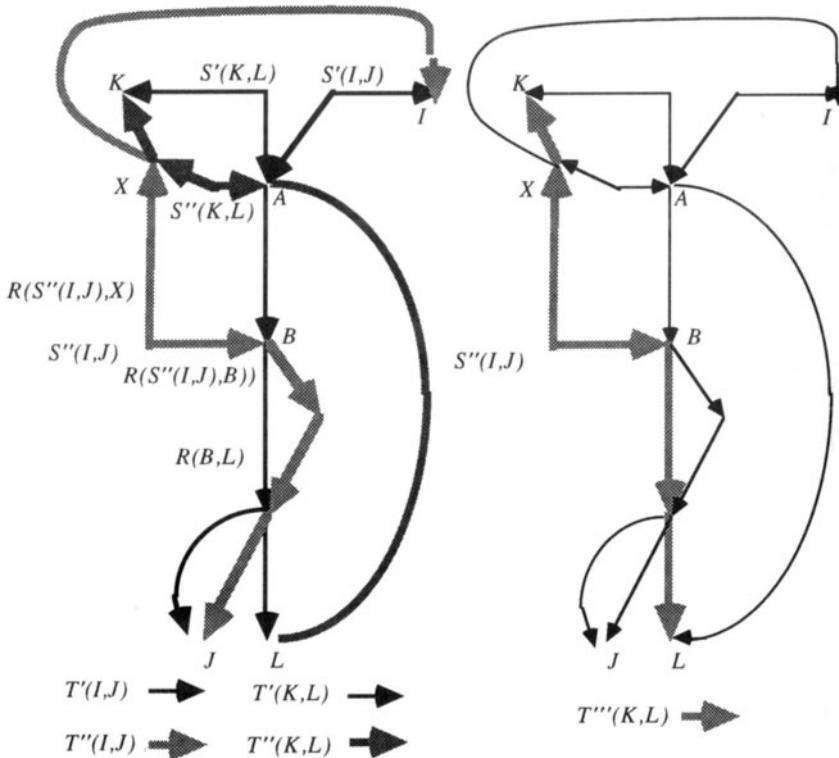
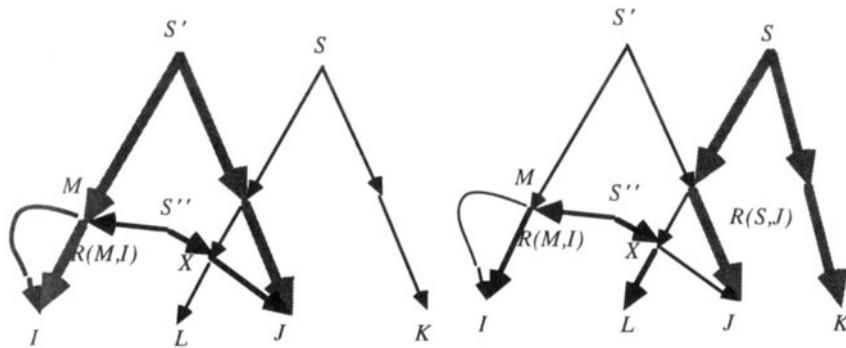


Figure 14

2. All $L(T(K,L))$ intersect $J(T(I,J))$, but not at a single vertex, or all $J(T(I,J))$ intersect $L(T(K,L))$ but not at a single vertex. Assume w.l.g. that the latter is the case. Let S' be the source of $T'(I,J)$ and S be the source of $T'(K,L)$. Let S be the set of sources of treks between I and J . By lemma 6.10.8, it is possible to form two paths $R(S'',L)$ and $R(S,J)$ or $R(S'',J)$ and $R(S,L)$ that don't intersect, where S'' is in S . Assume that it is possible to form the paths $R(S'',L)$ and $R(S,J)$ that don't intersect. (If the paths that don't intersect are $R(S'',J)$ and $R(S,L)$ the proof is the same except that the indices are permuted.) Let $T'(I,J)$ be a trek with

source S'' (See figure 15.) Let the first point of intersection of $I(T'(I,J))$ with $I(T(I,J))$ be M . There are two cases.

- a. Assume that $I(T''(I,J))$ does not intersect $K(T'(K,L))$ before it intersects $I(T'(I,J))$ at M . (See figure 15.) Let $R(M,I)$ be a final segment of $I(T(I,J))$ and $R(S'',M)$ be an initial segment of $I(T''(I,J))$. Let $J(T'(I,L)) = R(S'',M) \& R(M,I)$, $L(T'(I,L)) = R(S'',L)$, $J(T'(J,K)) = R(S,J)$ and $K(T'(J,K)) = K(T'(K,L))$. $R(S'',M)$ and $R(M,I)$ do not intersect $K(T'(K,L))$ by hypothesis. By lemma 6.10.7 $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .



$$T'(I,J) \rightarrow T'(K,L) \rightarrow T''(I,J) \rightarrow \quad T'(I,L) \rightarrow T'(J,K) \rightarrow$$

Figure 15

- b. Assume that $I(t''(I,J))$ does intersect $K(T'(K,L))$ before it intersects $I(T'(I,J))$, and the first point of intersection is Q . Let $R(Q,K)$ be a final segment of $K(T'(K,L))$ and $R(S'',Q)$ be an initial segment of $I(t''(I,J))$. Let Y be the first point of intersection of $R(S,J)$ and $J(T'(I,J))$, and $R(S',Y)$ be an initial segment of $J(T'(I,J))$. There are two cases.

1. Assume that $R(S'',L)$ intersects $R(S',Y)$ and the first point of intersection is Z . Let $R(S',Z)$ be an initial segment of $J(T'(I,J))$, $R(Z,L)$ be a final segment of $R(S'',L)$, $L(T'(I,L)) = R(S',Z) \& R(Z,L)$, $I(T'(I,L)) = I(T'(I,J))$, $J(T'(J,K)) = R(S,J)$, and $K(T'(J,K)) = K(T'(K,L))$. (See figure 16.)

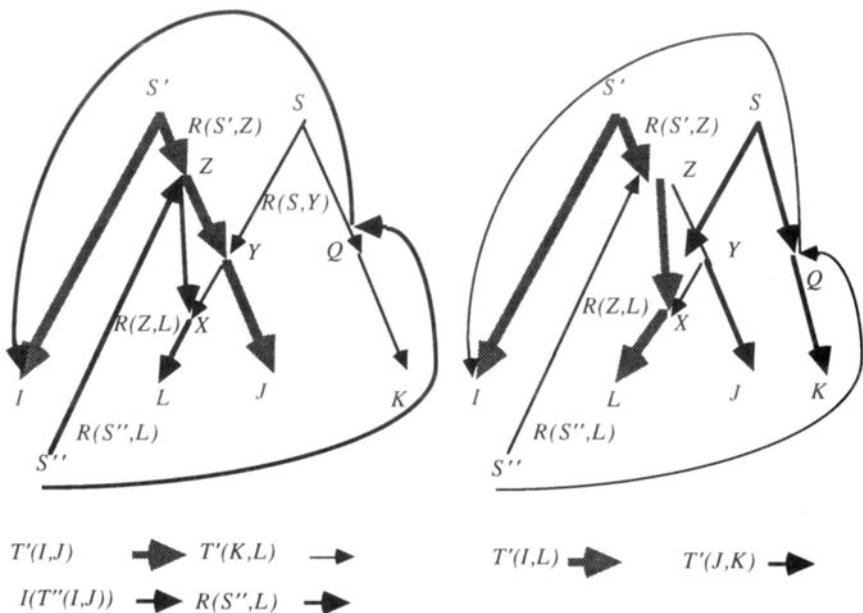


Figure 16

$K(T(J,K))$ does not intersect $I(T'(I,L))$ by hypothesis. $J(T'(J,K))$ does not intersect $L(T'(I,L))$ for the following reasons. $R(S',Z)$ does not intersect $R(S,J)$ because $R(S',Z)$ is a subpath of $J(T'(I,J))$, Z is before Y , and the first point of intersection of $J(T'(I,J))$ and $R(S,J)$ is Y . $R(Z,L)$ does not intersect $R(S,J)$ because it is a subpath of $R(S'',L)$ which does not intersect $R(S,J)$ by construction. By lemma 6.10.7. $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

2. Assume that $R(S'',L)$ does not intersect $R(S',Y)$. Let $L(T''(K,L)) = R(S'',L)$, $K(T''(K,L)) = R(S'',Q) \& R(Q,K)$, $I(T''(I,J)) = I(T'(I,J))$, and $J(T''(I,J)) = R(S',Y) \& R(Y,J)$. (See figure 17.) $K(T''(K,L))$ does not intersect $I(T''(I,J))$ for the following reasons. $R(S',Q)$ does not intersect $I(T'(I,J))$ since $R(S',Q)$ is an initial segment of $I(T'(I,J))$, and Q occurs before the first point of intersection of $I(T'(I,J))$ and $I(T'(I,J))$. $R(Q,K)$ does not intersect $I(T'(I,J))$ because it is a final segment of $K(T''(K,L))$, which does not intersect $I(T'(I,J))$ by hypothesis. $L(T''(K,L))$ does not intersect $J(T''(I,J))$ for the following reasons. $R(S',Y)$ does not intersect $R(S'',L)$ by hypothesis, and $R(Y,J)$ is a subpath of $R(S,J)$ which does not intersect $R(S'',L)$ by construction. By lemma 6.10.7 $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

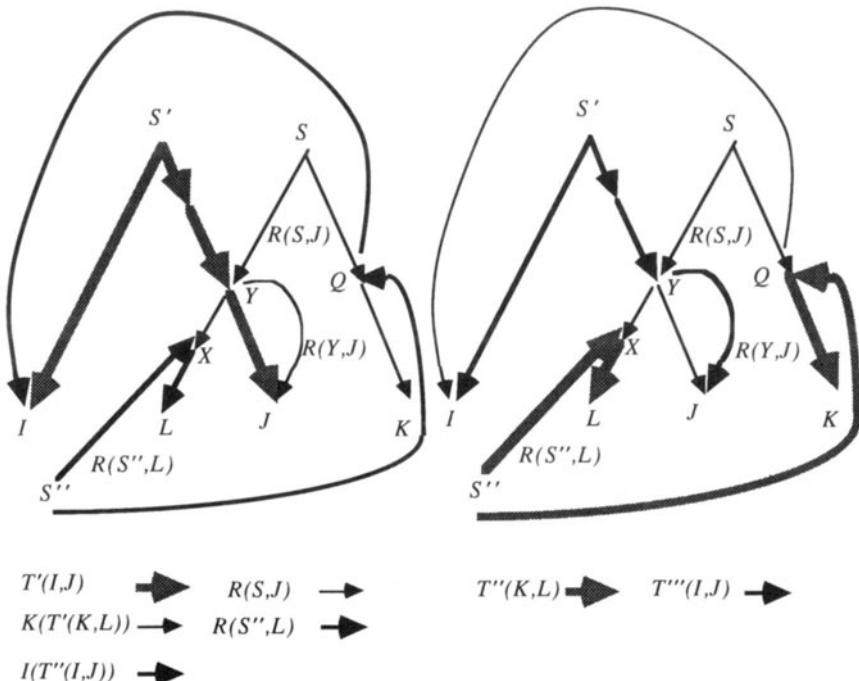


Figure 17

3. Either there is an $L(T''(K,L))$ that does not intersect $J(T''(I,J))$ or there is a $J(T''(I,J))$ that does not intersect $L(T''(K,L))$. Assume w.l.g. that $J(T''(I,J))$ with source $S''(I,J)$ does not intersect $L(T''(K,L))$. There are two cases.

a. Suppose that $I(T''(I,J))$ does not intersect $K(T''(K,L))$ before it intersects $I(T''(I,J))$ at vertex X . See figure 18.

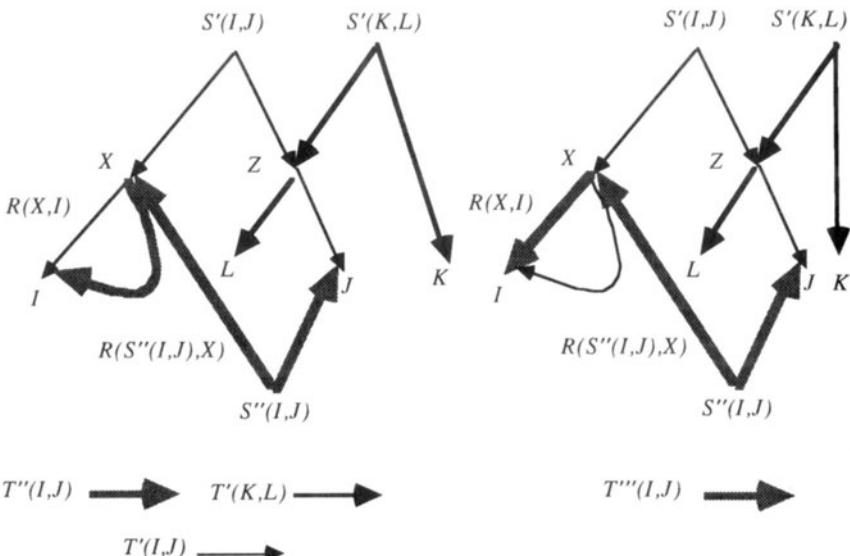


Figure 18

Let $R(X,I)$ be a final segment of $I(T'(I,J))$ and $R(S''(I,J),X)$ be an initial segment of $I(T''(I,J))$. The trek $T'''(I,J)$ can be formed as follows. $J(T''(I,J)) = J(T''(I,J))$ and $I(T''(I,J)) = R(S''(I,J),X) \& R(X,I)$. $R(S''(I,J),X)$ does not intersect $K(T'(K,L))$ because by hypothesis X occurs on $I(T''(I,J))$ before it intersects $K(T'(K,L))$. $R(X,I)$ does not intersect $K(T'(K,L))$ because it is a subpath of $I(T'(I,J))$ which does not intersect $K(T'(K,L))$ by hypothesis. Hence $I(T''(I,J))$ does not intersect $K(T'(K,L))$. $J(T''(I,J)) = J(T''(I,J))$ does not intersect $L(T'(K,L))$ by hypothesis. By lemma 6.10.7, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

b. Suppose $I(T'(I,J))$ intersects $K(T'(I,J))$ at Y before it intersects $I(T'(I,J))$ at X . Let Z be the first point of intersection of $J(T'(I,J))$ and $L(T'(K,L))$. (If no such vertex exists, then $J(T'(I,J))$ and $L(T'(K,L))$ do not intersect, $I(T'(I,J))$ and $K(T'(K,L))$ do not intersect by hypothesis, and by lemma 6.10.7 $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .) Let $R(S'(I,J),Z)$ be an initial segment of $I(T'(I,J))$, and $R(Z,L)$ be a final segment of $L(T'(K,L))$. There are two cases.

1. Suppose that $J(T'(I,J))$ does not intersect $R(S'(I,J),Z)$. See figure 19.

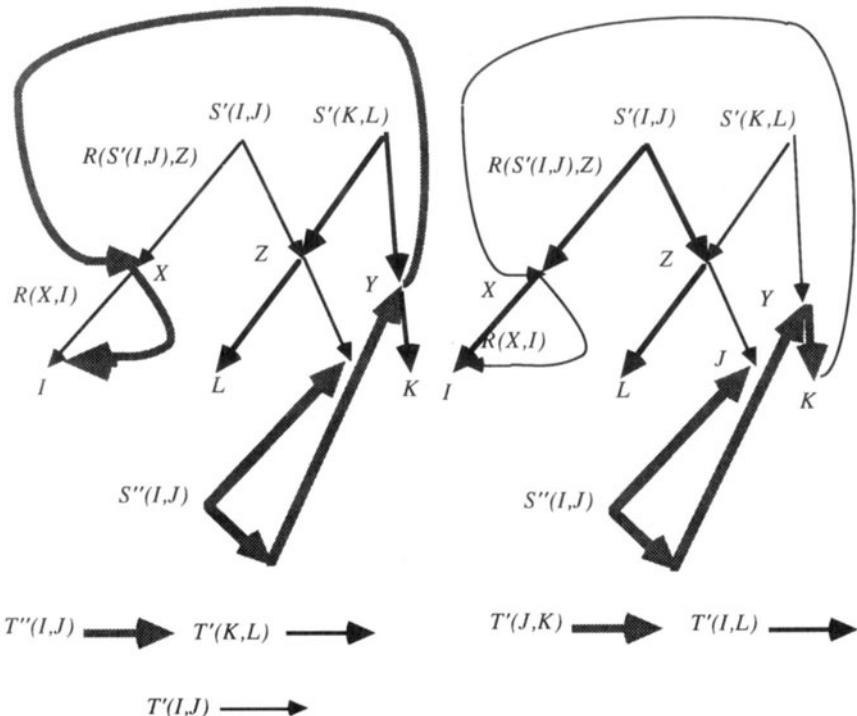


Figure 19

Let $R(Y,K)$ be a final segment of $K(T'(K,L))$ and $R(S''(I,J),Y)$ be an initial segment of $I(T''(I,J))$. Let $J(T'(J,K)) = J(T''(I,J))$, $K(T'(J,K)) = R(S''(I,J),Y) \& R(Y,K)$, $I(T'(I,L)) = I(T''(I,J))$, $L(T'(I,L)) = R(S'(I,J),Z) \& R(Z,L)$. $I(T'(I,L))$ and $K(T'(J,K))$ do not intersect for the following reasons. $I(T'(I,L))$ does not intersect $R(S''(I,J),Y)$ because by hypothesis, $I(T''(I,J))$ intersects $K(T'(K,L))$ at Y before it intersects $I(T'(I,J))$. $I(T'(I,L))$ does not intersect $R(Y,K)$ because $I(T'(I,L)) = I(T''(I,J))$ and $R(Y,K)$ is a subpath of $K(T'(K,L))$, which does not intersect $I(T''(I,J))$ by hypothesis. $J(T'(J,K))$ does not intersect $L(T'(I,L))$ for the following reasons. $J(T'(J,K))$ does not intersect $R(S'(I,J),Z)$ because $J(T'(J,K)) = J(T''(I,J))$, which does not intersect $R(S'(I,J),Z)$ by hypothesis. $J(T'(J,K))$ does not intersect $R(Z,L)$ because $J(T'(J,K)) = J(T''(I,J))$ which does not intersect $L(T'(K,L))$ (which contains $R(Z,L)$) by hypothesis. By lemma 6.10.7, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

2. Suppose that $J(T''(I,J))$ does intersect $R(S'(I,J),Z)$ and the first point of intersection is M . (See figure 20.) $M \neq Z$ because $J(T''(I,J))$ does not intersect $L(T(K,L))$ which contains Z . Let $R(S'(I,J),M)$ be an initial segment of $J(T'(I,J))$ and $R(M,J)$ be a final segment of $J(T''(I,J))$. Let $I(T'''(I,J)) = I(T''(I,J))$ and $J(T'''(I,J)) = R(S'(I,J),M) \& R(M,J)$. $I(T'''(I,J))$ does not intersect $K(T(K,L))$ by hypothesis. $J(T'''(I,J))$ does not intersect $L(T(K,L))$ for the following reasons. $R(S'(I,J),M)$ does not intersect $L(T(K,L))$ since M is before Z on $J(T'(I,J))$, and the first point of intersection of $J(T'(I,J))$ with $L(T(K,L))$ is Z . $R(M,J)$ does not intersect $L(T(K,L))$ because it is a subpath of $J(T''(I,J))$ which does not intersect $L(T(K,L))$ by hypothesis. By lemma 6.10.7, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

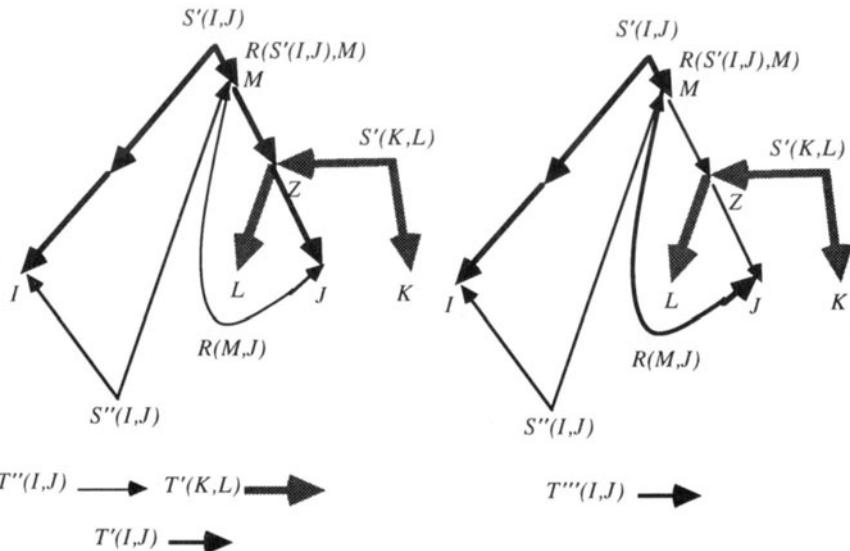


Figure 20

 \therefore

Lemma 6.10.17: In an acyclic LCF G , if there is no $LJ(T(I,L),T(J,K))$ choke point, and there is no $IK(T(I,L),T(J,K))$ choke point, then $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G .

Proof. The proof is the same as that of lemma 6.10.16, with the indices permuted. \therefore

Lemma 6.10.18: In an acyclic LCF G , if G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, then either there is an $LJ(T(I,J),T(K,L))$ choke point and an $LJ(T(I,L),T(J,K))$ choke point, or there is an $IK(T(I,J),T(K,L))$ choke point and an $IK(T(I,L),T(J,K))$ choke point.

Proof. Assume that G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$. By lemmas 6.10.16 and 6.10.17, if G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ then either there is an $LJ(T(I,J),T(K,L))$ choke point or an $IK(T(I,J),T(K,L))$ choke point, and there is either an $LJ(T(I,L),T(J,K))$ choke point or an $IK(T(I,L),T(J,K))$ choke point. If there is an $LJ(T(I,J),T(K,L))$ choke point and an $LJ(T(I,L),T(J,K))$ choke point, or there is an $IK(T(I,J),T(K,L))$ choke point and an $IK(T(I,L),T(J,K))$ choke point, the proof is done. Suppose then that there is an $LJ(T(I,J),T(K,L))$ choke point and an $IK(T(I,L),T(J,K))$ choke point, but no $IK(T(I,J),T(K,L))$ choke point and no $LJ(T(I,L),T(J,K))$ choke point. (The case where there is an $LJ(T(I,L),T(J,K))$ choke point and an $IK(T(I,J),T(K,L))$ choke point, but no $LJ(T(I,J),T(K,L))$ choke point and no $IK(T(I,L),T(J,K))$ choke point is essentially the same, with the indices permuted.)

By lemmas 6.10.9 through 6.10.14, if there is no $LJ(T(I,L),T(J,K))$ choke point, then either there is a pair of treks $T'(I,L)$ and $T'(J,K)$ such that $L(T'(I,L))$ does not intersect $J(T'(J,K))$ or $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ is not linearly implied by G . Since the latter possibility contradicts our hypothesis, assume that there is a pair of treks $T'(I,L)$ and $T'(J,K)$ such that $L(T'(I,L))$ does not intersect $J(T'(J,K))$. There are two cases.

If $I(T'(I,L))$ does not intersect $K(T'(J,K))$ then by lemma 6.10.7, G does not linearly imply $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, contrary to our hypothesis. Suppose then that $I(T'(I,L))$ does intersect $K(T'(J,K))$ at a vertex Y . (See figure 21.)

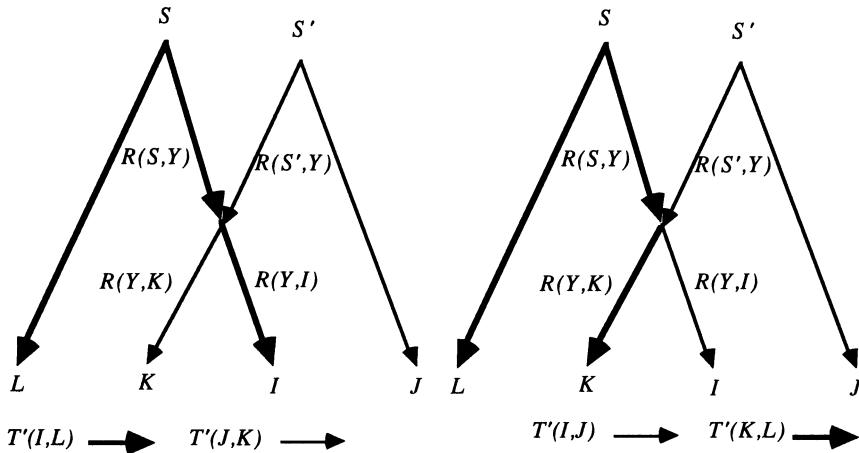


Figure 21

Let S be the source of $T'(I,L)$, S' the source of $T'(J,K)$, $R(S,Y)$ an initial segment of $I(T'(I,L))$, $R(Y,K)$ a final segment of $K(T'(J,K))$, $R(S',Y)$ an initial segment of $K(T'(J,K))$, $R(Y,I)$ a final segment of $I(T'(I,L))$, $I(T'(I,J)) = R(S',Y) \& R(Y,I)$, $J(T'(I,J)) = J(T'(J,K))$, $K(T'(K,L)) = R(S,Y) \& R(Y,K)$, and $L(T'(K,L)) = L(T'(I,L))$. But since $J(T'(I,J)) = J(T'(J,K))$ does not intersect $L(T'(K,L)) = L(T'(I,L))$, there is no $LJ(T(I,J),T(K,L))$ choke point, contrary to our hypothesis. ∴.

Lemma 6.10.19: In an acyclic LCF G , if G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, then either there is an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point, or there is an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point.

Proof. Assume that G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$. By lemma 6.10.18, either there is an $LJ(T(I,J),T(K,L))$ choke point and an $LJ(T(I,L),T(J,K))$ choke point, or there is an $IK(T(I,J),T(K,L))$ choke point and an $IK(T(I,L),T(J,K))$ choke point. Suppose w.l.g. that the former is the case. If some $LJ(T(I,J),T(K,L))$ choke point is also an $LJ(T(I,L),T(J,K))$ choke point, the proof is done. Suppose then that no $LJ(T(I,J),T(K,L))$ choke point is also an $LJ(T(I,L),T(J,K))$ choke point. Let C be an $LJ(T(I,J),T(K,L))$ choke point. By hypothesis C is not an $LJ(T(I,L),T(J,K))$ choke point, so there exist a pair of treks $T'(I,L)$ and $T'(J,K)$ with sources S and S' respectively, such that $L(T'(I,L))$ and $J(T'(J,K))$ do not intersect at C . (See figure 22.)

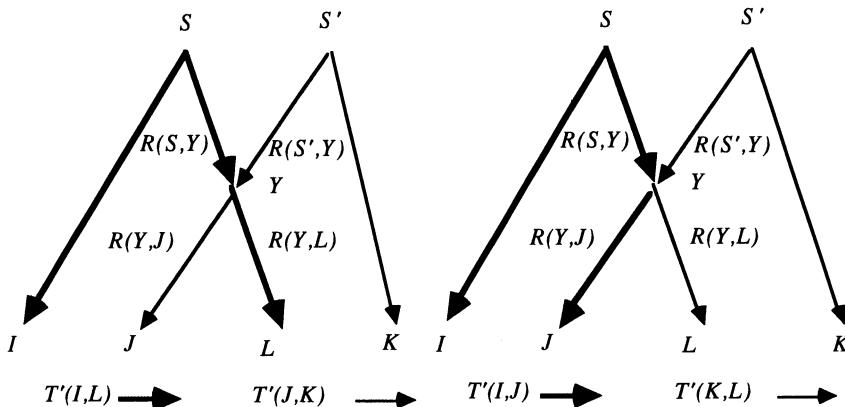


Figure 22

Hence there is at most one occurrence of C in the pair of paths $L(T(I,L))$ and $J(T(J,K))$. Since there is an $LJ(T(I,L),T(J,K))$ choke point, $L(T(I,L))$ and $J(T(J,K))$ intersect at a point Y . Let $R(S,Y)$ be an initial segment of $L(T(I,L))$, $R(Y,J)$ be a final segment of $J(T(J,K))$, $R(S',Y)$ an initial segment of $J(T(J,K))$, $R(Y,L)$ a final segment of $L(T(I,L))$, $I(T(I,J)) = I(T(I,L))$, $J(T(I,J)) = R(S,Y) \& R(Y,J)$, $K(T(K,L)) = K(T(J,K))$ and $L(T(K,L)) = R(S',Y) \& R(Y,L)$. Since $L(T(K,L))$ and $J(T(I,J))$ are rearrangements of the vertices in $J(T(J,K))$ and $L(T(I,L))$, the number of occurrences of any vertex in $L(T(K,L))$ and $J(T(I,J))$ is less than or equal to the number of occurrences of that vertex in $J(T(J,K))$ and $L(T(I,L))$. Since C occurs at most once in $J(T(J,K))$ and $L(T(I,L))$, it occurs at most once in $L(T(K,L))$ and $J(T(I,J))$. Hence $L(T(K,L))$ and $J(T(I,J))$ do not intersect at C , contrary to the hypothesis that C is an $LJ(T(I,J),T(K,L))$ choke point. \therefore

Lemma 6.10.20: For any probability distribution over a set of random variables \mathbf{W} , if there exists a subset \mathbf{P} of \mathbf{V} such that $\rho_{IJ} \cdot \rho_{PKL} - \rho_{IL} \cdot \rho_{PJK} = 0$, and for all variables U in \mathbf{P} and all subsets \mathbf{V} of \mathbf{P} not containing U , either $\rho_{IU} \cdot \mathbf{v} = 0$ and $\rho_{KU} \cdot \mathbf{v} = 0$, or $\rho_{JU} \cdot \mathbf{v} = 0$ and $\rho_{LU} \cdot \mathbf{v} = 0$, then $\rho_{IJ} \rho_{PKL} - \rho_{IL} \rho_{PJK} = 0$.

Proof. The proof is by induction on the cardinality of \mathbf{P} .

Base Case: Suppose the cardinality of \mathbf{P} is zero. Then $\rho_{IJ} \rho_{PKL} - \rho_{IL} \rho_{PJK} = 0$ is equivalent to $\rho_{IJ} \cdot \rho_{PKL} - \rho_{IL} \cdot \rho_{PJK} = 0$.

Induction Case: Suppose that the lemma is true for all sets of cardinality n or less. Let \mathbf{P} have cardinality $n+1$. Assume that $\rho_{IJ,\mathbf{P}}\rho_{KL,\mathbf{P}} - \rho_{IL,\mathbf{P}}\rho_{JK,\mathbf{P}} = 0$.

Let Y be a variable in \mathbf{P} , and $\mathbf{P}' = \mathbf{P} - \{Y\}$. Since $\rho_{IJ,\mathbf{P}}\rho_{KL,\mathbf{P}} - \rho_{IL,\mathbf{P}}\rho_{JK,\mathbf{P}}$, by the recursion formula for partial correlation,

$$\left(\frac{\rho_{IJ,\mathbf{P}'} - \rho_{IY,\mathbf{P}}\rho_{JY,\mathbf{P}'}^*}{(\sqrt{1-\rho_{IY,\mathbf{P}}^2})(\sqrt{1-\rho_{JY,\mathbf{P}'}^2})} \right) \left(\frac{\rho_{KL,\mathbf{P}'} - \rho_{KY,\mathbf{P}}\rho_{LY,\mathbf{P}'}^*}{(\sqrt{1-\rho_{KY,\mathbf{P}}^2})(\sqrt{1-\rho_{LY,\mathbf{P}'}^2})} \right) =$$

$$\left(\frac{\rho_{IL,\mathbf{P}'} - \rho_{IY,\mathbf{P}}\rho_{LY,\mathbf{P}'}^*}{(\sqrt{1-\rho_{IY,\mathbf{P}}^2})(\sqrt{1-\rho_{LY,\mathbf{P}'}^2})} \right) \left(\frac{\rho_{JK,\mathbf{P}'} - \rho_{JY,\mathbf{P}}\rho_{KY,\mathbf{P}'}^*}{(\sqrt{1-\rho_{JY,\mathbf{P}}^2})(\sqrt{1-\rho_{KY,\mathbf{P}'}^2})} \right)$$

The denominator of the l.h.s. equals the denominator of the r.h.s., so the numerator of the l.h.s. equals the numerator of the r.h.s. Expanding the numerators of each side,

$$\begin{aligned} & \rho_{IJ,\mathbf{P}}\rho_{KL,\mathbf{P}'} - \rho_{IJ,\mathbf{P}}\rho_{KY,\mathbf{P}'}\rho_{LY,\mathbf{P}'} - \rho_{KL,\mathbf{P}'}\rho_{IY,\mathbf{P}}\rho_{JY,\mathbf{P}'} - \rho_{IY,\mathbf{P}}\rho_{JY,\mathbf{P}'}\rho_{KY,\mathbf{P}'}\rho_{LY,\mathbf{P}'} = \\ & \rho_{IL,\mathbf{P}}\rho_{JK,\mathbf{P}'} - \rho_{IL,\mathbf{P}}\rho_{JY,\mathbf{P}}\rho_{KY,\mathbf{P}'} - \rho_{JK,\mathbf{P}'}\rho_{IY,\mathbf{P}}\rho_{LY,\mathbf{P}'} - \rho_{IY,\mathbf{P}}\rho_{JY,\mathbf{P}'}\rho_{KY,\mathbf{P}'}\rho_{LY,\mathbf{P}'} \end{aligned}$$

The fourth terms on both sides are equal. By hypothesis, either $\rho_{IY,\mathbf{P}'} = \rho_{KY,\mathbf{P}'} = 0$, or $\rho_{JY,\mathbf{P}'} = \rho_{LY,\mathbf{P}'} = 0$. In either case, the second and third terms on each side are equal to zero. It follows that $\rho_{IJ,\mathbf{P}}\rho_{KL,\mathbf{P}'} - \rho_{IL,\mathbf{P}}\rho_{JK,\mathbf{P}'} = 0$. Since \mathbf{P}' has one less member than \mathbf{P} , by the induction hypothesis, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$. ∴

Lemma 6.10.21: In an acyclic LCF G , if there exists an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point, then G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

Proof. Suppose w.l.g. that X is the last $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point. There are two cases.

First consider the case where there is no trek between at least one of the pairs I and J , and K and L , and there is no trek between at least one of the pairs I and L , and J and K . It follows that at least one of ρ_{IJ} and ρ_{KL} equals 0, and at least one of ρ_{IL} and ρ_{JK} is equal to zero. Hence $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

Next suppose w.l.g. that there are treks $T'(I,J)$ and $T'(K,L)$. We will prove that $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ by proving that there exists a set \mathbf{Q}' of variables such that $\rho_{IJ,Q'}\rho_{KL,Q'} - \rho_{IL,Q'}\rho_{JK,Q'} = 0$.

$\rho_{IL.Q'}\rho_{JK.Q'} = 0$, and for all variables U in Q' and all subsets V of Q' not containing U , either $\rho_{IU.V} = 0$ and $\rho_{KU.V} = 0$, or $\rho_{JU.V} = 0$ and $\rho_{LU.V} = 0$, and applying lemma 6.10.20.

Let $Q = \{\text{sources of treks between } X \text{ and } J \text{ or } X \text{ and } L\}$. Since X is on $J(T'(I,J))$ and $L(T'(K,L))$, and by definition the sink of $J(T'(I,J))$ is J , and the sink of $L(T'(K,L))$ is L , there are directed paths $R(X,J)$ and $R(X,L)$; hence X is in Q . We will now demonstrate that $I \perp\!\!\!\perp J|Q$ by showing that I and J are d-separated given Q . We will show that I and J are d-separated given Q by showing that every undirected path between I and J either contains a vertex V that is a collider that is not the source of a directed path from V to any vertex in Q , or it contains some vertex in Q that is not a collider.

Consider first the undirected paths between I and J without colliders. If there is an undirected path with no collider between I and J that does not contain X , there is a trek between I and J that does not contain X . But, every $T(I,J)$ contains X , since X is a choke point. Hence, there does not exist an undirected path between I and J without colliders that does not contain X . Since X is in Q , every undirected path that does not contain a collider contains a vertex in Q .

Consider now undirected paths between I and J that contain colliders. If some vertex W is a collider and is not the source of a directed path from W to some vertex in Q , the proof is done. Suppose then that every vertex W that is a collider is the source of a directed path from W to some vertex in Q . Consider w.l.g. an arbitrary undirected path $R(J,I)$ from J to I . Let Z be the first vertex on $R(J,I)$ that is a collider. By hypothesis, there is a directed path $R(Z,U)$ where U is a vertex in Q . Since the undirected path from J to Z does not contain any colliders, there is a vertex S that is the source of a pair of directed paths $R(S,J)$ and $R(S,Z)$. Since Z has an edge directed into it, $S \neq Z$. There are two cases.

- a. $S = J$. See figure 23. There is a directed path $R(J,Z)$. There is a directed path $R(Z,U)$. Since U is the source of a trek between X and J , there is a directed path $R(U,X)$. We have already shown that there is a directed path $R(X,J)$. Hence there is a cyclic path $R(J,Z) \& R(Z,U) \& R(U,X) \& R(X,J)$.

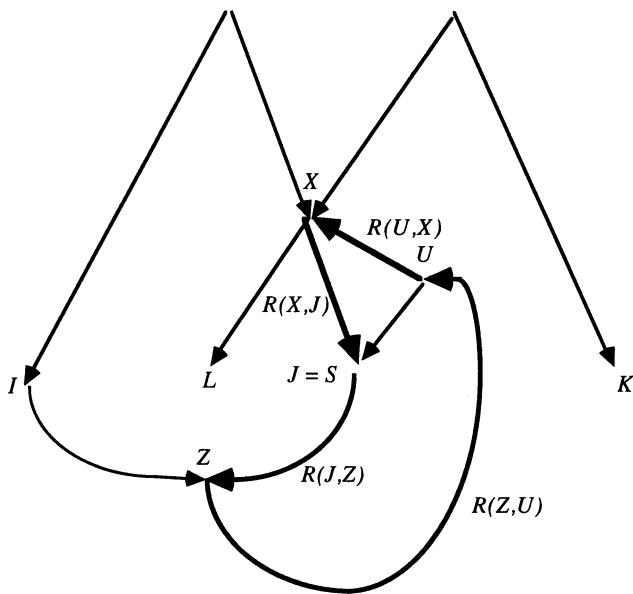


Figure 23

b. $S \neq J$. See figure 24. There is a directed path $R(S,J)$, and a directed path $R(S,Z) \& R(Z,U) \& R(U,X)$. By lemma 6.10.2 there is a trek $T(J,X)$ with source M , where M is the last point of intersection of $R(S,J)$ and $R(S,Z) \& R(Z,U) \& R(U,X)$, and $J(T(J,X))$ is a subpath of $R(S,J)$. Since M is on $R(S,J)$, and S occurs before Z on $R(J,I)$, M occurs before Z on $R(J,I)$. Hence there is no collision at M in $R(J,I)$. Also, M is in Q , since it is the source of a trek between X and J . The undirected path $R(J,I)$ contains a vertex in Q that is not a collider.

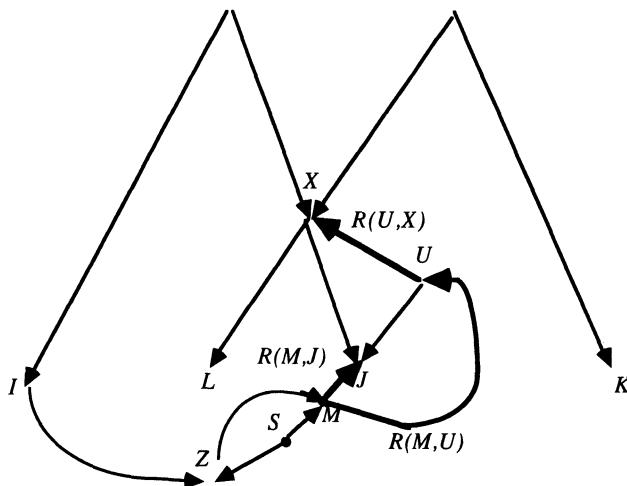


Figure 24

In either case \mathbf{Q} d-separates X and Y , so $I \perp\!\!\!\perp J | \mathbf{Q}$. Similarly, it can be shown that $K \perp\!\!\!\perp L | \mathbf{Q}, J \perp\!\!\!\perp L | \mathbf{Q} \perp\!\!\!\perp K | \mathbf{Q}$. It follows that $\rho_{IJ|\mathbf{Q}} = 0$, $\rho_{KL|\mathbf{Q}} = 0$, $\rho_{IL|\mathbf{Q}} = 0$, and $\rho_{JK|\mathbf{Q}} = 0$. Let $\mathbf{Q}' = \mathbf{Q} - \{X\}$. By the recursion formula for partial correlation, $\rho_{IJ|\mathbf{Q}'} = \rho_{IX|\mathbf{Q}'} \rho_{JX|\mathbf{Q}'}$, $\rho_{KL|\mathbf{Q}'} = \rho_{KX|\mathbf{Q}'} \rho_{LX|\mathbf{Q}'}$, $\rho_{IL|\mathbf{Q}'} = \rho_{IX|\mathbf{Q}'} \rho_{LX|\mathbf{Q}'}$, and $\rho_{JK|\mathbf{Q}'} = \rho_{JX|\mathbf{Q}'} \rho_{KX|\mathbf{Q}'}$. Hence $\rho_{IJ|\mathbf{Q}'} \rho_{KL|\mathbf{Q}'} = \rho_{IX|\mathbf{Q}'} \rho_{JX|\mathbf{Q}'} \rho_{KX|\mathbf{Q}'} \rho_{LX|\mathbf{Q}'} = \rho_{IX|\mathbf{Q}'} \rho_{LX|\mathbf{Q}'} \rho_{JX|\mathbf{Q}'} \rho_{KX|\mathbf{Q}'} = \rho_{IL|\mathbf{Q}'} \rho_{JK|\mathbf{Q}'} = \rho_{IL|\mathbf{Q}'} \rho_{JK|\mathbf{Q}'}$.

We will next demonstrate that for each variable U in \mathbf{Q}' , and each subset \mathbf{V} of \mathbf{Q}' not containing U , $I \perp\!\!\!\perp U | \mathbf{V}$, by showing that I and U are d-separated given \mathbf{V} . We will show that I and U are d-separated given \mathbf{V} by showing that every undirected path between I and U either contains a vertex W that is a collider that is not the source of a directed path from W to any vertex in \mathbf{V} , or it contains some vertex in \mathbf{V} that is not a collider.

For U in \mathbf{Q}' , consider an arbitrary undirected path $R(I,U)$ that contains colliders. Let Z be the first point of $R(I,U)$ after I that is a collider, and $R(I,Z)$ be an initial segment of $R(I,U)$. If Z is not the source of a path to some vertex M in \mathbf{Q}' , then I and U are d-separated given \mathbf{Q}' , and the proof is done. Suppose then that there is a directed path $R(Z,M)$ to some M in \mathbf{Q}' . Since $R(I,Z)$ contains no colliders, there is a vertex S on $R(I,Z)$ that is the source of directed paths $R(S,I)$ and $R(S,Z)$. Hence S is the source of directed paths to I and M , $R(S,I)$ and $R(S,M) = R(S,Z) \& R(Z,M)$ respectively. (If $R(I,U)$ is an undirected path that contains no colliders, then it still follows that there is a vertex S on $R(I,U)$ that is the source of directed paths $R(S,I)$ and

$R(S,U)$) M is either the source of a trek between X and J or X and L . Suppose w.l.g. that M is the source of a trek between X and J . Then M is the source of a directed path $R(M,J)$ and a directed path $R(M,X)$. M does not equal X by hypothesis. Hence $R(M,J)$ does not contain X , since $R(M,J)$ is a branch of a trek between J and X , and the two branches of the trek intersect only at M . $R(S,M)$ does not contain X , else there is a cycle. Either $R(S,I)$ contains X or it doesn't.

- a. If $R(S,I)$ does not contain X , then there are a pair of paths, $R(S,I)$ and $R(S,Z) \& R(Z,M) \& R(M,J)$ that do not contain X . See figure 25. Hence there is a trek between I and J that does not contain X . This contradicts the assumption that X is an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point.

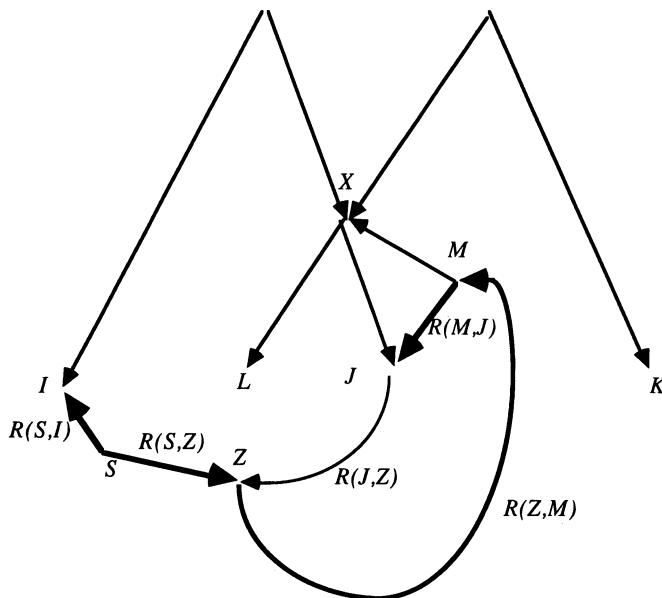


Figure 25

- b. Suppose $R(S,J)$ does contain X . See figure 26. Then there is a directed path $R(S,X)$ that is a subpath of $R(S,J)$. There is also a directed path $R(S,M) = R(S,Z) \& R(Z,M) \& R(M,J)$. Hence there is a trek between X and J whose source S' lies on $R(S,X)$, which is a subpath of $R(S,J)$. S' is in Q' since it is the source of a trek between X and J . S' lies on $R(S,X)$ which is a

subpath of $R(S,J)$. S' is not a collier on $R(S,J)$, since S' occurs before Z , which is the first collider on $R(I,U)$. Hence $R(S,I)$ contains a vertex in Q' that is not a collider.

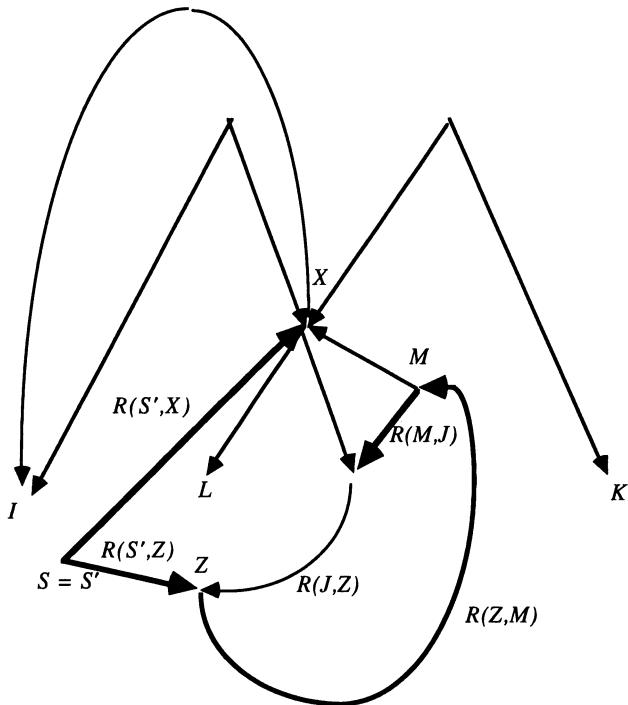


Figure 26

By lemma 6.10.20, $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$. ∴

Tetrad Representation Theorem 6.10: In an acyclic LCF G , there exists an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point iff G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

Proof. This follows directly from lemma 6.10.19 and lemma 6.10.21. ∴

Corollary 6.10.1: If an acyclic LCF G' is a subgraph of an acyclic LCF G , and G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, then G' linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$.

Proof. If G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, then by lemma 6.10.21 G has either an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point. If G has either an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point, then G' has either an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point. By lemma 6.10.21, G' linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$. \therefore

13.26 Theorem 6.11

Theorem 6.11: An acyclic LCF G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$ only if either ρ_{IJ} or $\rho_{KL} = 0$, and ρ_{IL} or $\rho_{JK} = 0$, or there exists a (possible empty) set \mathbf{Q} of random variables in G such that $\rho_{IJ.Q} = \rho_{KL.Q} = \rho_{IL.Q} = \rho_{JK.Q} = 0$.

Proof. By theorem 6.10, if G linearly implies $\rho_{IJ}\rho_{KL} - \rho_{IL}\rho_{JK} = 0$, then there is either an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point in G . In the proof of lemma 6.10.21 we demonstrated that the existence of an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point or an $IK(T(I,J),T(K,L),T(I,L),T(J,K))$ choke point then either ρ_{IJ} or $\rho_{KL} = 0$, and ρ_{IL} or $\rho_{JK} = 0$, or there exists a set \mathbf{Q} of random variables such that $\rho_{IJ.Q} = 0$, $\rho_{KL.Q} = 0$, $\rho_{IL.Q} = 0$, and $\rho_{JK.Q} = 0$. \therefore

13.27 Theorem 7.1

If G is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, \mathbf{W} is exogenous with respect to \mathbf{V} in G , \mathbf{Y} and \mathbf{Z} are disjoint subsets of \mathbf{V} , $P(\mathbf{V} \cup \mathbf{W})$ is a distribution that satisfies the Markov

condition for G , and $\text{Manipulated}(W) = X$, then $P(Y|Z)$ is invariant under direct manipulation of X in G by changing W from w_1 to w_2 if and only if $P(Y|Z, W = w_1) = P(Y|Z, W = w_2)$ wherever they are both defined. .

Theorem 7.1: If G_{Comb} is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G_{Comb} , Y and Z are disjoint subsets of V , $P(V \cup W)$ is a distribution that satisfies the Markov condition for G_{Comb} , no member of $X \cap Z$ is a member of $\text{IP}(Y, Z)$ in G_{Unman} , and no member of XZ is a member of $\text{IV}(Y, Z)$ in G_{Unman} , then $P(Y|Z)$ is invariant under a direct manipulation of X in G_{Comb} by changing W from w_1 to w_2 .

Proof. Suppose that G_{Comb} is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V , G_{Unman} is the subgraph of G_{Comb} over V , $P(V \cup W)$ is a distribution that satisfies the Markov condition for G_{Comb} , $X = \text{Manipulated}(W)$, $P(Y|Z, W = w_1) \neq P(Y|Z, W = w_2)$ when G_{Comb} is manipulated by changing the value of W from w_1 to w_2 , Y and Z are disjoint subsets of V , no member of $X \cap Z$ is a member of $\text{IP}(Y, Z)$ in G_{Unman} , and no member of XZ is a member of $\text{IV}(Y, Z)$ in G_{Unman} , but $P(Y|Z)$ is not invariant when X is manipulated. Hence there is an undirected path U in G_{Comb} that d-connects some R in W to some Y in Y given Z . Let W be the vertex on U closest to Y that is in W . By lemma 3.3.2, $U(W, Y)$ d-connects W and Y given $Z \setminus \{W, Y\} = Z$. Because $U(W, Y)$ contains no member of W except W , every subpath of $U(W, Y)$ that does not contain W is an undirected path in G_{Unman} . Because $U(W, Y)$ is an undirected path between W and Y , it contains some variable X in $\text{Manipulated}(W)$. There are two cases: either X is in Z or it is not in Z .

If X is in Z then X is a collider on U in G_{Unman} , and the vertex T adjacent to X on U and between X and Y is a parent of X , and hence not a collider on U . Because T is not a collider on U , T is not in Z , and $Z \setminus \{T\} = Z$. If T is in Y , then X is in $\text{IP}(Y, Z)$, contrary to our assumption. If T is not in Y , then $U(T, Y)$ d-connects T and Y given $Z \setminus \{T, Y\} = Z$ in G_{Unman} . T has a descendant (X) in Z in G_{Unman} , and hence T is in $\text{IV}(Y, Z)$ in G_{Unman} . But then X is in $\text{IP}(Y, Z)$ in G , contrary to our assumption.

If X is not in Z , then $U(X, Y)$ d-connects Y and X given $Z \setminus \{X\} = Z$ in G_{Unman} . If X is a collider on U then X has a descendant in Z in G_{Unman} . If X is not a collider on U then $U(X, Y)$ is out of X because X is a child of W . Either X is an ancestor of a collider on $U(X, Y)$, in which case it is an ancestor of some member of Z in G_{Comb} , or $U(X, Y)$ is a directed path to Y , in which case it is an ancestor of some member of Y in G_{Comb} . If X has a descendant in $Z \cup Y$ in G_{Comb} , then X has a descendant in $Z \cup Y$ in G_{Unman} , because W is exogenous with respect to

V . Hence X has a descendant in $Y \cup Z$ in G_{Unman} . It follows that X is in $\text{IV}(Y, Z)$ in G_{Unman} , contrary to our assumption. \therefore

13.28 Theorem 7.2

Theorem 7.2: If $P(O)$ is the marginal of a distribution faithful to G over V , π is a partially oriented inducing path graph of G over O , and Ord is an ordering of variables in O acceptable for some inducing path graph over O with partially oriented inducing path graph π , then there is a minimal I-map G_{Min} of $P(O)$ in which $\text{Definite-SP}(Ord, X)$ in π is included in $\text{Parents}(G_{Min}, X)$ which is included in $\text{Possible-SP}(Ord, X)$ in π .

Proof. Suppose that G_{IP} is an inducing path graph over O with partially oriented inducing path graph π . By lemma 6.2.4 if G_{IP} is an inducing path graph over O and Ord an acceptable total ordering of variables for G_{IP} , then $\text{Predecessors}(Ord, X) \setminus \text{SP}(Ord, G_{IP}, X)$ is d-separated from X given $\text{SP}(Ord, G_{IP}, X)$. Hence, if $\text{Parents}(G_{Min}, X) = \text{SP}(Ord, G_{IP}, X)$ then G_{Min} is an I-map of $P(O)$.

We will now show that no subgraph of G_{Min} is an I-map of $P(O)$. Suppose in G_{Sub} $\text{Parents}(G_{Sub}, X)$ is properly included in $\text{Parents}(G_{Min}, X)$ and hence properly included in $\text{SP}(Ord, G_{IP}, X)$. Let V be some variable in $\text{Parents}(G_{Min}, X) \setminus \text{Parents}(G_{Sub}, X)$. Because V is in $\text{SP}(Ord, G_{IP}, X)$ there is an undirected path U in G_{IP} between V and X on which all of the vertices except the endpoints are colliders, and precede X in Ord . Let W be the vertex on U closest to X but not equal to X that is in $\text{Parents}(G_{Min}, X) \setminus \text{Parents}(G_{Sub}, X)$. It follows that $U(W, X)$ is an undirected path in G_{IP} between W and X such that every vertex on $U(W, X)$ except for the endpoints is a collider and in $\text{Parents}(G_{Sub}, X)$. Hence W is in $\text{Predecessors}(Ord, X) \setminus \text{Parents}(G_{Sub}, X)$ and is d-connected to X given $\text{Parents}(G_{Sub}, X)$ in G_{IP} . Hence W is d-connected to X given $\text{Parents}(G_{Sub}, X)$ in G , and because $P(V)$ is faithful to G , W and X are dependent given $\text{Parents}(G_{Sub}, X)$. Hence $P(O)$ does not satisfy the Markov Condition for G_{Sub} .

For a partially oriented inducing path graph π and ordering Ord acceptable for π , V is in $\text{Possible-SP}(Ord, X)$ if and only if $V \neq X$ and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and no vertex on U except for the endpoints is a definite-non-collider on U . For a partially oriented inducing path graph π and ordering Ord acceptable for π , V is in $\text{Definite-SP}(Ord, X)$ if and only if $V \neq X$

and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and every vertex on U except for the endpoints is a collider on U . From these definitions and the definition of partially oriented inducing path graph it follows that **Definite-SP**(Ord, X) is included in **Parents**(G_{Min}, X) which is included in **Possible-SP**(Ord, X). \therefore

13.29 Theorem 7.3

Theorem 7.3: If G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , O is included in V , G_{Unman} is the subgraph of G over V , π is the FCI partially oriented inducing path graph over O of G_{Unman} , Y and Z are included in O , X is included in Z , Y and Z are disjoint, and no X in X is in **Possibly-IP**(Y, Z) in π , then $P(Y|Z)$ is invariant under direct manipulation of X in G by changing the value of W from w_1 to w_2 .

If A and B are not in Z , and $A \neq B$, then an undirected path U between A and B in a partially oriented inducing path graph π over O is a **possibly d-connecting** path of A and B given Z if and only if every collider on U is the source of a semi-directed path to a member of Z , and every definite non-collider is not in Z .

Lemma 7.3.1: If G is a directed acyclic graph, U is a path that d-connects V and Y given Z , X is in Z , and X is on U , then there is a path that d-connects X and Y given $Z \setminus \{X\}$ that is into X and that contains only edges that lie on a directed path to X , and a subpath of $U(X, Y)$.

Proof. Suppose that G is a directed acyclic graph, U is a path that d-connects V and Y given Z , X is in Z , and X is on U . Because X is in Z and on U , it follows that X is a collider on U , and hence $U(X, Y)$ is into X . No non-collider on $U(X, Y)$ except for the endpoints is in Z , so no non-collider on $U(X, Y)$ except for the endpoints is in $Z \setminus \{X\}$. Every collider on $U(X, Y)$ has a descendant in Z . If every collider on $U(X, Y)$ has a descendant in $Z \setminus \{X\}$ then $U(X, Y)$ d-connects X and Y given $Z \setminus \{X\}$. Suppose then that some collider on $U(X, Y)$ has X as a descendant but no other member of Z as a descendant, and let C be the closest such collider on U to Y . $U(C, Y)$ d-connects C and Y given $Z \setminus \{X\}$ because C is not in $Z \setminus \{X\}$, every collider on $U(C, Y)$ has a descendant in $Z \setminus \{X\}$, and no non-collider on $U(C, Y)$ is in $Z \setminus \{X\}$. There is a directed path from C to X that contains no member of $Z \setminus \{X\}$. Hence by lemma 3.3.3 X is d-connected to Y given $Z \setminus \{X\}$ by a path that is into X , and that contains only edges that lie on a directed path to X and a subpath of $U(X, Y)$.

Lemma 7.3.2: If G' is the inducing path graph for G over \mathbf{O} , X and Y are in \mathbf{O} , \mathbf{Z} is included in \mathbf{O} , and there is a path U d-connecting X and Y given \mathbf{Z} in G , then there is a path T d-connecting X and Y given \mathbf{Z} in G' such that if U is into X in G , then T is into X in G' and if U is into Y in G then T is into Y in G' .

Proof. Suppose that in G with inducing path graph G' that U is a path d-connecting X and Y given \mathbf{Z} . We will use the following algorithm to construct two sequences of vertices, *Ancestor*, and *D-Path*. (We are actually interested only in the undirected path *D-path*; *Ancestor* is used solely as a device to construct *D-path*.) The vertices in *D-Path* are always observed (i.e. vertices in \mathbf{O}), but might not be on U ; vertices in *Ancestor* are always on the path U , but might not be observed. For any sequence of vertices R of vertices, $R(n)$ refers to the n^{th} vertex in R . We will say that for any pair of variables V and W on U that W is after V on U if V is between W and X on U or $V = X$.

Algorithm D-Path

```

Ancestor(0) = < $X$ >.
D-path(0) = < $X$ >.
n = 0.
repeat
  if Ancestor(n) = D-path(n) then
    if there is no collider between Ancestor(n) and the next observed variable  $V$  on  $U$ ,
    Ancestor(n+1) = D-path(n+1) =  $V$ ;
    else Ancestor(n+1) = first collider on  $U$  after Ancestor(n) and D-path(n+1) = first
    observed variable on a path from Ancestor(n+1) to a member of  $\mathbf{Z}$ ;
    else if Ancestor(n) ≠ D-path(n) then
      if on  $U$  there is no collider  $C$  after Ancestor(n) that has D-path(n) as the first
      observed variable on a directed path from  $C$  to a member of  $\mathbf{Z}$ , then Ancestor(n+1)
      = D-path(n+1) = first observed variable on  $U$  after Ancestor(n)
      else
        let  $C_2$  be the collider closest to  $Y$  that has D-path(n) as the first observed variable
        on a directed path from  $C_2$  to a member of  $\mathbf{Z}$ ;
        if there is no collider between  $C_2$  and the first observed variable after  $C_2$  on  $U$ 
        then Ancestor(n+1) = D-path(n+1) = first observed variable after  $C_2$  on  $U$ ;
        else let  $C_1$  be the first collider after  $C_2$ , let Ancestor(n+1) =  $C_1$  and D-path(n+1)
        = the first observed variable on a directed path from  $C_1$  to a member of  $\mathbf{Z}$ ;
    else n = n + 1.
until  $Y$  is in D-path.

```

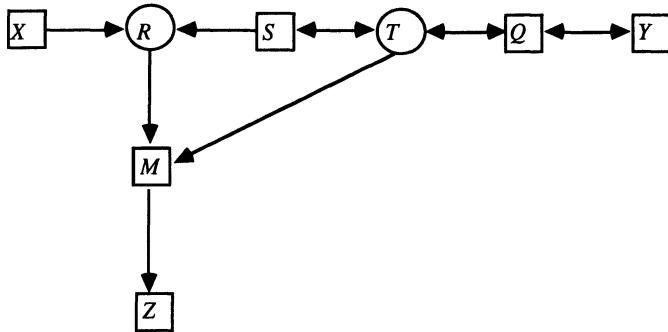


Figure 27

For example, when the algorithm is applied to the graph in figure 27 (where the circled vertices are not observed), for $U = \langle X, R, S, T, Q, Y \rangle$, and the result is $Ancestor = \langle X, R, Q, Y \rangle$ and $D\text{-path} = \langle X, M, Q, Y \rangle$.

We will now show that either $D\text{-path}$ d-connects X and Y given Z in G' , or some other path in G' d-connects X and Y given Z .

All of the vertices in $D\text{-path}$ are observed variables, and hence in G' . By the way that $D\text{-path}$ is constructed, each adjacent pair of vertices A and B in $D\text{-path}$ is connected in G by a trek $T(A,B)$ that contains no observed variables, except for the endpoints. If A and B are both on U then $T(A,B)$ contains the edges in $U(A,B)$; if A is on U and B is not then $T(A,B)$ contains the edges in $U(A,Ancestor(B))$ and a directed path from $Ancestor(B)$ to B ; if A is not on U and B is, then $T(A,B)$ consists of a directed path from $Ancestor(A)$ to A and $U(Ancestor(A),B)$; and if neither is on U , then $T(A,B)$ contains the edges in a directed path from $Ancestor(A)$ to A , $U(Ancestor(A),Ancestor(B))$, and a directed path from $Ancestor(B)$ to B . $T(A,B)$ is constructed out of subpaths of U , and subpaths of directed paths from colliders on U to vertices in Z . $T(A,B)$ is an inducing path in G , and hence each adjacent pair of vertices in $D\text{-path}$ is adjacent in G' . The method of construction of $D\text{-path}$ makes $D\text{-path}$ acyclic. It follows that $D\text{-path}$ is an acyclic undirected path from X to Y in G' .

If W is on $D\text{-path}$, but is not a collider on $D\text{-path}$, then W is on U in G , and is not a collider on U . It follows that W is not in Z .

We will now show that we can transform *D-path* into a path *D-path'* in G' such that every collider B on *D-path'* has a descendant in \mathbf{Z} in G . Let B be the vertex on *D-Path* closest to X that is a collider on *D-path* but that in G does not have a descendant in \mathbf{Z} , and A be the predecessor of B on *D-path*, and C be the successor of B on *D-path*. If in G $T(A,B)$ and $T(B,C)$ are both into B , then by the construction of *D-path*, B has a descendant in \mathbf{Z} in G . Hence at least one of $T(A,B)$ and $T(B,C)$ is out of B in G . Suppose without loss of generality that $T(B,C)$ is out of B in G , and B is between X and C on *D-path*. It follows that B is an ancestor of C in G . In addition since there is an arrowhead at B in G' , there is an inducing path between B and C that is into B and C . By lemma 6.6.2, there is a vertex V on *D-path*(X,C) such that there is an edge between V and C in G' that is substitutable for *D-path*(V,C). Let *D-path'* be the concatenation of *D-path*(X,V) with the edge between V and C . By lemma 6.6.2, *D-path'* is into X if *D-path* is. Every collider on *D-path'* is a collider on *D-path*, and every non-collider on *D-path'* is a non-collider on *D-path*. Furthermore, *D-path'* does not contain the vertex B which in G does not have a descendant in \mathbf{Z} . Repeat this process until every vertex on the modified *D-path* that in G does not have a descendant in \mathbf{Z} has been removed from the path. Call the result *D-path'*.

Suppose now that some collider B on *D-path'* has a descendant in \mathbf{Z} in G but not in G' . We will show how to transform *D-path'* into a path in G' in which every collider has a descendant in \mathbf{Z} in G' . Let P be a directed path in G from B to some Z that is a member of \mathbf{Z} . In G' , let P' be the undirected path from B to Z that consists of the observed variables on P in the order in which they occur. P' is an undirected path in G' because in G the directed path between any two observed variables on P is an inducing path. Let S be the vertex on P' closest to Z such that there is no directed path from B to S in G' . Let R be the predecessor of S on P' . If $P'(B,R)$ is not a directed path from B to R then form P'' by substituting some directed path from B to R in G' for $P'(B,R)$ in P' . There is an inducing path between R and S in G that is into S , so in G' the edge between R and S is into S . Because $P''(B,S)$ is not a directed path from B to S , but $P''(B,R)$ is a directed path from B to R , it follows that $R \leftrightarrow S$ in G' .

We will now demonstrate that there is an edge $B \leftrightarrow S$ in G' . If $B = R$, it follows from what we have just shown. Suppose then that $R \neq B$. In that case let Q be the predecessor of R on P'' . Because $P''(B,R)$ is a directed path from B to R , $Q \rightarrow R$ in G' . By lemma 6.6.2, there is a vertex E on $P''(B,R)$ such that there is an edge between E and S that is into S and is substitutable for $P''(E,S)$ in $P''(B,S)$. If the edge between E and S is out of E , then there is a directed path from B to S in G' , contrary to our assumption. It follows that the edge between E and S is into E . But because $P''(B,R)$ is a directed path from B to R , if the edge between E and

S is into E , the edge between E and S is not substitutable for $P''(E,S)$ in $P''(B,S)$ unless $E = B$. It follows then that $B \leftrightarrow S$ in G' .

We will now form a path $D\text{-path}''$ between X and Y by the following iteration, where at each stage of the iteration the vertices B and S are defined as above. Let the 0th stage $D\text{-path}''$ equal $D\text{-path}'$. If S is on the $n-1$ th stage $D\text{-path}''(X,B)$ let the n th stage $D\text{-path}''(X,S)$ equal the $n-1$ th stage $D\text{-path}''(X,S)$. If S is not on the $n-1$ th stage $D\text{-path}''(X,B)$ let V equal the concatenation of the $n-1$ th stage $D\text{-path}''(X,B)$ and $B \leftrightarrow S$. By lemma 6.6.2 there is a vertex E on V that is not equal to B and not equal to S such that there is an edge from E to S that is into S , and is a collider on V if and only if it is a collider on the concatenation of $V(X,E)$ with the edge between E and S . Let the n th stage $D\text{-path}''(X,S)$ equal the concatenation of $V(X,E)$ and the edge between E and S . Similarly, form the n th stage $D\text{-path}''(Y,S)$. The n th stage $D\text{-path}''(X,S)$ does not intersect the n th stage $D\text{-path}''(Y,S)$ except at S because except for the edges containing S , they are subpaths of paths that do not intersect except possibly at S . Let the n th stage $D\text{-path}''$ be the concatenation of $D\text{-path}''(X,S)$ and $D\text{-path}''(Y,S)$. If S does not have a descendant in Z in G' , repeat this process until some vertex M on P' that does have a descendant in Z in G' is on $D\text{-path}''$. See figure 28, where $D\text{-path}'$ is $\langle X, E, B, F, Y \rangle$ and $D\text{-path}''$ consists of the edges in boldface.

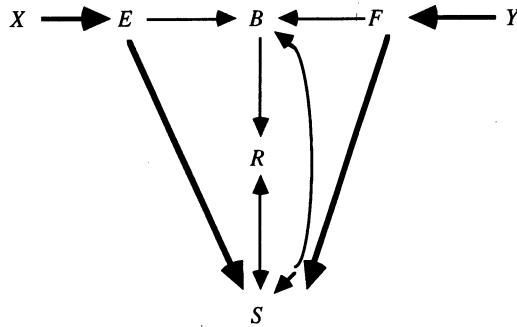


Figure 28

The n th stage $D\text{-path}''$ is into X if the $n-1$ th stage $D\text{-path}''$ is, and into Y if the $n-1$ th stage $D\text{-path}''$ is. Moreover, the 0th stage $D\text{-path}''$ ($D\text{-path}'$) is into X if U is, and into Y if U is. Every non-collider on the n th stage $D\text{-path}''$ is a non-collider on the $n-1$ th stage $D\text{-path}''$. Because every non-collider on $D\text{-path}'$ is not in Z , every non-collider on the n th stage $D\text{-path}''$ is not in Z . Every collider on the n th stage $D\text{-path}''$ with the possible exception of M is a collider on the $n-1$ th stage $D\text{-path}''$, and hence a collider on $D\text{-path}'$. M is a collider on the n th stage $D\text{-path}''$,

but it has a descendant in \mathbf{Z} . There is at least one fewer collider on n^{th} stage $D\text{-path}''$ that does not have a descendant in \mathbf{Z} than there is on $D\text{-path}'$ (because $D\text{-path}'$ contains B , and the n^{th} stage $D\text{-path}''$ does not.) This process can be repeated until every collider on $D\text{-path}''$ has a descendant in \mathbf{Z} . The resulting path d-connects X and Y given \mathbf{Z} in G' , is into X if U is, and into Y if U is. \therefore

Lemma 7.3.3: If G is a directed acyclic graph over \mathbf{V} , π is the FCI partially oriented inducing path graph of G over \mathbf{O} , and some path U in G d-connects X and Y given \mathbf{Z} , then there is a path U'' in π that possibly d-connects X and Y given \mathbf{Z} . Furthermore if U is into X , then U'' is not out of X .

Proof. Suppose that some path U in G d-connects X and Y given \mathbf{Z} . Let G' be the inducing path graph of G . By lemma 7.3.2, there is a path U' in G' that d-connects X and Y given \mathbf{Z} , and if U is into X then U' is into X . Let U'' be the path in π that corresponds to U' in G' . If R is a collider on U'' , then by the definition of partially oriented inducing path graph R is a collider on U' . Because R is a collider on U' , and U' d-connects X and Y given \mathbf{Z} , R has a descendant in \mathbf{Z} in G' . By Theorem 6.6, there is a semi-directed path from R to a member of \mathbf{Z} in π . If R is a definite non-collider on U'' , then by definition of partially oriented inducing path graph R is a non-collider on U' . Because R is a non-collider on U' , and U' d-connects X and Y given \mathbf{Z} , R is not in \mathbf{Z} . Hence U'' is a possibly d-connecting path between X and Y given \mathbf{Z} . Furthermore, if U' is into X , then by definition of partially oriented inducing path graph U'' is not out of X . \therefore

If π is a partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IV(Y,Z)** if and only if X is not in \mathbf{Z} , there is a possibly d-connecting path between X and some Y in \mathbf{Y} given \mathbf{Z} , and there is a semi-directed path from X to a member of $\mathbf{Y} \cup \mathbf{Z}$. If π is a partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IP(Y,Z)** if and only if \mathbf{Y} and \mathbf{Z} are disjoint, X is in \mathbf{Z} , and there is a possibly d-connecting path between X and some Y in \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ that is not out of X . If π is the FCI partially oriented inducing path graph of G over \mathbf{O} , then X is in **Definite-Non-Descendants(Y)** if and only if there is no semi-directed path from any member of \mathbf{Y} to X in π .

Lemma 7.3.4: If X is in **IV(Y,Z)** in directed acyclic graph G , \mathbf{Y} and \mathbf{Z} are disjoint subsets of \mathbf{O} , X is in \mathbf{O} , and π is the FCI partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IV(Y,Z)** in π .

Proof. Suppose that X is in **IV(Y,Z)** in G , \mathbf{Y} and \mathbf{Z} are disjoint subsets of \mathbf{O} , X is in \mathbf{O} , and π is the FCI partially oriented inducing path graph of G over \mathbf{O} . Because X is in **IV(Y,Z)** in G , X has a descendant in $\mathbf{Y} \cup \mathbf{Z}$ in G . Hence, by Theorem 6.6, there is a semi-directed path from

X to a member of $\mathbf{Y} \cup \mathbf{Z}$ in π . Also, there is a path that d-connects X and some member Y of \mathbf{Y} given \mathbf{Z} in G . Hence, by lemma 7.3.3 there is a path that possibly d-connects X and some member Y of \mathbf{Y} given \mathbf{Z} in π . By definition X is in **Possibly-IV(Y,Z)** in π . \therefore

Lemma 7.3.5: If X is in **IP(Y,Z)** in directed acyclic graph G , \mathbf{Y} and \mathbf{Z} are disjoint subsets of \mathbf{O} , and π is the FCI partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IP(Y,Z)** in π .

Proof. Suppose that X is in **IP(Y,Z)** in G , \mathbf{Y} and \mathbf{Z} are disjoint subsets of \mathbf{O} , and π is the FCI partially oriented inducing path graph of G over \mathbf{O} . Because X is in **IP(Y,Z)** in G , some variable T in G is a parent of X and in **IV(Y,Z)** or \mathbf{Y} . If T is in \mathbf{Y} then there is a directed path from a member T of \mathbf{Y} to X that d-connects T and X given $\mathbf{Z} \setminus \{X\}$. If T is in **IV(Y,Z)** then T is d-connected to some Y in \mathbf{Y} given \mathbf{Z} by some path U . If X is on U then X is a collider on U and $U(X,Y)$ is into X ; furthermore, by lemma 7.3.1 there is an undirected path that d-connects X and \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ that is into X . If X is not on U then the concatenation of the edge from T to X and U is a path that d-connects X and \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ and is into X . Hence, by lemma 7.3.3 there is a path that possibly d-connects X and \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ in π that is not out of X . By definition X is in **Possibly-IP(Y,Z)** in π . \therefore

Theorem 7.3: If G is a directed acyclic graph over $\mathbf{V} \cup \mathbf{W}$, \mathbf{W} is exogenous with respect to \mathbf{V} in G , \mathbf{O} is included in \mathbf{V} , G_{Unman} is the subgraph of G over \mathbf{V} , π is the FCI partially oriented inducing path graph over \mathbf{O} of G_{Unman} , \mathbf{Y} and \mathbf{Z} are included in \mathbf{O} , X is included in \mathbf{Z} , \mathbf{Y} and \mathbf{Z} are disjoint, and no X in \mathbf{X} is in **Possibly-IP(Y,Z)** in π , then $P(\mathbf{Y}|\mathbf{Z})$ is invariant under direct manipulation of \mathbf{X} in G by changing the value of \mathbf{W} from w_1 to w_2 .

Proof. Suppose that G is a directed acyclic graph over $\mathbf{V} \cup \mathbf{W}$, \mathbf{O} is included in \mathbf{V} , \mathbf{W} is exogenous with respect to \mathbf{V} in G , G_{Unman} is the subgraph of G over \mathbf{V} , π is the FCI partially oriented inducing path over \mathbf{O} of G_{Unman} , \mathbf{Y} and \mathbf{Z} are included in \mathbf{O} , X is included in \mathbf{Z} , \mathbf{Y} and \mathbf{Z} are disjoint, and no X in \mathbf{X} is in **Possibly-IP(Y,Z)** in π . If $P(\mathbf{Y}|\mathbf{Z})$ is not invariant when \mathbf{X} is manipulated by changing the value of \mathbf{W} from w_1 to w_2 then \mathbf{W} is d-connected to \mathbf{Y} given \mathbf{Z} in G . Suppose that \mathbf{W} is d-connected to \mathbf{Y} given \mathbf{Z} in G . Let W be a member of \mathbf{W} that is d-connected to some Y in \mathbf{Y} by an undirected path U in G that contains no other member of \mathbf{W} . No non-collider on U is in \mathbf{Z} , and every collider on U has a descendant in \mathbf{Z} .

Note that if R and N are in \mathbf{V} and R is a descendant of N in G , then R is a descendant of N in G_{Unman} , because there is no edge from any member of \mathbf{V} into a member of \mathbf{W} . In G , U contains some X in \mathbf{X} . Because X is in \mathbf{Z} , X is a collider on U , and $U(X,Y)$ is into X . By lemma 7.3.1 in G there is an undirected path M that d-connects X and Y given $\mathbf{Z} \setminus \{X\}$, is into X , and contains only edges that lie on a directed path to X and a subpath of $U(X,Y)$. Hence M

is an undirected path in G_{Unman} , no non-collider on M is in $Z \setminus \{X\}$, and every collider on M has a descendant in $Z \setminus \{X\}$ in G , and hence in G_{Unman} . It follows that M d-connects X and Y given $Z \setminus \{X\}$ in G_{Unman} . Let T be the vertex adjacent to X on M . If $T = Y$ then X is in $\text{IP}(Y, Z)$ in G_{Unman} . If $T \neq Y$ then T has a descendant in Z (namely X) in G_{Unman} . Also T is not a collider on $U(X, Y)$, and hence not in Z . By lemma 3.3.2 T is d-connected to Y given $Z \setminus \{T\} = Z$ in G_{Unman} . It follows that T is in $\text{IV}(Y, Z)$ in G_{Unman} , and hence X is in $\text{IP}(Y, Z)$ in G_{Unman} . In either case X is in $\text{IP}(Y, Z)$ in G_{Unman} and by lemma 7.3.5, X is in **Possibly-IP**(Y, Z) in π , contrary to our assumption. \therefore

13.30 Theorem 7.4

Theorem 7.4: If G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , O is included in V , G_{Unman} is the subgraph of G over V , π is the FCI partially oriented inducing path graph over O of G_{Unman} , X , Y and Z are included in O , X , Y and Z are pairwise disjoint, and no X in X is in **Possibly-IV**(Y, Z) in π , then $P(Y|Z)$ is invariant under direct manipulation of X in G by changing the value of W from w_1 to w_2 .

Proof. Suppose G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , O is included in V , G_{Unman} is the subgraph of G over V , π is the FCI partially oriented inducing path over O of G_{Unman} , Y and Z are included in O , X , Y and Z are pairwise disjoint, and no X in X is in **Possibly-IV**(Y, Z). If $P(Y|Z)$ is not invariant when X is manipulated by changing the value of W from w_1 to w_2 then W is d-connected to Y given Z in G . Let W be a member of W that is d-connected to some Y in Y given Z by an undirected path U in G that contains no other member of W .

Because U d-connects W and Y given Z , no non-collider on U is in Z , and every collider on U has a descendant in Z . U contains some X in X . By lemma 3.3.2 $U(X, Y)$ is an undirected path that d-connects X and Y given Z in G . There is a path $U'(X, Y)$ in G_{Unman} with the same edges as $U(X, Y)$ in G , because $U(X, Y)$ contains no member of W . No non-collider on $U'(X, Y)$ is in Z . In G , every collider on $U(X, Y)$ has a descendant in Z ; hence every collider on $U'(X, Y)$ has a descendant in Z in G_{Unman} . Hence $U(X, Y)$ d-connects X and Y given Z in G_{Unman} . By lemma 7.3.3 there is a possibly d-connecting path between X and some Y in Y given Z in π .

Now we will show that X has a descendant in $Y \cup Z$ in G_{Unman} . If X is a collider on U , then X has a descendant in Z in G , and hence in G_{Unman} . Suppose then that X is not a collider on

U . The edge from W to X on U is into X , so the edge containing X on $U(X,Y)$ is out of X . If $U(X,Y)$ contains no colliders then Y is a descendant of X . If $U(X,Y)$ contains a collider, then the collider on $U(X,Y)$ closest to X is a descendant of X , and an ancestor of a member of Z . Hence X is an ancestor of a member of Z . In either case, X has a descendant in $Y \cup Z$ in G , and hence in G_{Unman} .

It follows that X is in $\mathbf{IV}(Y, Z)$ in G_{Unman} , and hence by lemma 7.3.4 X is in $\mathbf{Possibly-IV}(Y, Z)$, contrary to our assumption. \therefore

13.31 Theorem 7.5

Theorem 7.5: If G is a directed acyclic graph over $V \cup W$, W is exogenous with respect to V in G , G_{Unman} is the subgraph of G over V , $P_{Unman(W)}(V) = P(V|W = w_1)$ is faithful to G_{Unman} , and changing the value of W from w_1 to w_2 is a direct manipulation of X in G , then the Prediction Algorithm is correct.

Proof. Let G_{Man} be the manipulated graph, and F the minimal I-map of $P_{Unman(W)}(V)$ constructed by the algorithm for the given ordering of variables Ord . Step A) is trivial. Step B) is correct by Theorem 6.4. Step C1) is correct by Theorem 7.2. In step C2, by lemma 3.3.5, for all values of V for which the conditional distributions in the factorization are defined,

$$P_{Unman(W)}(Y|Z) = \frac{\sum_{\substack{\rightarrow \\ \mathbf{IV}(Y, Z) V \in \mathbf{IV}(Y, Z) \cup \mathbf{IP}(Y, Z) \cup Y}} \prod_{V \in \mathbf{IV}(Y, Z) \cup \mathbf{IP}(Y, Z) \cup Y} P_{Unman(W)}(V|\mathbf{Parents}(F, V))}{\sum_{\substack{\rightarrow \\ \mathbf{IV}(Y, Z) \cup Y V \in \mathbf{IV}(Y, Z) \cup \mathbf{IP}(Y, Z) \cup Y}} \prod_{V \in \mathbf{IV}(Y, Z) \cup \mathbf{IP}(Y, Z) \cup Y} P_{Unman(W)}(V|\mathbf{Parents}(F, V))}$$

for all values z of Z such that $P_{Man}(z) \neq 0$.

Because G_{Man} is a subgraph of G_{Unman} , if F is an I-map of $P_{Unman(W)}(V)$ then F is an I-map of $P_{Man(W)}(V)$. Hence $P_{Man(W)}(V)$ satisfies the Markov condition for F , and by lemma 3.3.5

$$(1) \quad P_{Man(W)}(Y|Z) = \frac{\sum_{\substack{IV(Y,Z) \\ V \in IV(Y,Z) \cup IP(Y,Z) \cup Y}}^{\rightarrow} P_{Man(W)}(V|\text{Parents}(F,V))}{\sum_{\substack{IV(Y,Z) \cup Y \\ V \in IV(Y,Z) \cup IP(Y,Z) \cup Y}}^{\rightarrow} P_{Man(W)}(V|\text{Parents}(F,V))}$$

for all values z of Z such that $P_{Man}(z) \neq 0$, and for all values for which the conditional distributions in the factorization exist.

$P_{Man(W)}(V)$ satisfies the Markov condition for G_{Man} by hypothesis. Hence in $P_{Man(W)}(V)$ X is independent of its non-parental non-descendants in G_{Man} given $\text{Parents}(G_{Man},X)$. The predecessors of X in Ord by hypothesis are either in **Definite-Non-Descendants**(π,X), in which case they are in **Non-Descendants**(G_{Unman},X) or they are in **Parents**(G_{Man},X). G_{Man} is a subgraph of G_{Unman} , so any vertex that is a non-descendant of X in G_{Unman} is a non-descendant of X in G_{Man} . Hence each predecessor of X in Ord is a non-descendant of X in G_{Man} . The algorithm guarantees that **Parents**(G_{Man},X) is included in **Predecessors**(Ord,X). It follows that **Parents**(G_{Man},X) is a subset of **Predecessors**(Ord,X) such that **Predecessors**($Ord,X \setminus \text{Parents}(G_{Man},X)$) is independent of X given **Parents**(G_{Man},X) in $P_{Man(W)}(V)$. Hence, if **Parents**(G_{Man},X) is substituted for **Parents**(F,X) in F , the resulting graph is still an I-map of $P_{Man(W)}(V)$, by lemma 3.7.1. So in (1) we can substitute $P(X|\text{Parents}(G_{Man},X))$ for $P(X|\text{Parents}(F,X))$. By assumption the algorithm returns a value only if $P_{Man(W)}(V|\text{Parents}(F,V)) = P_{Unman(W)}(V|\text{Parents}(F,V))$ for each $V \neq X$, so we can substitute $P_{Unman(W)}(V|\text{Parents}(F,V))$ for $P_{Man(W)}(V|\text{Parents}(F,V))$ in (1). ∴

13.32 Theorem 9.1

Theorem 9.1 If $P(S)$ is faithful to $G(S)$, and X and Y are sets of variables in $G(S)$ not containing S , then $P(Y|X) = P(Y|X,S)$ if and only if X d-separates Y and S in $G(S)$.

Proof. This follows from Theorem 3.3. ∴

13.33 Theorem 9.2

Theorem 9.2: For a joint distribution, P , faithful to graph G , exactly one of $\langle Y \perp\!\!\!\perp X|Z; Y \perp\!\!\!\perp X|Z \cup \{S\} \rangle$ is true in P if and only if the corresponding member and only that member of $\langle Z \text{ d-separates } X; Y; Z \cup \{S\} \text{ d-separates } X, Y \rangle$ is true in G .

Proof. This follows from Theorem 3.3. \therefore

13.34 Theorem 10.1

Theorem 10.1: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, and each latent variable in T has at least two measured indicators, then latent variables T_1 and T_3 , whose measured indicators include J and L respectively, are d-separated given latent variable T_2 , whose measured indicators include I and K , if and only if G linearly implies $\rho_{J|T} = \rho_{JL\rho_{LK}} = \rho_{JK\rho_{IL}}$.

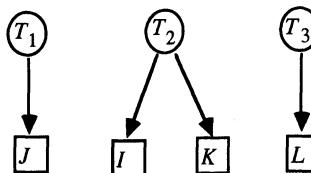


Figure 29

We say that a measurement model is **almost pure** if the only kind of impurities among the measured variables are common cause impurities. An **almost pure latent variable graph** is one in which the measurement model is almost pure.

Lemma 10.1.1: If G' is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, and each latent variable in T has at least two measured indicators, and latent variables T_1 and T_3 , whose measured indicators include J and L respectively, are d-separated given latent variable T_2 , whose measured indicators include I and K , then G' linearly implies $\rho_{J|T} = \rho_{JL\rho_{LK}} = \rho_{JK\rho_{IL}}$.

Proof. Let G be a pure latent variable subgraph of G' , formed by removing the sources of all treks creating common cause impurities. If T_1 and T_3 are d-separated given T_2 in G' then they are d-separated given T_2 in G . Because I and K are pure indicators of T_2 in G , and thus children

only of T_2 , T_2 is a non-collider on all undirected paths between I and any other indicator or K and any other indicator. Therefore J and I are d-separated given T_2 , K and L are d-separated given T_2 , and K and I are d-separated given T_2 .

Since T_1 and T_3 are d-separated given T_2 , and again J and L are children only of T_1 and T_3 respectively, then J and L are d-separated given T_2 . X and Z are d-separated given Y if and only if G linearly implies $\rho_{XZ|Y} = 0$. Hence G linearly implies $\rho_{IJ|T_2} = 0$, and $\rho_{IJ} = \rho_{IT_2} \times \rho_{JT_2}$. Similarly, G linearly implies $\rho_{KL} = \rho_{KT_2} \times \rho_{LT_2}$, $\rho_{JL} = \rho_{JT_2} \times \rho_{LT_2}$ and $\rho_{IK} = \rho_{IT_2} \times \rho_{KT_2}$. Hence G linearly implies $\rho_{JI}\rho_{LK} = \rho_{JT_2} \times \rho_{IT_2} \times \rho_{LT_2} \times \rho_{KT_2} = \rho_{JT_2} \times \rho_{LT_2} \times \rho_{KT_2} \times \rho_{IT_2} = \rho_{JL}\rho_{KI}$. G linearly implies the same vanishing tetrad differences as G' , so G' linearly implies $\rho_{JI}\rho_{LK} = \rho_{JL}\rho_{KI}$. The proof that $\rho_{JL}\rho_{KI} = \rho_{JK}\rho_{IL}$ is linearly implied by G' is essentially the same. ∴

Lemma 10.1.2: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, and each latent variable in T has at least two measured indicators, then latent variables T_1 and T_3 , whose measure indicators respectively include J and L , are d-separated given latent variable T_2 , whose measured indicators include I and K , if G linearly implies $\rho_{JI}\rho_{LK} = \rho_{JL}\rho_{KI}$.

Proof. Suppose that G linearly implies $\rho_{JI}\rho_{LK} = \rho_{JL}\rho_{KI}$ but T_1 and T_3 are not d-separated given T_2 .

By the Tetrad Representation Theorem, if G linearly implies $\rho_{JI}\rho_{LK} = \rho_{JL}\rho_{KI}$ then either there is an $IL(T(I,J),T(L,K),T(L,J),T(I,K))$ choke point, or there is a $JK(T(I,J),T(L,K),T(L,J),T(I,K))$ choke point.

Let $T(I,K)$ be the trek consisting of the edges from T_2 to I and T_2 to K . Suppose first that there is an $IL(T(I,J),T(L,K),T(L,J),T(I,K))$ choke point. The choke point is either I or T_2 because those are the only vertices in $I(T(I,K))$. I is not the choke point because it does not lie on any trek between L and K . Hence T_2 is the choke point. Similarly, if there is a $JK(T(I,J),T(L,K),T(L,J),T(I,K))$ choke point it is T_2 . Hence, in either case T_2 is a choke point.

There are two ways that T_1 and T_3 might fail to be d-separated given T_2 . Either there is a trek between T_1 and T_3 that does not contain T_2 , or there is some undirected path U between T_1 and T_3 such that T_2 is a descendent of every collider on U , and T_2 is not a non-collider on U .

First assume that there is some trek between T_1 and T_3 that does not contain T_2 . Then there is a trek between J and L that does not contain T_2 . But then T_2 is not a choke point, contrary to what we have just proved.

Now assume that there is some undirected path U between T_1 and T_3 such that T_2 is a descendent of every collider on U , and T_2 is not a non-collider on U . In that case U d-connects T_1 and T_3 given T_2 . Again there are two cases.

Suppose first that T_2 is an $IL(T(I,J), T(L,K), T(L,J), T(I,K))$ choke point. Let C be the collider on the undirected path U that is closest to T_3 . (See figure 30.)

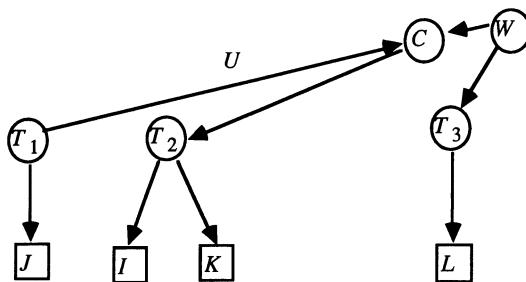


Figure 30

$U(T_3, C)$ does not contain any colliders on U except C because C is the closest collider to T_3 on U ; hence $U(T_3, C)$ is a trek between T_3 and C . There is a vertex W on $U(T_3, C)$ that is the source of a trek between T_3 and C . $W \neq C$ because W is not a collider on U , but C is. Hence $U(W, T_3)$ contains no colliders on U . It follows that $U(W, T_3)$ does not contain T_2 , because T_2 is not a non-collider on U . Hence there is a trek $T(K, L)$ between K and L whose K branch consists of the concatenation of $U(W, C)$, a directed path from C to T_2 , and the edge from T_2 to K , and whose L branch consists of the concatenation of $U(W, T_3)$ and the edge from T_3 to L . Because neither $U(W, T_3)$ nor the edge from T_3 to L contains T_2 , T_2 is not in $L(T(K, L))$, and hence is not an $IL(T(I,J), T(L,K), T(L,J), T(I,K))$ choke point, contrary to our hypothesis.

A similar argument shows that if there is some undirected path U between T_1 and T_3 such that T_2 is a descendent of every collider on U and T_2 is not a non-collider on U , then there is no $JK(T(I,J), T(L,K), T(L,J), T(I,K))$ choke point.

Therefore T_1 and T_3 are d-separated given T_2 . \therefore

Theorem 10.1: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each latent variable in T has at least two measured indicators, then latent variables T_1 and T_3 , whose measured indicators include J and L respectively, are d-separated given latent variable T_2 , whose measured indicators include I and K , if and only if G linearly implies $\rho_{JI}\rho_{LK} = \rho_{JL}\rho_{KI} = \rho_{JK}\rho_{IL}$.

Proof. The theorem follows from lemmas 10.1.1 and 10.1.2.

13.35 Theorem 10.2

Theorem 10.2: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , and Π is the output of MIMBuild then

- A-1) If X and Y are not adjacent in Π , then they are not adjacent in G .
- A-2) If X and Y are adjacent in Π and the edge is not labelled with a "?", then X and Y are adjacent in G .
- O-1) If $X \rightarrow Y$ is in Π , then every trek in G between X and Y is into Y .
- O-2) If $X \rightarrow Y$ is in Π and the edge between X and Y is not labelled with a "?", then $X \rightarrow Y$ is in G .

Lemma 10.2.1: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , Π is the output of MIMBuild, and X and Y are not adjacent in Π , then they are not adjacent in G .

Proof. This follows directly from Theorem 3.4.

Lemma 10.2.2: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , Π is the output of MIMBuild, and $X \rightarrow Y$ is in Π , then every trek in G between X and Y is into Y .

Proof. Suppose $X \rightarrow Y$ is in Π . The proof is by induction on the number of iterations of the repeat loop in step D) in the PC Algorithm.

Base Case: There is a trek between X and Y in G , because otherwise X and Y are d-separated given the empty set and therefore not adjacent in Π . Suppose that $X \rightarrow Y$ is oriented as $X \rightarrow Y \leftarrow Z$ by step C) of the PC Algorithm (i.e. X and Z are d-separated by some set not containing Y .) If in G , there is a trek between X and Y , and a trek between Y and Z that are not both into Y , then there is a trek between X and Z and hence X and Z are not d-separated given the empty set. Suppose then that X and Z are d-separated by some $W \neq Y$ in G . Because X and Y are adjacent in Π , W does not d-separate X and Y in G . Similarly, W does not d-separate Y and Z . If there is a trek in G between X and Y that is out of Y then there is a directed path U from Y to X in G . If U does not contain W then U d-connects X and Y given W in G . There is also a path V in G that d-connects Y and Z given W . Because U is out of Y , U and V do not collide at Y in G . Hence by lemma 3.3.1 X and Z are d-connected given W in G , contrary to our assumption. If U does contain W , then W is a descendant of Y , and by lemma 3.3.1 X and Z are d-connected given W , contrary to our assumption. Hence no trek in G between X and Y is out of Y .

Induction Case: Suppose after $n-1$ iterations of the repeat loop in step D) of the PC Algorithm, if $Z \rightarrow X$ in Π , then every trek between Z and X in G is into X . Suppose that the $X \rightarrow Y$ edge is oriented because there is some vertex Z such that $Z \rightarrow X - Y$ in Π and Z is not adjacent to Y in Π . Because the edge between X and Y in Π was not oriented into Y , X and Z are d-separated given Y . There are treks between X and Y , and between Y and Z in G , because they are adjacent in Π . If there is a trek between Y and X that is into X , then by lemma 3.3.1, X and Z are d-connected given Y , contrary to our assumption. \therefore

Y is a **definite non-collider** on an undirected path U in pattern Π if and only if either $X *-* Y \rightarrow Z$, or $X \leftarrow Y *-* Z$ are subpaths of U , or X and Z are not adjacent and not $X \rightarrow Y \leftarrow Z$ on U .

Lemma 10.2.3: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , Π is the output of MIMBuild, and Y is a definite non-collider on undirected path U in P , and the corresponding path U' exists in G , then Y is a non-collider on U' .

Proof. If U contains $X *-* Y \rightarrow Z$ in Π , then by lemma 10.2.2, if the corresponding path U' exists in G , then the edge between Y and Z in G is out of Y ; hence Y is not a collider on U' .

Similarly, if $X <- Y *-* Z$ in Π , then Y is not a collider on U . Suppose then that X and Z are not adjacent and not $X \rightarrow Y <- Z$ on U in Π . It follows that X and Z are d-separated given Y in G . Hence if the edges between X and Y and between Y and Z exist in G , they do not collide at Y .

Lemma 10.2.4: If G is an almost pure latent variable graph over $V \cup T \cup C$, T is causally sufficient, each variable in T has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , Π is the output of MIMBuild, and $X - Y$ or $X \rightarrow Y$ is in Π , and the edge is not labelled by a "?", then X and Y are adjacent in G .

Proof. Suppose that $X - Y$ or $X \rightarrow Y$ is in Π , the edge is not labelled by a "?", but that X and Y are not adjacent in G . Then there is some set S that d-separates X and Y in G . Let P be the set of undirected paths in P between X and Y of length ≥ 2 . Any such S has cardinality ≥ 2 , because otherwise MIMBuild would have found it with some test of vanishing zero or first order partial correlations. $X - Y$ or $X \rightarrow Y$ was not labelled with a "?" so either (i) P is empty, or (ii) every path in P contains a collider, or (iii) there is some vertex Z that is a definite non-collider on every path in P , or (iv) every path in P contains some subpath $\langle A, B, C \rangle$.

Suppose P is empty. Because by lemma 10.2.1 non-adjacencies in Π are non-adjacencies in G , the adjacencies in Π are a superset of those in G , and thus the set of undirected paths in Π is a superset of the undirected paths in G . It follows that there is no undirected path of length ≥ 2 in G . If in G there is also no edge between X and Y , then X and Y are d-separated given the empty set in G . But since there is an edge between X and Y in Π , X and Y are not d-separated given the empty set in G . Hence there is an edge between X and Y in G .

Suppose every path in P contains a collider and there is no edge between X and Y in G . By lemmas 10.2.1 and 10.2.2 every path in G between X and Y contains a collider. Hence there is no trek between X and Y in G . But then there is no edge between X and Y in Π , contrary to our assumption.

Suppose there is some vertex Z that is a definite non-collider on every path in P . It follows from lemma 10.2.1, 10.2.2, and 10.2.3 that if there is no edge between X and Y in G , then Z is a non-collider on every undirected path between X and Y in G . Hence X and Y are d-separated by Z . It follows that there is no edge between X and Y in Π , contrary to our assumption.

Suppose every path in \mathbf{P} contains some subpath $\langle A, B, C \rangle$. If there is no edge between X and Y in G , then every undirected path in G between X and Y contains $\langle A, B, C \rangle$. It follows that B is either a collider on every path between X and Y in G , in which case X and Y are d-separated given the empty set, or B is a non-collider on every path between X and Y in G , in which case, X and Y are d-separated given B in G . In either case, there is no edge between X and Y in Π , contrary to our assumption. \therefore

Lemma 10.2.5: If G is an almost pure latent variable graph over $\mathbf{V} \cup \mathbf{T} \cup \mathbf{C}$, \mathbf{T} is causally sufficient, each variable in \mathbf{T} has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , Π is the output of MIMBuild, and $X \rightarrow Y$ is in Π , and the edge is not labelled by a "?", then $X \rightarrow Y$ is in G .

Proof. This follows from lemmas 10.2.2 and 10.2.4. \therefore

Theorem 10.2: If G is an almost pure latent variable graph over $\mathbf{V} \cup \mathbf{T} \cup \mathbf{C}$, \mathbf{T} is causally sufficient, each variable in \mathbf{T} has at least two measured indicators, the input to MIMBuild is a list of all vanishing zero and first order correlations among the latent variables linearly implied by G , and Π is the output of MIMBuild then

- A-1) If X and Y are not adjacent in Π , then they are not adjacent in G .
- A-2) If X and Y are adjacent in Π and the edge is not labelled with a "?", then X and Y are adjacent in G .
- O-1) If $X \rightarrow Y$ is in Π , then every trek in G between X and Y is into Y .
- O-2) If $X \rightarrow Y$ is in Π and the edge between X and Y is not labelled with a "?", then $X \rightarrow Y$ is in G .

Proof. This follows from lemmas 10.2.1 through 10.2.5. \therefore

13.36 Theorem 11.1

Theorem 11.1. If G is a subgraph of directed acyclic graph G' , than the set of tetrad equations among variables of G that are linearly implied by G' is a subset of those linearly implied by G .

Proof. If G is a subgraph of directed acyclic graph G' , then the treks in G are a subset of the treks in G' . Hence if there is a choke point in G' , there is a choke point in G . By the Tetrad

Representation Theorem, if G' linearly implies that a tetrad difference t vanishes, then G linearly implies t vanishes. \therefore

Glossary

A: In a graph G , Let $\mathbf{A}(A,B)$ be the union of the ancestors of A or B .

Acceptable: Let a total order Ord of variables in a graph G' be **acceptable** for G if and only if whenever $A \neq B$ and there is a directed path from A to B in G' , A precedes B in Ord .

After: In a graph G , vertex X is **after** vertex Y if and only if there is a directed path from Y to X in G .

Almost Pure: We say that a measurement model is **almost pure** if the only kind of impurities among the measured variables are common cause impurities. An **almost pure latent variable graph** is one in which the measurement model is almost pure.

Before: In a graph G , vertex X is **before** vertex Y if and only if there is a directed path from X to Y in G .

C.F: See constant factor.

Choke point: In a directed acyclic graph G , if for all $T(K,L)$ in $\mathbf{T}(K,L)$ and all $T(I,J)$ in $\mathbf{T}(I,J)$, $L(T(K,L))$ and $J(T(I,J))$ intersect at a vertex Q , then Q is an $LJ(T(I,J),T(K,L))$ **choke point**. Similarly, if for all $T(K,L)$ in $\mathbf{T}(K,L)$ and all $T(I,J)$ in $\mathbf{T}(I,J)$, $L(T(K,L))$ and all $J(T(I,J))$ intersect at a vertex Q , and for all $T(I,L)$ in $\mathbf{T}(I,L)$ and all $T(J,K)$ in $\mathbf{T}(J,K)$, $L(T(I,L))$ and $J(T(J,K))$ also intersect at Q , then Q is an $LJ(T(I,J),T(K,L),T(I,L),T(J,K))$ **choke point**. Also see the definition of trek.

Combined graph: See manipulation.

Constant factor: In an LCF or LCT T , if an expression is equal to ce , where c is a non-zero constant, and e is a product of equation coefficients raised to positive integral powers, then c is the **constant factor** of (c.f.) ce .

Contains: In a directed acyclic graph, directed paths $R(U,I)$ and $R(U,J)$ **contain trek** T iff $I(T(I,J))$ is a final segment of $R(U,I)$ and $J(T(I,J))$ is a final segment of $R(U,J)$.

D: Given a directed acyclic graph G , $\mathbf{D}(X_i, X_j)$ is the set of all directed paths from X_i to X_j .

D-connection: See D-separation.

Definite discriminating path: In a partially oriented inducing path graph π , U is a **definite discriminating path** for B if and only if U is an undirected path between X and Y containing B , $B \neq X$, $B \neq Y$, every vertex on U except for B and the endpoints is a collider or a definite non-collider on U , and

- (i) if V and V' are adjacent on U , and V' is between V and B on U , then $V \rightarrow V'$ on U ,
- (ii) if V is between X and B on U and V is a collider on U then $V \rightarrow Y$ in π , else $V \leftarrow Y$ in π ,
- (iii) if V is between Y and B on U and V is a collider on U then $V \rightarrow X$ in π , else $V \leftarrow X$ in π ,
- (iv) X and Y are not adjacent in π .

Definite non-collider: A vertex B is a **definite non-collider** on undirected path U if and only if either B is an endpoint of U , or there exist vertices A and C such that U contains one of the subpaths $A \leftarrow B \rightarrow C$, $A \rightarrow B \rightarrow C$, or $A \leftarrow B \rightarrow C$.

Definite non-descendant: If π is the FCI partially oriented inducing path graph of G over \mathbf{O} , then X is in **Definite-Non-Descendants**(\mathbf{Y}) if and only if there is no semi-directed path from any member of \mathbf{Y} to X in π .

Definite-SP: For a partially oriented inducing path graph π over \mathbf{O} and ordering Ord acceptable for π , V is in **Definite-SP**(Ord, X) if and only if $V \neq X$ and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and every vertex on U except for the endpoints is a collider on U .

Dependent: In an LCT or LCFS, a variable X_i is **dependent** iff X_i does not have zero indegree.

Det: $\text{Det}(\mathbf{Z})$ is the set of variables determined by any subset of \mathbf{Z} .

Determines: A set of variables **Z determines** the set of variables **A**, when every variable in **A** is a deterministic function of the variables in **Z**, and not every variable in **A** is a deterministic function of any proper subset of **Z**.

Det-connected: See Det-separation.

Det-separated: If G is a directed acyclic graph over V , Z is a subset of V that does not contain X or Y , and $X \neq Y$, then X and Y are **det-separated** given Z and **Deterministic(V)** if and only if either X and Y are d-separated given $Z \cup \text{Det}(Z)$ in some **Mod(G)** relative to **Deterministic(V)** and Z , or X or Y is in **Det(Z)**; otherwise if $X \neq Y$ and X and Y are not in Z , then X and Y are **det-connected** given Z and **Deterministic(V)**. If X , Y and Z are disjoint sets of variables in V , and X and Y are non-empty, then X and Y are **det-separated** given Z if and only if every member X of X and every member Y of Y are det-separated given Z ; otherwise if X , Y and Z are disjoint sets of variables in V , and X and Y are non-empty, then X and Y are **det-connected** given Z and **Deterministic(V)**.

Discriminating path: In an inducing path graph G' , U is a **discriminating path** for B if and only if U is an undirected path between X and Y containing B , $B \neq X$, $B \neq Y$, and

- (i) if V and V' are adjacent on U , and V' is between V and B on U , then $V \xrightarrow{*} V'$ on U ,
- (ii) if V is between X and B on U and V is a collider on U then $V \rightarrow Y$ in G' , else $V \leftarrow Y$ in G' ,
- (iii) if V is between Y and B on U and V is a collider on U then $V \rightarrow X$ in G' , else $V \leftarrow X$ in G' ,
- (iv) X and Y are not adjacent in G' .

Distributed form: The **distributed form** of an expression or equation E is the result of carrying out every multiplication, but no additions, subtractions, or divisions in E . If there are no divisions in an equation then its distributed form is a sum of terms. For example, the distributed form of the equation $u = (a + b)(c + d)v$ is $u = acv + adv + bcv + bdv$.

D-map: An acyclic graph G over V is a **D-map** of probability distribution $P(V)$ iff for every X , Y , and Z that are disjoint sets of random variables in V , if X is not d-separated from Y given Z in G then X is not independent of Y given Z in $P(V)$. However, when D-map is applied to the graph in an LCT, the quantifiers in the definitions apply only to sets of *non-error* variables.

D-Sep: If G' is an inducing path graph over \mathbf{O} and $A \neq B$, let $V \in \mathbf{D\text{-}SEP}(A,B)$ if and only if $A \neq V$ and there is an undirected path U between A and V such that every vertex on U is an ancestor of A or B , and (except for the endpoints) is a collider on U .

D-separated: If G is a directed acyclic graph with vertex set \mathbf{V} , \mathbf{Z} is a set of vertices not containing X or Y , $X \neq Y$, and X and Y are not in \mathbf{Z} , then X and Y are **D-separated** given \mathbf{Z} and **Deterministic(V)** if and only if there is no undirected path U in G between X and Y such that each collider on U has a descendant in \mathbf{Z} , and no other vertex on U is in $\text{Dett}(Z)$; otherwise if $X \neq Y$ and X and Y are not in \mathbf{Z} , then X and Y are **D-connected** given \mathbf{Z} and **Deterministic(V)**. Similarly, if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint sets of variables, and \mathbf{X} and \mathbf{Y} are non-empty, then \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} and **Deterministic(V)** if and only if each pair $\langle X, Y \rangle$ in the Cartesian product of \mathbf{X} and \mathbf{Y} are D-separated given \mathbf{Z} and **Deterministic(V)**; otherwise if \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are disjoint, and \mathbf{X} and \mathbf{Y} are non-empty, then \mathbf{X} and \mathbf{Y} are D-connected given \mathbf{Z} and **Deterministic(V)**. (Note that this is different from d-separation, which begins with a lowercase "d", and d-connection, which also begins with a lowercase "d".)

e: In an LCF F , $e(S)$ is equal to S if S is an independent variable, and it is equal to the error variable into S if S is not an independent variable.

E: If X is a random variable, $E(X)$ is the expected value of X .

Equiv(G'): If G' is an inducing path graph over \mathbf{O} , **Equiv(G')** is the set of inducing path graphs over the same vertices with the same d-connections as G .

E.C.F: See equation coefficient factor.

Equation coefficient: See linear causal theory, linear causal form.

Equation coefficient factor: In an LCF or LCT T , if an expression is equal to ce , where c is a non-zero constant, and e is a product of equation coefficients raised to positive integral powers, then e is the **equation coefficient factor(e.c.f.)** of ce .

Equivalent to a polynomial: In an LCF, a quantity (e.g. a covariance) X is **equivalent to a polynomial in the coefficients and variances of exogenous variables** if and only if for each LCF $F = \langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle$, C , V , $\mathbf{EQ}, \mathbf{L}, \mathbf{Err}$ and in every LCT $S = \langle \mathbf{R}', \mathbf{M}', \mathbf{E}' \rangle$, $(\Omega f, P)$, $\mathbf{EQ}', \mathbf{L}', \mathbf{Err}' \rangle$ that is an instance of F , there is a polynomial in the

variables in \mathbf{C} and \mathbf{V} such that X is equal to the result of substituting the linear coefficients of S in as values for the corresponding variables in \mathbf{C} , and the variances of the exogenous variables in S as values for the corresponding variables in \mathbf{V} .

Error variable: See linear causal theory, linear causal form.

Exogenous: If G is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, and $\mathbf{V} \cap \mathbf{W} = \emptyset$, then \mathbf{W} is **exogenous with respect to \mathbf{V}** in G if and only if there is no directed edge from any member of \mathbf{V} to any member of \mathbf{W} .

Faithfully indistinguishable: We will say that two directed acyclic graphs, G, G' are **faithfully indistinguishable** (f.i.) if and only if every distribution faithful to G is faithful to G' and vice-versa.

F.I.: See faithfully indistinguishable.

Final segment: In a graph G , a path U of length n is a **final segment** of path V of length m iff $m \geq n$, and for $1 \leq i \leq n+1$, the i^{th} vertex of V equals the $(m-n+i)^{\text{th}}$ vertex of U .

I-Map: An acyclic directed graph G over \mathbf{V} is an **I-map** of probability distribution $P(\mathbf{V})$ iff for every \mathbf{X}, \mathbf{Y} , and \mathbf{Z} that are disjoint sets of random variables in \mathbf{V} , if \mathbf{X} is d-separated from \mathbf{Y} given \mathbf{Z} in G then \mathbf{X} is independent of \mathbf{Y} given \mathbf{Z} in $P(\mathbf{V})$. However, when I-map is applied to the graph in an LCT, the quantifiers in the definitions apply only to sets of *non-error* variables.

Ind: For a directed acyclic graph G , **Ind** is the set of independent variables in G .

Ind_{IJ} : Ind_{IJ} is the coefficient of J in the independent equational for I . See also independent equational.

Independent: In an LCT or LCFS, a variable X_i is **independent** iff X_i has zero indegree (i.e. there are no edges directed into it). Note that the *property* of independence is completely distinct from the *relation* of statistical independence. The context will make clear in which of these senses the term is used.

Independent equational: In an LCF $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle\rangle, \mathbf{C}, \mathbf{V}, \mathbf{EQ}, \mathbf{L}, \mathbf{S} \rangle$ an equation is an **independent equational for a dependent variable X_j** if and only if it is implied by \mathbf{EQ} and the variables in \mathbf{R} which appear on the r.h.s. are independent and occur at most once

Inducing path: If G is a directed acyclic graph over a set of variables V , O is a subset of V containing A and B , and $A \neq B$, then an undirected path U between A and B is an **inducing path relative to O** if and only if every member of O on U except for the endpoints is a collider on U , and every collider on U is an ancestor of either A or B . We will sometimes refer to members of O as **observed** variables.

Inducing path graph: G' is an **inducing path graph over O for directed acyclic graph G** if and only if O is a subset of the vertices in G , there is an edge between variables A and B with an arrowhead at A if and only if A and B are in O , and there is an inducing path in G between A and B relative to O that is into A . (Using the notation of Chapter 2, the set of marks in an inducing path graph is $\{>, EM\}$.)

Initial segment: In a graph G , a path U of length n is an **initial segment** of path V of length m iff $m \geq n$, and for $1 \leq i \leq n+1$, the i^{th} vertex of V equals the i^{th} vertex of U .

Into: In a graph G , an edge between A and B is into A if and only if the mark at the A end of the edge is an " $>$ ". If an undirected path U between A and B contains an edge into A we will say that U is **into A** .

Invariant: If G is a directed acyclic graph over a set of variables $V \cup W$, W is exogenous with respect to V in G , Y and Z are disjoint subsets of V , $P(V \cup W)$ is a distribution that satisfies the Markov condition for G , and **Manipulated(W) = X** , then $P(Y|Z)$ is **invariant** under direct manipulation of X in G by changing W from w_1 to w_2 if and only if $P(Y|Z, W = w_1) = P(Y|Z, W = w_2)$ wherever they are both defined.

Instance: An LCT S is an **instance** of an LCF F if and only if the graph of S is isomorphic to the graph of F .

IP: In a directed acyclic graph G , if $Y \cap Z = \emptyset$, W is in **IP(Y, Z)** (W has a parent that is an informative variable for Y given Z) if and only if W is a member of Z , and W has a parent in **IV(Y, Z)**.

IV: In a directed acyclic graph G , if $Y \cap Z = \emptyset$, then V is in **IV(Y, Z)** (informative variables for Y given Z) if and only if V is d-connected to Y given Z , and V is not in **ND(Y, Z)**. (This entails that V is not in $Y \cup Z$.)

Label: See linear causal theory, linear causal form.

Length: In a graph G , the **length** of a path equals the number of vertices in the path minus one.

Last point of intersection: In a directed acyclic graph G , the **last point of intersection** of directed path $R(U,I)$ with directed path $R(V,J)$ is the last vertex on $R(U,I)$ that is also on $R(V,J)$. Note that if G is a directed acyclic graph, the last point of intersection of directed path $R(U,I)$ with directed path $R(V,J)$ equals the last point of intersection of $R(V,J)$ with $R(U,I)$; this is not true of directed cyclic paths.

LCF: See linear causal form.

LCT: See linear causal theory.

Linear causal form: A **linear causal form** is an unestimated LCT in which the linear coefficients and the variances of the exogenous variables are real variables instead of constants. This entails that an edge label in an LCF is a real variable instead of a constant (except that the label of an edge from an error variable is fixed at one.) More formally, let a linear causal form (LCF) be $\langle\langle R, M, E \rangle, C, V, EQ, L, Err \rangle$ where

- i. $\langle R, M, E \rangle$ is a directed acyclic graph. Err is a subset of R called the **error variables**. Each error variable is of indegree 0 and outdegree 1. For every X_i in R of indegree $\neq 0$ there is exactly one error variable with an edge into X_i .
- ii. c_{ij} is a unique real variable associated with an edge from X_j to X_i , and C is the set of c_{ij} . V is the set of variables σ_i^2 , where X_i is an exogenous variable in $\langle R, M, E \rangle$ and σ_i^2 is a variable that ranges over the positive real numbers.
- iii. L is a function with domain E such that for each e in E , $L(e) = c_{ij}$ iff $head(e) = X_i$ and $tail(e) = X_j$. $L(e)$ will be called the **label** of e . By extension, the product of labels of edges in any acyclic undirected path U will be denoted by $L(U)$, and $L(U)$ will be called the **label** of U . The label of an empty path is fixed at 1.
- iv. EQ is a consistent set of independent homogeneous linear equationals in variables in R . For each X_i in R of positive indegree there is an equation in EQ of the form

$$X_i = \sum_{X_j \in \text{Parents}(X_i)} c_{ij} X_j$$

where each c_{ij} is a real variable in \mathbf{C} and each X_i is in \mathbf{R} . There are no other equations in \mathbf{EQ} . c_{ji} is the **equation coefficient** of X_j in the equation for X_i .

Linear causal theory: Let a **linear causal theory** be **(LCT)** be $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle\rangle$, (Ω, f, P) , $\mathbf{EQ}, \mathbf{L}, \mathbf{Err}$ where

- i. (Ω, f, P) is a probability space, where Ω is the sample space, f is a sigma-field over Ω , and P is a probability distribution over f .
- ii. $\langle\langle \mathbf{R}, \mathbf{M}, \mathbf{E} \rangle\rangle$ is a directed acyclic graph. \mathbf{R} is a set of random variables over (Ω, f, P) .
- iii. The variables in \mathbf{R} have a joint distribution. Every variable in \mathbf{R} has a non-zero variance. \mathbf{E} is a set of directed edges between variables in \mathbf{R} . (\mathbf{M} is the set of marks that occur in a directed graph, i.e. $\{\mathbf{EM}, >\}$).
- iv. \mathbf{EQ} is a consistent set of independent homogeneous linear equations in random variables in \mathbf{R} . For each X_i in \mathbf{R} of positive indegree there is an equation in \mathbf{EQ} of the form

$$X_i = \sum_{X_j \in \mathbf{Parents}(X_i)} a_{ij} X_j$$

where each a_{ij} is a non-zero real number and each X_i is in \mathbf{R} . This implies that each vertex X_i in \mathbf{R} of positive indegree can be expressed as a linear function of all and only its parents. There are no other equations in \mathbf{EQ} . A non-zero value of a_{ij} is the **equation coefficient** of X_j in the equation for X_i .

- v. If vertices (random variables) X_i and X_j are exogenous, then X_i and X_j are pairwise statistically independent.
- vi. L is a function with domain E such that for each e in E , $L(e) = a_{ij}$ iff **head**(e) = X_i and **tail**(e) = X_j . $L(e)$ will be called the **label** of e . By extension, the product of labels of edges in any acyclic undirected path U will be denoted by $L(U)$, and $L(U)$ will be called the **label** of U . The label of an empty path is fixed at 1.
- vii. There is a subset of \mathbf{S} of \mathbf{R} called the **error variables**, each of indegree 0 and outdegree Note that the variance of any endogenous variable I conditional on any set of variables that does not contain the error variable of I is not equal to zero.

Linear Representation: A directed acyclic graph G over V **linearly represents** a distribution $P(V)$ if and only if there exists a directed acyclic graph G' over V' and a distribution $P''(V')$ such that

- (i) V is included in V' ;

- (ii) for each endogenous (that is, with positive indegree) variable X in \mathbf{V} , there is a unique variable ε_X in $\mathbf{V} \setminus \mathbf{V}$ with zero indegree, positive variance, outdegree equal to one, and a directed edge from ε_X to X ;
- (iii) G is the subgraph of G' over \mathbf{V} ;
- (iv) each endogenous variable in G is a linear function of its parents in G' ;
- (v) in $P'(\mathbf{V}')$ the correlation between any two exogenous variables in G' is zero;
- (vi) $P(\mathbf{V})$ is the marginal of $P'(\mathbf{V}')$ over \mathbf{V} .

The members of $\mathbf{V} \setminus \mathbf{V}$ are called **error variables** and we call G' the **expanded graph**.

Linearly implies: A directed acyclic graph G **linearly implies** $\rho_{AB,H} = 0$ if and only if $\rho_{AB,H} = 0$ in all distributions linearly represented by G . (We assume all partial correlations are defined for the distribution.)

Manipulate: See manipulation.

Manipulated graph: See manipulation.

Manipulation: If G is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, and $\mathbf{V} \cap \mathbf{W} = \emptyset$, then \mathbf{W} is **exogenous with respect to \mathbf{V}** in G if and only if there is no directed edge from any member of \mathbf{V} to any member of \mathbf{W} . If G_{Comb} is a directed acyclic graph over a set of variables $\mathbf{V} \cup \mathbf{W}$, and $P(\mathbf{V} \cup \mathbf{W})$ satisfies the Markov condition for G_{Comb} , then changing the value of \mathbf{W} from \mathbf{w}_1 to \mathbf{w}_2 is a **manipulation** of G_{Comb} with respect to \mathbf{V} if and only if \mathbf{W} is exogenous with respect to \mathbf{V} , and $P(\mathbf{V}|\mathbf{W} = \mathbf{w}_1) \neq P(\mathbf{V}|\mathbf{W} = \mathbf{w}_2)$. We define $P_{Unman(\mathbf{W})}(\mathbf{V}) = P(\mathbf{V}|\mathbf{W} = \mathbf{w}_1)$, and $P_{Man(\mathbf{W})}(\mathbf{V}) = P(\mathbf{V}|\mathbf{W} = \mathbf{w}_2)$, and similarly for various marginal and conditional distributions formed from $P(\mathbf{V})$. We refer to G_{Comb} as the **combined graph**, and the subgraph of G_{Comb} over \mathbf{V} as the **unmanipulated graph** G_{Unman} . V is in **Manipulated(\mathbf{W})** (that is, V is a variable directly influenced by one of the manipulation variables) if and only if V is in **Children(\mathbf{W}) \cap \mathbf{V}** ; we will also say that the variables in **Manipulated(\mathbf{W})** have been **directly manipulated**. We will refer to the variables in \mathbf{W} as **policy variables**. The **manipulated graph**, G_{Man} is a subgraph of G_{Unman} for which $P_{Man(\mathbf{W})}(\mathbf{V})$ satisfies the Markov Condition and which differs from G_{Unman} in at most the parents of members of **Manipulated(\mathbf{W})**.

Minimal I-map: An acyclic graph G is a **minimal I-map** of probability distribution P iff G is an I-map of P , and no subgraph of G is an I-map of P . However, when minimal I-map is

applied to the graph in an LCT, the quantifiers in the definitions apply only to sets of *non-error* variables.

Mod: If G is a directed acyclic graph over V , and Z is included in V , then G' is in $\text{Mod}(G)$ relative to **Deterministic(V)** and Z if and only if for each V in V

- (i) if there exists a set of vertices included in Z that are non-descendants of V in G and that determine V , then $\text{Parents}(G', V) = X$, where X is some set of vertices included in Z that are non-descendants of V in G and that determine V ;
- (ii) if there is no set X of vertices included in Z that are non-descendants of V in G and that determine V , then $\text{Parents}(G', V) = \text{Parents}(G, V)$.

ND: In a directed acyclic graph G , $\text{ND}(Y)$ is the set of all vertices that do not have a descendant in Y .

Non-Descendants: In a directed acyclic graph G , X is in **Non-Descendants(Y)** if and only if there is no directed path from any member of Y to X in G .

Observed: See inducing path graph, inducing path.

Out of: In a graph G , an edge between A and B is out of A if and only if the mark at the A endpoint is the empty mark. If an undirected path U between A and B contains an edge out of A we will say that U is **out of A**.

Parallel embedding: Directed acyclic graphs G_1 and G_2 with common vertex set O have a **parallel embedding** in directed acyclic graphs H_1 and H_2 having a common set U of vertices that includes O if and only if

- (i) G_1 is the subgraph of H_1 over O and G_2 is the subgraph of H_2 over O ;
- (ii) every directed edge in H_1 but not in G_1 is in H_2 and every directed edge in H_2 but not in G_2 is in H_1 .

Path form: If G is a directed acyclic graph, let Let P_{XY} be the set of all directed paths in G from X to Y . In an LCF S , the **path form of a product of covariances** $\gamma_{IJ}\gamma_{KL}$ is the distributed form of

$$\left(\sum_{U \in \mathbf{U}_{IJ}} \left(\sum_{R \in \mathbf{P}_{UI}} \sum_{R' \in \mathbf{P}_{UJ}} L(R)L(R')\sigma_U^2 \right) \right) \left(\sum_{V \in \mathbf{U}_{KL}} \left(\sum_{R'' \in \mathbf{P}_{VK}} \sum_{R''' \in \mathbf{P}_{VL}} L(R'')L(R''')\sigma_V^2 \right) \right)$$

$\gamma_{JJKL} - \gamma_{LYJK}$ is in **path form** iff both terms are in path form. $\gamma_{JJKL} - \gamma_{LYJK}$ is in **path form** iff both terms are in path form.

Policy variables: See manipulate.

Possible-D-SEP(A,B): If $A \neq B$ in partially oriented inducing path graph π , V is in **Possible-D-Sep(A,B)** in π if and only if $V \neq A$, and there is an undirected path U between A and V in π such that for every subpath $\langle X,Y,Z \rangle$ of U either Y is a collider on the subpath, or Y is not a definite non-collider on U , and X, Y , and Z form a triangle in π .

Possibly d-connecting: If A and B are not in \mathbf{Z} , and $A \neq B$, then an undirected path U between A and B in a partially oriented inducing path graph π over \mathbf{O} is a **possibly d-connecting** path of A and B given \mathbf{Z} if and only if every collider on U is the source of a semi-directed path to a member of \mathbf{Z} , and every definite non-collider is not in \mathbf{Z} .

Possibly-IP: If π is a partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IP(\mathbf{Y},\mathbf{Z})** if and only if \mathbf{Y} and \mathbf{Z} are disjoint, X is in \mathbf{Z} , and there is a possibly d-connecting path between X and some Y in \mathbf{Y} given $\mathbf{Z} \setminus \{X\}$ that is not out of X .

Possibly-IV: If π is a partially oriented inducing path graph of G over \mathbf{O} , then X is in **Possibly-IV(\mathbf{Y},\mathbf{Z})** if and only if X is not in \mathbf{Z} , there is a possibly d-connecting path between X and some Y in \mathbf{Y} given \mathbf{Z} , and there is a semi-directed path from X to a member of $\mathbf{Y} \cup \mathbf{Z}$.

Possible-SP: For a partially oriented inducing path graph π and ordering Ord acceptable for π , let V be in **Possible-SP(Ord,X)** if and only if $V \neq X$ and there is an undirected path U in π between V and X such that every vertex on U except for X is a predecessor of X in Ord , and no vertex on U except for the endpoints is a definite-non-collider on U .

Predecessors: For inducing path graph G' and acceptable total ordering Ord , let **Predecessors(Ord,V)** equal the set of all variables that precede V (not including V) according to Ord .

Proper final segment: A path U of length n is a **proper final segment** of path V of length m iff U is a final segment of V and $U \neq V$.

Proper initial segment: A path U of length n is a **proper initial segment** of path V of length m iff U is an initial segment of V and $U \neq V$.

$P_{Man(W)}(V)$: See manipulate.

$P_{Unman(W)}(V)$: See manipulate.

Pure Latent Variable Graph: A **pure latent variable graph** is a directed acyclic graph in which each measured variable is a child of exactly one latent variable, and a parent of no other variable.

Random coefficient linear causal theory: The definition of a **random coefficient linear causal theory** is the same as that of a linear causal theory except that each linear coefficient is a random variable independent of the set of all other random variables in the model.

Rigidly statistically indistinguishable: If directed acyclic graphs G and G' are strongly statistically indistinguishable and every parallel embedding of G and G' is strongly statistically indistinguishable then structures G and G' are **rigidly statistically indistinguishable (r.s.i.)**.

R.S.I.: See rigidly statistically indistinguishable.

Semi-directed: A **semi-directed path from A to B** in partially oriented inducing path graph π is an undirected path U from A to B in which no edge contains an arrowhead pointing towards A , that is, there is no arrowhead at A on U , and if X and Y are adjacent on the path, and X is between A and Y on the path, then there is no arrowhead at the X end of the edge between X and Y .

Source: See trek.

SP: For inducing path graph G' and acceptable total ordering Ord , W is in $SP(Ord, G', V)$ (separating predecessors of V in G' for ordering Ord) if and only if $W \neq V$ and there is an undirected path U between W and V such that each vertex on U except for V precedes V in Ord and every vertex on U except for the endpoints is a collider on U .

S.S.I.: See strongly statistically indistinguishable.

Strongly statistically indistinguishable: Two directed acyclic graphs G, G' are **strongly statistically indistinguishable** if and only if they have the same vertex set V and every distribution P on V satisfying the Minimality and Markov Conditions for G satisfies those conditions for G' , and vice-versa.

Substitutable: In an inducing path or directed acyclic graph G that contains an undirected path U between X and Y , the edge between V and W is **substitutable** for $U(V,W)$ in U if and only if V and W are on U , V is between X and W on U , G contains an edge between V and W , V is a collider on the concatenation of $U(X,V)$ and the edge between V and W if and only if it is a collider on U , and W is a collider on the concatenation of $U(Y,W)$ and the edge between V and W if and only if it is a collider on U .

T: See trek.

Termini: See trek.

Trek: A **trek** $T(I,J)$ between two distinct vertices I and J is an unordered pair of acyclic directed paths from some vertex K to I and J respectively that intersect only at K . The source of the paths in the trek is called the **source** of the trek. I and J are called the **termini of the trek**. Given a trek $T(I,J)$ between I and J , $I(T(I,J))$ will denote the path in $T(I,J)$ from the source of $T(I,J)$ to I and $J(T(I,J))$ will denote the path in $T(I,J)$ from the source of $T(I,J)$ to J . One of the paths in a trek may be an empty path. However, since the termini of a trek are distinct, only one path in a trek can be empty. $\mathbf{T}(I,J)$ is the set of all treks between I and J . $T(I,J)$ will represent a trek in $\mathbf{T}(I,J)$. $S(T(I,J))$ represents the source of the trek $T(I,J)$.

Undirected: In a graph G , Let V be in **Undirected**(X,Y) if and only if V lies on some undirected path between X and Y .

Unmanipulated graph: See manipulation.

U_X : In an LCF S , U_X is the set of all independent variables that are the source of a directed path to X . (Note that if X is independent then $X \in U_X$ since there is an empty path from every vertex to itself.)

U_{XY} : In an LCF S , U_{XY} is $U_X \cap U_Y$.

Weakly faithfully indistinguishable: Two directed acyclic graphs are **weakly faithfully indistinguishable** (w.f.i.) if and only if there exists a probability distribution faithful to both of them.

Weakly statistically indistinguishable: Two directed acyclic graphs are **weakly statistically indistinguishable** (w.s.i.) if and only if there exists a probability distribution meeting the Minimality and Markov Conditions for both of them.

W.F.I.: See weakly faithfully indistinguishable.

W.S.I.: See weakly statistically indistinguishable.

Bibliography

- Aigner, D. and Goldberger, A. (1977). Latent Variables in Socio-economic Models. North-Holland, Amsterdam.
- Aitkin, M. (1979). A simultaneous test procedure for contingency table models. *Appl. Statist.* 28, 233-242.
- Allen, D. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 17, 125-127.
- Anderson, J., and Gerbing, D. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research*. 19, 453-60.
- Anderson, T. (1984). An Introduction to Multivariate Statistical Analysis. Wiley, New York.
- Asher, H. (1976). Causal Modeling. Sage Publications, Beverly Hills, CA.
- Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* 70, 567-578.
- Bagozzi, Richard P. (1980). Causal Models in Marketing. Wiley, New York.
- Bartlett, M. (1935). Contingency table interaction. *J. Roy. Statist. Soc. Suppl.* 2, 248-252.
- Bartlett, M. (1954). A note on the multiplying factors for various chi-squared approximations. *J. Roy. Statist. Soc. Ser. B* 16, 296-298.
- Basman, R. (1965). A note on the statistical testability of 'explicit causal chains' against the class of 'interdependent' models. *JASA*, 60, 1080-1093.
- Beale, E., Kendall, M., and Mann, D. (1967). The discarding of variables in multivariate analysis. *Biometrika* 54, 357-366.

- Becker, G. (1964). *Human Capital*. National Bureau of Economic Research, New York.
- Beinlich, I., Suermondt, H., Chavez, R., and Cooper, G. (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. Proc. Second European Conference on Artificial Intelligence in Medicine, London, England. 247-256.
- Bentler, P. (1985). Theory and Implementation of EQS: A Structural Equations Program. BMDP Statistical Software Inc., Los Angeles.
- Bentler, P. (1980). Multivariate analysis with latent variables: causal modeling. *Annual Review of Psychology* 31, 419-456.
- Bentler, P. and Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88, 588-606.
- Bentler, P., and Peeler, W. (1979). Models of female orgasm. *Archives of Sexual Behavior* 8, 405-423.
- Birch, M. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc.* 25, 220-223.
- Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.
- Blalock, H. (1961). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill, NC.
- Blalock, H. (1969). *Theory Construction: From Verbal to Mathematical Formulations*. Prentice-Hall, Englewood Cliffs, NJ.
- Blalock, H. (1971). *Causal Models in the Social Sciences*. Aldine-Atherton, Chicago.
- Blau, P. and Duncan, O. (1967). *The American Occupational Structure*. Wiley, New York.

- Blum, R. (1984). Discovery, confirmation and incorporation of causal relationships from a time-oriented clinical data base: The RX project. Readings in Medical Artificial Intelligence, Clancey, W. and Shortliffe, E. (eds.). Addison-Wesley, Reading, MA.
- Blyth, C. (1972). On Simpson's paradox and the sure-thing principle. JASA 67, 364-366.
- Bollen, K. (1989). Structural Equations with Latent Variables. Wiley, New York.
- Bollen, K. (1990). Outlier screening and a distribution-free test for vanishing tetrads. Sociological Methods and Research 19, 80-92.
- Bowden, R. and Turkington, D. (1984). Instrumental Variables. Cambridge University Press, New York.
- Breslow, N. and Day, N. (1980). Statistical Methods in Cancer Research, Vol 1: The Analysis of Case-Control Studies. IARC, Lyon.
- Brownlee, K. (1965). A review of "Smoking and Health." JASA 60, 722-739.
- Bunker, J., Forrest, W., Mosteller, F. and Vandam, L. (1969). The National Halothane Study: Report of the Subcommittee on the National Halothane Study of the Committee on Anesthesia, Division of Medical Sciences, National Academy of Sciences. National Research Council, Washington, D.C. U.S. Government Printing Office.
- Burch, P. (1978). Smoking and lung cancer: The problem of inferring cause (with discussion). J. Roy. Statist. Soc. Ser. A 141, 437-477.
- Burch, P. (1983). The Surgeon General's "Epidemiologic Criteria for Causality." A critique. Journal of Chronic Diseases 36, 821-836.
- Burch, P. (1984). The Surgeon General's "Epidemiologic Criteria for Causality." Reply to Lilienfeld. Journal of Chronic Diseases, 37, 148-157.
- Byron, R. (1972). Testing for misspecification in econometric systems using full information. International Economic Review 28, 138-151.

Callahan, J., and Sorensen, S., (1992) Using TETRAD II as an automated exploratory tool. *Sociological Methods and Research*. Fall

Campbell, D. and Stanley, J. (1963). Experimental and Quasi-Experimental Designs. Rand McNally, Chicago.

Campbell, D., Schwartz, R., Sechrest, L., and Webb, E. (1966). Unobtrusive Measures: Nonreactive Research in the Social Sciences. Rand McNally, Chicago.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive processes from patterns of impaired performance: the case for single patient studies. *Brain and Cognition*, 5, 41-66.

Cartwright, N. (1983). How the Laws of Physics Lie. Oxford University Press, New York.

Cartwright, N. (1989). Nature's Capacities and Their Measurement. Clarendon Press, Oxford.

Cavallo, R. and Klir, G. (1979). Reconstructability analysis of multi-dimensional relations: A theoretical basis for computer-aided determination of acceptable systems models. *International Journal of General Systems* 5, 143-171.

Cederlof, R., Friberg, L., Lundman, T. (1972). The interactions of smoking, environment and heredity and their implications for disease etiology. *Acta Med Scand*. 612 (Suppl).

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* IT-14, 462-467.

Chow, C. and Wagner, T. (1973). Consistency of an estimate of tree-dependent probability distributions. *IEEE Trans. on Info. Theory* IT-19, 369-371.

Christensen, R. (1990). Log-Linear Models. Springer-Verlag, New York.

Coleman, J. (1964). Introduction to Mathematical Sociology. Free Press, New York.

Cooper, G. and Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief networks from databases. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. LA, CA. 86-94.

- Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* (to appear).
- Cooper, G. (1989). Current research in the development of expert systems based on belief networks. *Applied Stochastic Models and Data Analysis* 5, 39-52.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22, 173-203.
- Costner, H. (1971). Theory, deduction and rules of correspondence. *Causal Models in the Social Sciences*, Blalock, H. (ed.). Aldine, Chicago.
- Costner, H. and Schoenberg, R. (1973). Diagnosing indicator ills in multiple indicator models, in *Structural Equation Models in the Social Sciences*, Goldberger, A. and Duncan, O. (eds.). Seminar Press, New York.
- Costner, H. and Herting, J. (1985). Respecification in multiple indicator models, in *Causal Models in the Social Sciences*, 2nd ed., Blalock, H. (ed.), 321-393. Aldine, New York.
- Cox, D. (1958). *Planning of Experiments*. Wiley, New York.
- Crawford, S. and Fung, R. (1990). An Analysis of Two Probabilistic Model Induction Techniques. Third International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL.
- Darroch, J., Lauritzen, S., and Speed, T. (1980). Markov fields and log linear interaction models for contingency tables. *Ann. Stat.* 8, 522-539.
- Davis, W. (1988). Probabilistic theories of causation. *Probability and Causality*, James Fetzer (ed.). D. Reidel, Dordrecht.
- Dawid, A. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* 41, 1-31.
- Dempster, A. (1972). Covariance selection. *Biometrics* 28, 157-175.

- Doll, R. and Hill, A. (1952). A study of the aetiology of carcinoma of the lung. *Brit. Med. J.* 2, 1271-1286.
- Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.
- Duncan, O., Featherman, D., and Duncan, B. (1972). *Socioeconomic Background and Achievement*. Seminar Press, New York.
- Edwards, A. (1976). *An Introduction to Linear Regression and Correlation*. W. H. Freeman, New York.
- Edwards, D. and Havranek, T. (1985). A fast procedure for model search in multi-dimensional contingency tables. *Biometrika* 72, 339-351.
- Edwards, D. and Havranek, T. (1987). A fast model selection procedure for large families of models. *J. Amer. Statist. Assoc.* 82, 205-211.
- Edwards, D. and Kreiner, S. (1983). The analysis of contingency tables by graphical models. *Biometrika* 70, 553-565.
- Elby, A. (1992) Should we explain the EPR causally? *Philosophy of Science* 59, 16-25.
- Fienberg, S. (1977). *The Analysis of Cross-classified Categorical Data*. MIT Press, Cambridge, MA.
- Fine, A. (1982). Hidden variables, joint probability, and the Bell inequalities. *Physical Review Letters* 48, 291-295.
- Fine, T. (1973). *Theories of Probability*. Academic Press, New York.
- Fisher, F. (1966). *The Identification Problem in Economics*. McGraw-Hill, New York.
- Fisher, R. (1951). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- Fisher, R. (1959). *Smoking. The Cancer Controversy*. Oliver and Boyd, Edinburgh.

- Flack, V. and Chang, P. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *American Statistician* 41, 84-86.
- Forbes, H. and Tufte, E. (1968). A note of caution in causal modeling. *American Political Science Review* 62, 1258-1264.
- Fox, J. (1984). *Linear Statistical Models and Related Methods*. Wiley, New York.
- Freedman, D. (1983a). Structural-equation models: A case study. Report, Department of Statistics, University of California, Berkeley.
- Freedman, D. (1983b). A note on screening regression equations. *American Statistician* 37, 152-155.
- Freedman, D., Navidi, W., and Peters, S. (1986). On the impact of variable selection in fitting regression equations. In *Model Uncertainty and its Statistical Implications*. Lecture Notes in Economics and Mathematical Systems 307, Dijkstra, T. (ed.). Springer-Verlag, Berlin.
- Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton University Press, Princeton, NJ.
- Fung, R. and Crawford, S. (1990). Constructor: A system for the induction of probabilistic models. *Proceedings of the Eighth National Conference on AI*, Boston, AAAI.
- Furnival, G., and Wilson, R. (1974). Regression by leaps and bounds. *Technometrics* 16, 4990-5111
- Geiger, D. (1990). Graphoids: A Qualitative Framework for Probabilistic Inference. Ph.D. Thesis, University of California, Los Angeles.
- Geiger, D., and Heckerman, D. (1991) Advances in probabilistic reasoning. Proc. Seventh Conference on Uncertainty in AI, B. D'Ambrosio et. al. (ed.). Morgan Kauman, Los Angeles, California.
- Geiger, D. and Pearl, J. (1989a). Logical and Algorithmic Properties of Conditional Independence and Qualitative Independence. Report CSD 870056, R-97-IIL, Cognitive Systems Laboratory, University of California, Los Angeles.

- Geiger, D. and Pearl, J. (1989b). Axioms and Algorithms for Inferences Involving Conditional Independence. Report CSD 890031, R-119-I, Cognitive Systems Laboratory, University of California, Los Angeles.
- Geiger, D., Verma, T., and Pearl, J. (1990) Identifying independence in Bayesian Networks. *Networks* 20, 507-533.
- Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *JASA* 74, 153-160.
- Geweke, J., Meese, R., and Dent, W. (1983). Comparing alternative tests of causality in temporal systems. *Journal of Econometrics* 21, 161-194.
- Glymour, C. (1983). Social science and social physics. *Behavioral Science* 28, 126-133.
- Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure*. Academic Press, San Diego, CA.
- Glymour, C., Scheines, R., and Spirtes, P. (1989). Why Aviators Leave the Navy: Applications of Artificial Intelligence Procedures in Manpower Research. Report to the Naval Personnel Research Development Center, January.
- Glymour, C., Spirtes, P., and Scheines, R. (1991a). Independence relations produced by parameter values. *Philosophical Topics*, v. 18, no. 2, Fall.
- Glymour, C., Spirtes P., and Scheines, R. (1991b). From probability to causality. *Philosophical Studies*, v. 64, no. 1, 1-36.
- Gold, E. (1967). Language identification in the limit. *Information and Control* 10, 447-474.
- Gold, E. (1965). Limiting recursion. *Journal of Symbolic Logic* 30, 27- 48.
- Goldberg, A., Duncan, O. (eds.) (1973). *Structural Equation Models in the Social Sciences*. Seminar Press, New York.

- Goodman, L. (1973a). Causal analysis of data from panel studies and other kinds of surveys. *Amer. J. Sociol.* 78, 1135-1191.
- Goodman, L. (1973b). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika* 60, 179-192.
- Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424-438.
- Greenland, S. (1989). Modelling variable selection in epidemiologic analysis. *American Journal of Public Health* 79, 340-349.
- Griffiths, W., Hill, R., and Pope, P. (1987). Small sample properties of probit model estimators. *JASA* 82, 929-937.
- Haberman, S.J. (1979). *Analysis of Qualitative Data. Volume 2.* Academic Press.
- Harary, F., Norman R., and Cartwright, D. (1965). *Structural Models: An Introduction to the Theory of Directed Graphs.* Wiley, New York.
- Harary, F., and Palmer, E. (1973). *Graphical Enumeration.* Academic Press, New York.
- Hausman, D. (1984). Causal priority. *Nous* 18, 261-279.
- Havranek, T. (1984). A procedure for model search in multi-dimensional contingency tables. *Biometrics* 40, 95-100.
- Heise, D. (1975). *Causal Analysis.* Wiley, New York.
- Herskovits, E., and Cooper, G. (1990). Kutato: An entropy-driven system for construction of probabilistic expert systems from databases. Proc. Sixth Conf. Uncertainty in AI. Association for Uncertainty in AI, Inc., Mountain View, CA.
- Herskovits, E. (1992). Computer Based Probabilistic-Network Construction. Ph.D Thesis, Departments of Computer Science and Medicine, Stanford University.

- Herting, J. and Costner, J. (1985). Respecification in multiple indicator models. *Causal Models in the Social Sciences*, Blalock, H. (ed.). Aldine, NY.
- Hocking, R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics* 9, 531-540.
- Holland, P. (1986). Statistics and causal inference. *JASA* 81, 945-960.
- Holland, P. and Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Stat.* 14, 1523-1543.
- Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New York.
- Howson, C., and Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, Illinois.
- Hsiao, C. (1981). Autoregressive modelling and money-income causality deduction. *Journal of Monetary Economics* 7, 85-106.
- James, L., Mulaik, S., and Brett, J. (1982). *Causal Analysis: Assumptions, Models and Data*. Sage Publications, Beverly Hills, CA.
- Jeffreys, H. (1957). *Scientific Inference*. Cambridge University Press, New York.
- Joreskog, K. (1973). A general method for estimating a linear structural equation. *Structural Equation Models in the Social Sciences*, Goldberger, A., and Duncan, O. (eds.). Seminar Press, New York.
- Joreskog, K. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* 43, 443-447.
- Joreskog, K. (1981). Analysis of covariance structures. *Scandinavian Journal of Statistics* 8, 65-92.
- Joreskog, K. and Sorbom, D. (1984). *LISREL VI User's Guide*. Scientific Software, Inc., Mooresville, IN.

Joreskog, K. and Sorbom, D. (1990). Model search with TETRAD II and LISREL. *Sociological Methods and Research* 19, 93-106.

Kadane, J. and Sedransk, N. (1980). Toward a more ethical clinical trial. *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain), May 28 to June 2, 1979*, edited by J. Bernardo et al., University Press, 229-238.

Kadane, J. and Sedransk, N. (1992). Results of the clinical trial. *Toward a More Ethical Clinical Trial*, J. Kadane (ed.), John Wiley & Sons, NY, forthcoming.

Kadane, J. and Seidenfeld, T. (1990). Randomization in a Bayesian Perspective. *Journal of Statistical Planning and Inference*, 25, North-Holland, 329-345.

Kadane, J. and Seidenfeld, T. (1992). Statistical issues in the analysis of data gathered in the new designs. *Toward a More Ethical Clinical Trial*, J. Kadane (ed.), John Wiley & Sons, NY, forthcoming.

Kelley, T. (1928). *Crossroads in the Mind of Man*. Stanford University Press, Stanford.

Kendall, M. (1948). *The Advanced Theory of Statistics*. Charles Griffin and Co., London.

Kenny, D. (1979). *Correlation and Causality*. Wiley, New York.

Kiiveri H. (1982). A Unified Approach to Causal Models. Ph.D. thesis, Univ. of Western Australia, in preparation.

Kiiveri, H. and Speed, T. (1982). Structural analysis of multivariate data: A review. *Sociological Methodology*, Leinhardt, S. (ed.). Jossey-Bass, San Francisco.

Kiiveri, H., Speed, T., and Carlin, J. (1984). Recursive causal models. *Journal of the Australian Mathematical Society* 36, 30-52.

Klein, L., (1961). *An Econometric Model of the United Kingdom*. Oxford University, Institute of Statistics, Oxford.

Kleinbaum, D, Kupper, L., and Morgenstern, H. (1982). *Epidemiologic Research*. Lifetime Learning Publications, Belmont, CA.

- Klepper, S. (1988). Regressor diagnostics for the classical errors in variables model. *Journal of Econometrics* 37, 225-250.
- Klir, G., and Parviz, B. (1986). General reconstruction characteristics of probabilistic and possibilistic systems. *International Journal of Man-Machine Studies* 25, 367-397.
- Kullback, S. (1959, 1968). *Information Theory and Statistics*. Wiley, New York.
- Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Information Theory* 13, 126-127.
- Kullback, S. (1968). Probability densities with given marginal. *Ann. Math. Statist.* 39, 79-86.
- Kohn, M. (1969). *Class and Conformity*. Dorsey Press, Homewood, IL.
- Lauritzen, S., Speed, T., and Vijayan, K. (1978). Decomposable Graphs and Hypergraphs. Preprint 9, Institute of Mathematical Statistics, University of Copenhagen.
- Lauritzen, S. and Wermuth, N. (1984). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Stat.* 17, 31-57.
- Lawley, D., and Maxwell, A. (1971). *Factor Analysis as a Statistical Method*. Butterworth, London.
- Lazarsfeld, P., and Henry, N. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.
- Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Non-experimental Data*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, NY.
- Lee, S. (1985). Analysis of covariance and correlation structures. *Computational Statistics and Data Analysis* 2, 279-295.
- Lee, S. (1987). Model Equivalence in Covariance Structure Modeling. Ph.D. Thesis, Department of Psychology, Ohio State University.
- Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, MA.

- Lewis, D. (1973). Causation. *Journal of Philosophy* 70, 556-572.
- Lilienfeld, A. (1983). The Surgeon General's "Epidemiologic Criteria for Causality." A Criticism of Burch's Critique. 36, 837-845.
- Linthurst, R. A. (1979). Aeration, nitrogen, pH and salinity as factors affecting Spartina alterniflora growth and dieback. Ph.D. thesis, North Carolina State University.
- Lohmoller, J. (1989). Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg.
- Long, J. (1983a). Qualitative Applications in the Social Sciences. Vol. 33: Confirmatory Factory Analysis. Sage Publications, Beverly Hills, CA.
- Long, J. (1983b). Qualitative Applications in the Social Sciences. Vol. 34: Covariance Structure Models. Sage Publications, Beverly Hills, CA.
- Luijben, T., Boomsma, A., and Molenaar, I. (1986). Modification of factor analysis models in covariance structure analysis. A Monte Carlo study. On Model Uncertainty and its Statistical Implications. Lecture Notes in Economics and Mathematical Systems 307, Dijkstra, T. (ed.). Springer-Verlag, Berlin.
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin* 100, 107-120.
- Mackie, J. (1974). The Cement of the Universe. Oxford University Press, New York.
- Mallows, C. (1973). Some comments on C_p . *Technometrics* 15, 661-676.
- Mardia, K., Kent, J., and Bibby, J. (1979). Multivariate Analysis. Academic Press, New York.
- Maruyama, G. and McGarvey, B. (1980). Evaluating causal models: An application of maximum likelihood analysis of structural equations. *Psychological Bulletin* 87, 502-512.

- McPherson, J., Welch, S., and Clark, C. (1977). The stability and reliability of political efficacy: Using path analysis to test alternative models. *American Political Science Review* 71: 509-21.
- Miller, R. Jr. (1981). *Simultaneous Statistical Inference*, 2nd ed. McGraw-Hill, New York.
- Miller, W. and Stokes, D. (1963). Constituency influence in Congress. *American Political Science Review* 1963, 45-456.
- Miller, J., Slomczynski, K., and Schoenberg, R. (1981). Assessing comparability of measurement in cross-national research: Authoritarian-conservatism in different sociocultural settings. *Social Psychology Quarterly* 44, 178-191.
- Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *JASA* 83, 1023-1032.
- Mitchell, T. (1977). Version spaces: A candidate elimination approach to rule learning. *Proceedings Fifth International Joint Conference on AI. IJCAI - 77*, Pittsburgh, PA.
- Mosteller, F., and Tukey, J. (1977). *Data Analysis and Regression, A Second Course in Regression*. Addison-Wesley, Massachusetts.
- MRFIT Research Group (1982). Multiple risk factor intervention trial; risk factor changes and mortality results. *JAMA* 248, 1465-1477.
- Neopolitan, R. (1990) *Probabilistic Reasoning in Expert Systems*, Wiley, New York.
- Neyman, J. (1935) Statistical problems with agricultural experimentation. *J. Roy. Stat. Soc. Suppl.* 2, 107-180.
- Osherson, D., Stob, T., and Weinstein, S. (1986). *Systems That Learn*. MIT Press, Cambridge, MA.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan and Kaufman, San Mateo.

Pearl, J. and Dechter, R. (1989). Learning structure from data: A survey. *Proceedings COLT '89*, 30-244.

Pearl, J., Geiger, D. and Verma, T. (1990). The logic of influence diagrams. *Influence Diagrams, Belief Nets and Decision Analysis*. R. Oliver and J. Smith, editors. John Wiley & Sons Ltd.

Pearl, J. and Tarsi, M. (1986). Structuring causal trees. *Journal of Complexity* 2, 60-77.

Pearl, J. and Verma, T. (1987). The Logic of Representing Dependencies by Directed Graphs. Report CSD 870004, R-79-II, University of California at Los Angeles Cognitive Systems Laboratory.

Pearl, J. and Verma, T. (1990). A Formal Theory of Inductive Causation. Technical Report R-155, Cognitive Systems Laboratory, Computer Science Dept. UCLA.

Pearl, J. and Verma, T. (1991). A theory of inferred causation. *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Morgan Kaufmann, San Mateo, CA.

Pratt, J. and Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics* 39, 23-52.

Putnam, H. (1965). Trial and error predicates and a solution to a problem of Mostowski. *Journal of Symbolic Logic* 30, 49-57.

Rawlings, J. (1988). *Applied Regression Analysis*. Wadsworth, Belmont, CA.

Reichenbach, H. (1956). *The Direction of Time*. Univ. of California Press, Berkeley, CA.

Reiss, I., Banwart, A., and Forman, H. (1975). Premarital contraceptive usage: A study and some theoretical explorations. *J. Marriage and the Family* 37, 619-630.

Risannen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. Ser. B* 49, 223-239.

Rindfuss, R., Bumpass, L., and St. John, C. (1980). Education and fertility: Implications for the roles women occupy. *American Sociological Review* 45, 431-447.

- Rodgers, R. and Maranto, C. (1989). Causal models of publishing productivity in psychology. *Journal of Applied Psychology* 74, 636-649.
- Rose, G. Hamilton, P. Colwell, L., and Shipley, J. (1982). A randomised controlled trial of anti-smoking advice: 10-year results. *Journal of Epidemiology and Community Health* 36, 102-108.
- Rosenbaum, P. (1984). From association to causation in observational studies. *JASA* 79, 41-48.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 688-701.
- Rubin, D. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2, 1-26.
- Rubin, D. (1978). Bayesian inference for causal effects: The role of randomizations. *Ann. Stat.* 6, 34-58.
- Rubin, D. (1986). Comment: Which ifs have causal answers. *JASA* 81, 396.
- Salmon, W. (1980). Probabilistic causality. *Pacific Philosophical Quarterly* 61, 50-74.
- Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton Univ. Press, Princeton, NJ.
- Saris, W. and Stronkhorst, H. (1984). *Causal Modeling in Nonexperimental Research*. Sociometric Research Foundation, Amsterdam.
- Scheines, R. (1988). Automating creativity. *Aspects of Artificial Intelligence*, J. Fetzer (ed.). Kluwer, Boston.
- Scheines, R., and Spirtes, P. (1992). Finding latent variable models in large data bases. *International Journal of Intelligent Systems*, G. Piatetski-Shapiro, (ed.).

- Scheines, R., Spirtes, P., Glymour, G., and Sorensen, S. (1990). Causes of Success and Satisfaction Among Naval Recruiters. Report to the Navy Personnel Research Development Center, San Diego, CA
- Scheines, R., Spirtes, P., and Glymour, C. (1990). A qualitative approach to causal modeling. Qualitative Simulation Modeling and Analysis, Fishwick, P. and Luker, P. (eds.). Advances in Simulation 5, 72-97. Springer-Verlag, New York.
- Sclove, S. (undated). On Criteria for Choosing a Regression Equation for Prediction. Technical Report 28, Department of Statistics, Carnegie-Mellon University.
- Sewell, W. and Shah, V. (1968). Social class, parental encouragement, and educational aspirations. American Journal of Sociology. 73, 559-572.
- Simon, H. (1953). Causal ordering and identifiability. Studies in Econometric Methods. Hood and Koopmans (eds). 49-74. Wiley, NY.
- Simon, H. (1954). Spurious correlation: a causal interpretation. JASA. 49, 467-479.
- Simon, H. (1977). Models of Discovery. D. Reidel, Dordrecht.
- Simpson, C. (1951). The interpretation of interaction in contingency tables. J. Roy. Statist. Soc. Ser. B 13, 238-241.
- Sims, C. (1972). Money, income, and causality. American Economic Review 62, 540-552.
- Skyrms, B. (1980). Causal Necessity: A Pragmatic Investigation of the Necessity of Laws. Yale University Press, New Haven.
- Sober, E. (1987). The principle of the common cause. Probability and Causality, Fetzer, J. (ed.). D. Reidel, Dordrecht.
- Sorbom., D. (1975). Detection of correlated errors in longitudinal data. British Journal of Mathematical and Statistical Psychology 28, 138-151.
- Spearman, C. (1904). General intelligence objectively determined and measured. American Journal of Psychology 15, 201-293.

- Spence, M. (1973). Job market signalling. *Quarterly Journal of Economics* 87, 355-379.
- Spiegelhalter, D. (1986). Probabilistic reasoning in predictive expert systems. Uncertainty in Artificial Intelligence, Kanal, K. and Lemmer, J. (eds.). North-Holland, Amsterdam.
- Spiegelhalter, D., and Knell-Jones, R. (1984). Statistical and knowledge-based approaches to clinical decision-support systems. *J. Royal Statist. Soc. Ser. A* 147, 35-77.
- Spirites, P. (1989a). A Necessary and Sufficient Condition for Conditional Independencies to Imply a Vanishing Tetrad Difference. Technical Report CMU-LCL-89-3, Laboratory for Computational Linguistics, Carnegie Mellon University, Pgh, PA.
- Spirites, P. (1989b). Fast Geometrical Calculations of Overidentifying Constraints. Technical Report CMU-LCL-89-3, Laboratory for Computational Linguistics, Carnegie Mellon University, Pgh, PA.
- Spirites, P. (1992). Building causal graphs from statistical data in the presence of latent variables", forthcoming in Proceedings of the IX International Congress on Logic, Methodology, and the Philosophy of Science, B. Skyrms, ed., Uppsala, Sweden, 1991.
- Spirites, P. and Glymour, C. (1988). Latent variables, causal models and overidentifying constraints. *Journal of Econometrics* 39, 175-198.
- Spirites, P. and Glymour, C. (1990). Causal Structure Among Measured Variables Preserved with Unmeasured Variables. Technical Report CMU-LCL-90-5, Laboratory for Computational Linguistics, Carnegie Mellon University.
- Spirites, P., Glymour, C., and Scheines, R. (1990b). Causality from probability. Evolving Knowledge in Natural Science and Artificial Intelligence, Tiles, J. et. al. (eds.). Pitman, London, 181-199.
- Spirites P., Glymour C., and Scheines, R. (1990c). Causality from probability. Conference Proceedings: Advanced Computing for the Social Sciences, Williamsburgh, VA.
- Spirites, P., Glymour, C., and Scheines, R. (1991a). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, vol. 9, 1991, pp. 62-72.

- Spirites P., Glymour C., and Scheines, R. (1991b). From probability to causality. *Philosophical Studies*, 64, 1-36.
- Spirites P., Glymour C., Scheines, R., Meek, C., Fienberg, S., and Slate, E. (1992). Prediction and Experimental Design with Graphical Causal Models. Technical Report CMU-Phil-32, Philosophy Department, Carnegie-Mellon University.
- Spirites, P., Glymour, C., Scheines, R., and Sorensen, S. (1990). TETRAD Studies of Data for Naval Air Traffic Controller Trainees. Report to the Navy Personnel Research Development Center, San Diego, CA.
- Spirites, P., Scheines, R., and Glymour, C. (1990a). Simulation studies of the reliability of computer aided specification using the TETRAD II, EQS, and LISREL Programs. *Sociological Methods and Research*, 19, 3-66.
- Spirites P., and Verma, T. (1992). Equivalence of Causal Models with Latent Variables. Technical Report CMU-Phil-33, Philosophy Department, Carnegie-Mellon University.
- Spohn, W. (1983). Deterministic and probabilistic reasons and causes. *Methodology, Epistemology, and Philosophy of Science: Essays in Honour of Wolfgang Stegmuller on the Occasion of his 60th Birthday*, C. G. Hempel, H. Putnma, and W.K. Essler (eds.), D. Reidel, Dordrecht, Holland, 371-396.
- Spohn, W. (1990). Direct and indirect causes. *Topoi*, 9, 125-145.
- Spohn, W. (1991). On Reichenbach's principle of the common cause. *Proceedings of the First Pittsburgh-Konstanz Colloquium*, W. Salmon, G. Walters (eds.).
- Spohn, W. (1992). Causal laws are objectifications of inductive schemes. *Theory of Probability*, J. Dubucs (ed.), Kluwer, Dordrecht, Holland.
- Stein, C. (1960). Multiple regression. Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling, Olkin, I. (ed.). Stanford Univ. Press, Stanford, CA.
- Stetzl, I. (1986). Changing causal relationships without changing the fit: Some rules for generating equivalent LISREL models. *Multivariate Behavior Research* 21, 309-331.

- Suppes, P. (1970). A Probabilistic Theory of Causality. North-Holland, Amsterdam.
- Suppes, P. and Zanotti, M. (1981). When are probabilistic explanations possible. *Synthese* 48, 191-199.
- Surgeon General of the United States (1964). Smoking and Health. U.S. Government Printing Office.
- Surgeon General of the United States (1979). Smoking and Health. U.S. Government Printing Office.
- Swamy, P. (1971). Statistical Inference in Random Coefficient Regression Models. Springer-Verlag, Berlin.
- Thomson, G. (1916). A hierarchy without a general factor. *British Journal of Psychology* 8, 271-281.
- Thomson, G. (1935). On complete families of correlation coefficients and their tendency to zero tetrad-differences: Including a statement of the sampling theory of abilities. *British Journal of Psychology* 26, 63-92.
- Thurstone, L. (1935). The Vectors of Mind. Univ. of Chicago Press, Chicago.
- Timberlake, M. and Williams, K. (1984). Dependence, political exclusion, and government repression: Some cross-national evidence. *American Sociological Review* 49, 141-146.
- Verma, T. (1987). Causal networks: semantics and expressiveness. Technical Report R-65-I, Cognitive Systems Laboratory, University of California, Los Angeles.
- Verma, T. and Pearl, J. (1990a). On Equivalence of Causal Models. Technical Report R-150, Cognitive Systems Laboratory, University of California, Los Angeles.
- Verma, T. and Pearl, J. (1990b). Equivalence and synthesis of causal models. Proc. Sixth Conference on Uncertainty in AI. Association for Uncertainty in AI, Inc., Mountain View, CA, 220-227.

- Verma, T. and Pearl J. (1990c). Causal networks: semantics and expressiveness. *Uncertainty in Artificial Intelligence 4*, R. Shacter, T. Levitt, L. Kanal, J. Lemmer (eds.) Elsevier Science Publishers, North-Holland.
- Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. Technical Report R-150, Cognitive Systems Laboratory, University of California, Los Angeles.
- Wallace, C. and Freeman, P (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. Ser. B* 49, 240-265.
- Weisberg, S. (1985). Applied Linear Regression, 2nd ed. Wiley, New York.
- Wermuth, N. (1976). Model search among multiplicative models. *Biometrika* 32, 253-363.
- Wermuth, N. (1980). Linear recursive equations, covariance selection and path analysis. *JASA* 75, 963-972.
- Wermuth, N. and Lauritzen, S. (1983). Graphical and recursive models for contingency tables. *Biometrika* 72, 537-552.
- Wermuth, N. and Lauritzen, S. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *J. Roy. Statist. Soc. Ser. B* 52, 21-50.
- Wheaton, B., Muthen, B., Alwin, D., and Summers, G. (1977). Assessing Reliability and Stability in Panel Models. *Sociological Methodology 1977*, Heise, D. (ed.). Jossey-Bass, San Francisco.
- Whitney, H. (1957). Elementary structures of real algebraic varieties. *Ann. Math.* 66, 545-556.
- Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Wiley, New York.
- Wise, D. (1975). Academic achievement and job performance. *American Economic Review* 65, 350-366.
- Wishart, J. (1928). Sampling errors in the theory of two factors. *British Journal of Psychology*, Vol. 19, pp. 180-187.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Stat.* 5, 161-215.

Younger, M. (1978). *Handbook for Linear Regression*. Wadsworth, California.

Yule, G. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* 2, 121-134.

Yule, G. (1926). Why do we sometimes get nonsensical relations between time-series? -- A study in sampling and the nature of time series. *Journal of the Royal Statistical Society*, 89, 1-64.

Index

- 2x2 Foursomes 312-314
- 3x1 Foursomes 314
- Acceptable Ordering 221
- Acyclic Path 29
- Adjacency 29
- Adjacent 28
- Aitkin, M. 105
- Almost Pure Latent Variable Graph 309
- Almost Pure Measurement Model 309, 473
- Ancestor 31
- Anderson, T. 128
- Asmussen, S. 4
- Asymptotically Reliable iv
- Axioms
 - Causal Markov Condition 54-55, 57-64
 - Statement 54
 - Causal Minimality Condition 55-56
 - Statement 55
 - Faithfulness 35-36
 - Faithfulness Condition
 - Statement 56
 - Markov 33
- Basmann, R. 5, 101
- Bayesianism 23, 68, 70, 277-286, 323
 - Search Strategies 109-111
- Bell, J. 63
- Bentler, P. 104, 140-142, 323
- Bibby, J. 197
- Birch, M. 105, 107
- Bishop, E. 39
- Bishop, Y. 4, 105
- Blalock, H. i, 6, 75, 86
- Blau, P. 142-144
- Blyth, C. 267
- Bollen, K. i, 107, 195, 325
- Bootstrap Procedures 365
- Bouck, L. ii
- Bowden, R. 2
- Breslow, N. 164
- Brownlee, K. 17, 291, 297-298, 302
- Bumpass, L. 139-140
- Bunker, J. 3
- Burch, P. 291, 298-300, 301, 302
- Byron, R. 104
- Califano, J. 291
- Callahan, J. ii
- Cambell, D. 338
- Carlin, J. i, 40, 162
- Cartwright, N. 96
- Causal Chain 44
- Causal Graph 47
- Causal Inference Algorithm 181-184
 - Statement 183
- Causal Markov Condition 54-55, 57-64
 - Statement 54
- Causal Mediary 43
- Causal Minimality Condition 55-56
 - Statement 55
- Causal Representation Convention 47
- Causal Structure 45
 - Deterministic 49
 - Generates Probability Distribution 50, 54
 - Generation 45
 - Genuinely Indeterministic 53
 - Indeterministic 51
 - Isomorphism 45
 - Linear Deterministic 49
 - Linear Pseudo-indeterministic 52
- Causal System

- Causally Sufficient 45
- Causally Connected 45
- Causally Sufficient 45
- Causation
 - Reciprocal 354
- Cause
 - Boolean 43
 - Common 44, 47
 - Direct 43
 - Indirect 44
 - Representation 47
 - Scaled Variable 44
- Cederlof, R. 298
- Child 28
- Children 28
- Choke Point 196
- Chordless cycle 361
- Christensen, R. ii, 23, 105, 150
- CI (see Causal Inference Algorithm)
- CI Partially Oriented Inducing Path Graph 190
- Clique 31
 - Maximal 31
- Coleman, J. 147-149
- Collapsibility 3-4
- Collider 31, 181
- Combined Graph 78
- Common Cause 44, 47
- Common Cause Impure 309
- Complete Graph 30
- Concatenation 29
- Conditional 41-42
- Conditional Independence 32
- Conditional Independence Tests 366
- Connected Graph 30
- Constraints
 - Non-independence 191-193
- Cooper, G. i, 11, 110, 146, 323
- Cornfield, J. 291, 293-296
- Costner, H. i, 86, 338
- Covariance Structures 104
- Crawford, S. 108, 125
- Cross-Construct Foursomes 314
- Cross-Construct Impure 309
- Cross-Validation 365
- Cycle
 - Chordless 361
- Cyclic 29
- Cyclic Directed Graph 354-360
- d-connection 36, 74, 83
- D-SEP 176
- d-separation 36, 73-75, 82, 83, 113
 - Cyclic Graph 359
 - Faithful 113
 - Order 316
- DAG-Isomorph 36
- Darroch, J. 39
- Data Sets
 - Abortion Opinions 150
 - AFQT 243-244
 - Alarm Network 11, 145-147
 - American Occupational Structure 142-145
 - College Plans 149-150
 - Education and Fertility 139-140
 - Female Orgasm 140-142
 - Leading Crowd 147-149
 - Mathematical Marks 197-200
 - Political Exclusion 248-250
 - Publishing Productivity 13-14, 133-138, 166
 - Rat Liver 252-257
 - Simulation Studies 161, 250-251, 320, 332, 351

- Results 154-161
Sample Generation
Spartina grass 244-248
Daughter 29
Davis, W. 61
Day, N. 164
Definite d-connection 189
Definite Discriminating Path 181
Definite Non-Collider 180, 224, 319
Definite-Non-Descendant 221
Definite-SP 223
Degree 29
Dempster, A. 104
Descendant 31
Descendants 33
Det 83
Deterministic 83
Deterministic Causal Structure 49
Deterministic Graph 38
Deterministic System 38
Deterministic Variable 82
Direct Cause 43
Direct Manipulation 78
Directed Acyclic Graph 30, 32
 DAG-Isomorph 36
 I-map 33
 Linear Implication 196
 Minimal I-map 34
 Perfect Map 36
 Represents Probability Distribution 34
Directed Edge 28
 Head 30
 Tail 30
Directed Graph 25, 26, 28
 Acyclic 30
Directed Independence Graph 34-35, 111
Directed Path 25, 29
Discovery 103-104
Discovery Problems 103-104
Doll, R. 291
Druzzel, M. ii
Duncan, O. 142-144
Earman, J. ii
Edge 28, 29
 Directed 28
 Edge-end 28
 Endpoint 28
 Into 28
 Out of 28
 Edge-end 28
Edwards, A. 128
Edwards, D. 4, 109
Empty Path 29
Endogenous Variable 49
Endpoint 28
EQS 104, 109, 138, 143, 144, 323
 Automatic Model Modification 329,
 332
 Boldness 346
 Lagrange Multiplier Statistic 331
 Reliability 346
Equiv 177
Error
 Edge Direction of Commision 146
 Edge Direction of Omission 146
 Edge Existence of Commision 146
 Edge Existence of Omission 146
Error Variable 37
Estimation 132
 Multinomial Distributions 132
Ethics
 Experimental Design 286-290
Exogenous Variable 49, 78
Expanded Graph 37

- Experimental design 17, 23, 277-286
 Ethical 286-290
 Prospective 274
 Randomized 261
 Retrospective 274
 F.I. (see Faithful Indistinguishability)
 Factor Analysis 107, 198
 Faithful Indistinguishability 89-90, 361-363
 Faithfulness 36
 Faithfulness Condition 9, 22, 35, 66
 Statement 56
 Fast Causal Inference Algorithm 184
 Statement 188
 FCI (see Fast Causal Inference Algorithm)
 FCI Partially Oriented Inducing Path
 Graph 190
 Feedback 354-360
 Fienberg, S. i, 4, 39, 86, 105, 106, 147, 237
 Fisher, F. 101
 Fisher, R. iii, 17, 259, 260, 291, 292-293, 302
 Forrest, W. 3
 Foursomes
 Cross-Construct 310
 Intra-Construct 310
 Fox, J. 238
 Freedman, D. 142, 365
 Friberg, L. 298
 Friedman, M. 126
 Functional Determination 82
 Fung, R. 108, 125
 Geiger, D. i, 48, 71, 86, 224
 Generation 45, 50, 54
 Genuinely Indeterministic Causal Structure 53
 Glymour, C. 86, 200
 Gold, M. 162
 Granger Causality 363-365
 Graph 25, 28
 Adjacent 28
 Causal 47
 Clique 31
 Combined 78
 Complete 30
 Connected 30
 Cyclic Directed 354-360
 d-connection 74, 83
 d-separation 73, 83
 Deterministic 38
 Directed 25, 26
 Directed Acyclic 30, 32
 Directed Independence 34-35, 111
 Edge 28
 Expanded 37
 Faithful Representation of D-separations 113
 Inducing Path 25, 174-177
 Manipulated 78
 Over 28
 Parallel Embedding 94
 Partially Oriented Inducing Path 25, 177, 181
 Pattern 27, 90, 113
 Subgraph 30
 Trek 97, 196
 Trek Sum 97
 Triangle 31
 Underlying 37
 Undirected 25
 Undirected Independence 37-38, 108
 Unmanipulated 78
 Haenszel, W. 291

- Hammond, E. 291
Hausman, D. ii
Havranek, T. 109
Head 30
Heckerman, D. 48
Heise, D. 75, 355
Henry, N. 86
Herskovits, E. 11, 110, 146, 323
Hierarchical Models 106
Hill, A. 291
Hill, J. ii
Holland, P. i, 4, 8-9, 39, 105, 164
Howson, C 260
I-map 33
IG Algorithm 124-125
Implication
 Linear 37
Impure Indicators
 Common Cause 309
 Cross-Construct 309
 Intra-Construct 308, 311
 Latent-Measured 308
Indegree 29
Independence 32
Indeterministic Causal Structure 51
Indeterministic System 38
Indirect Cause 44
Indistinguishability
 Faithful 361-363
Indistinguishable
 Linearly Faithful 361
Inducing Path Graph 25, 26, 174-177
D-separability 176
SP 223
Informative Parents 208
Informative Variables 208
Instrumental Variable 101
Intersect 29
Into 28, 181
Intra-Construct Foursomes 310-311
Intra-Construct Impure 308, 311
Invariance 208, 461, 486
IP 208
Isomorphic Causal Structures 45
IV 208
Jackknife Procedures 365
Jeffreys, H. 162
Joint Independence 32
Joreskog, K. 104, 107, 132
Kadane, J. ii, 261, 277-286
Kelly, K. 86, 200
Kelly, T. 164
Kendall, M. 64-66, 67
Kent, J. 197
Kiiveri, H. i, 7, 39, 132, 162
Klein, L. 163
Kleinbaum, D. 164
Klepper, S. ii, 126
Klir, G. 105
Kohn, M. 307, 333
Kullback, S. 106
Kupper, L. 164
Lagrange Multiplier Statistic 331
Latent Variable Graph
 Almost Pure 309
Latent Variable Models
 Cross-Construct Foursome 310
 Intra-Construct Foursome 310
Latent Variables 21, 306
 Causal structure among 19
Latent-Measured Impure 308
Lauritzen, S. i, iv, 7, 37, 39, 111
Lazarsfeld, P. 86
Leamer, E. 109

- Lee, S. 101
- Lilienfeld, A. 291, 293, 300-301, 302
- Linear Deterministic Causal Structure 49
- Linear Faithfulness 74
- Linear Implication 37, 196
- Linear Model 196
- Linear Pseudo-Indeterministic 39
- Linear Pseudo-indeterministic Causal Structure 52
- Linear Pseudo-Indeterministic Models 307
- Linear Regression 101, 108
- Linear Representation 36
- Linear Statistical Indistinguishability 94, 99
- Linearly Faithful Indistinguishable 361
- Linthurst, R. ii, 16, 244-248
- LISREL 104, 107, 109, 143, 323
 - Automatic Model Modification 329-332
 - Boldness 346
 - Modification Indices 331
 - Reliability 346
- Log-linear models 4, 104, 105-107
 - Hierarchical Models 106, 108
 - Representing Colliders 106
- Lohmöller, J. 132
- Lundman, T. 298
- Lung Cancer 291-302
- Manipulated 78
- Manipulated Graph 78
- Manipulation 7-9, 22, 78
 - Direct 78
 - Invariance 208, 461, 486
- Manipulation Theorem 9, 75-81
 - Statement 79
- Maranto, C. 13, 133-138, 166
- Mardia, K. 197
- Marginal 32
- Markov Condition 7, 8, 9, 22, 33, 359-360
- Maruyama, G. 338
- Mason's Theorem 355
- Mason, S. 355
- Maximal Clique 31
- Maximally Oriented Partially Oriented Inducing Path Graph 180
- Maximum Likelihood Estimation Fitting Function 330
- McGarvey, B. 338
- McPherson J., 333
- Measurement Model
 - Almost Pure 309, 473
- Meek, C. i, 86
- Meyer, M. ii
- MIMBuild Algorithm
 - Complexity 320
 - Reliability 320
 - Simulation Study Results 320
- Minimal I-map 34
- MINITAB 240
- Mitchell, T. 109
- Mixture 58
- Model selection 4
- Model Specification 365
- Modification Indices 331
- Modified PC Algorithm 165
 - Statement 166
- Morgenstern, H. 164
- Mosteller, F. 3, 202, 239
- Multiple Risk Factor Intervention Trial Research Group 301
- National Halothane Study 3
- Navidi, W. 365
- ND 208
- Neyman, J. 7, 291

- No Descendants 208
Node
 Visit 327
Noncollider 31
Notational Conventions viii
Observed 173
Order
 d-separability 316
Osherson, D. 162
Out of 28, 181
Outdegree 29
Over 28
Papineau, D. ii
Parallel Embedding 94
Parameter Estimation 365
Parent 25, 28, 181
Parents 28
Partially Oriented Inducing Path Graph 25, 26, 177, 181
 Acceptable Ordering 221
 CI 190
 Collider 181
 Definite d-connection 189
 Definite Discriminating Path 181
 Definite Non-Collider 180, 224
 Definite-SP 223
 FCI 190
 Into 181
 Maximally Oriented 180
 Out Of 181
 Parent 181
 Possible-D-SEP 187
 Possible-SP 223
 Possibly d-connecting 224
 Possibly-IP 224
 Possibly-IV 224
 Semi-Directed Path 30, 190
Path
 Acyclic 29
 Adjacency 29
 Collider on 31
 Concatenation of 29
 Cyclic 29
 Definite Discriminating 181
 Directed 25, 29
 Empty 29
 Intersect 29
 Into 29
 Noncollider on 31
 Out of 29
 Point of Intersection 29
 Undirected 25, 29
 Unshielded Collider on 31
Pattern
 Definite Non-Collider 319
 Output 146
 True 146
Pattern Graph 27, 90, 113
PC Algorithm 116-122
 Applications 132
 Discrete Distributions 147
 Linear Structural Equation Models 133
 Applied to Latent Variables 316
 Complexity 119-120
 Stability 120-122
 Statement 117, 118
 PC* Algorithm 122-123
 Heuristics 123-124
 Statement 123
 Pearl, J. i, iv, 36, 40, 68, 71, 86, 101, 124, 126, 162, 200, 224, 258, 260
 Pearson, K. 21
 Peeler, W. 140-142

- Perfect Map 36
- Peters, S. 365
- P_{Man} 78
- Point of Intersection 29
- Policy Variable 78
- Political Exclusion 248-250
- Population
 - Manipulated 76
 - Unmanipulated 76
- Possible-D-SEP 187
- Possible-SP 223
- Possibly D-connecting 224
- Possibly-IP 224
- Possibly-IV 224
- Power 254
- Pratt, J. i, iv, 8-9, 22, 202, 203-212, 238
- Predictable 213
- Prediction Algorithm 227
 - Examples 227-237
 - Statement 221-222
- Probability 31
 - Conditional Independence 32
 - Independence 32
 - Interpretations 5
- Probability Distribution
 - Generated by Causal Structure 50, 54
- Prospective Design 274
- Pseudo-Indeterministic Causal Structure 52
- Pseudo-Indeterministic System 38, 39
- Pseudocorrelation Matrix 340
- P_{Unman} 78
- Putnam, H. 162
- R.S.I. (see Rigid Statistical Indistinguishability)
- Rawlings, J. 2, 16, 244-248
- Reconstructability Analysis 105
- Recursive Diagram 111
- Regression 1-2, 238-241, 363
 - Logistic 105
 - Selection of Regressors 14-17
 - Stepwise 105
- Reichenbach, H. 6
- Representation
 - Linear 36
- Represents 34
- Retrospective Design 274
- Rigid Statistical Indistinguishability 94
- Rindfuss, R. 139-140
- Rodgers, R. 13, 133-138, 166
- Rose, G. 300
- Rosenbaum, P. 164
- Rubin, D. i, iv, vi, 8, 22, 202, 203-212, 237
- S.S.I. (see Strong Statistical Indistinguishability)
- Salmon, W. 62-63
- Sampling 272-276
- SAS 245
- Scheines, R. 86, 200
- Schlaifer, R. i, iv, 8-9, 22, 202, 203-212, 238
- Schoenberg, R. 338
- Scott, D. ii
- Search Algorithms
 - Incorporating Background Knowledge 127
 - Probabilities of Error 130-132
 - Statistical Decisions 128
 - Variable Selections 125
- Search Strategies 104-111
- Sedransk, N. 277-286
- Seidenfeld, T. ii, 261, 277-286
- Semi-Directed Path 30, 190
- Seneca, E. ii

- Sewell, W. 149-150
SGS Algorithm 114-116
 Complexity 115
 Stability 115
 Statement 114
Shah, V. 149-150
Shimkin, M. 291
Simon, H. i, 6, 75, 86
Simpson's Paradox 64-68, 92, 267, 273
Simpson, C. 64, 67-68, 267
Simultaneous Equation Model 101
Skyrms, B. ii
Slate, E. i, 86
Slezak, P. ii
Smoking 291-302
 Surgeon General's Report 291, 296-297, 298, 302
Sober, E. 63
Sorbom, D. 104, 107
Sorensen, S. ii
Source 34
SP 223
Spearman, C. i, 86, 162, 164, 200
Specification Searches 4
 Error Probabilities 252-257
Speed, T. i, iv, 7, 39, 132, 162
Spirtes, P. 86, 200
St. John, C. 139-140
Stationary 363
Statistical Indistinguishability 21, 87
 Faithful 89-90
 Linear 99
 Rigid 94
 Strong 88
 Weak 90-93
Statistical Tests
 Degrees of Freedom 129
 G² 129
 Vanishing Partial Correlation 128
 Vanishing Tetrad Difference 325
 X² 129
Stepwise Regression 105
Stetzl, I. 5, 101
Stob, T. 162
Strong Statistical Indistinguishability 88
Structural Equation Models 307
 Indistinguishability 5
 Measurement Model 307
 Structural Model 307
Subgraph 30
Suppes, P. 6, 101, 164
Surgeon General's Report on Smoking and Health 291
Tail 30
Tetrad Differences 325
 Vanishing 194
 Variance of Sampling Distribution 325
TETRAD II 10, 127, 132, 138, 144, 146
 Search Procedure 324
 Boldness 346
 Implied_H 326
 Implied_H 326
 Limitations 351
 Reliability 346
 Scoring Principles 324-326
 T-maxscore 327
 Tetrad-Score 326
 Weight 326
 Tetrad Representation Theorem 22, 196-197
 Statement 196, 315
Thomson, G. 200
Thurstone, L. 107, 162, 198
Timberlake, M. 248-250

- Time Series 356-358, 363-365
 Trek 34, 97, 196
 Source 34
 Trek Sum 97
 Triangle 31
 Tukey, J. 202, 239
 Turkington, D. 2
 Type 1 Error 254
 Type 2 Error 254
 Underdetermination 91
 Underlying Graph 37
 Undirected Graph 25, 28
 Undirected Independence Graph 37-38, 108
 Algorithm 124
 Undirected Path 25, 29
 Unmanipulated Graph 78
 Unshielded Collider 31
 Urbach, P. 260
 Vandam, L. 3
 Vanishing Tetrad Differences 194, 310, 314, 315
 Linear Implication 196
 Variable
 Determined 81
 Deterministic 82
 Endogenous 49
 Error 37
 Exogenous 49
 Functionally Determined 82
 Instrumental 101
 Observed 173
 Policy 78
 Redefining 99-101
 Variable Aggregation 125
 Variable Redefinition 99-101
 Variable Selection 271-272
 Verma, T. i, 71, 86, 101, 124, 162, 191, 200, 224, 258, 260
 Version Spaces 109
 Vertex
 Ancestor 31
 Child 28
 Children 28
 Daughter 29
 Degree 29
 Descendant 31
 Descendants 33
 Indegree 29
 Outdegree 29
 Parent 28
 Parents 28
 Vijayan, K. 39
 Visit 327
 W.F.I. (see Weak Faithful Indistinguishability)
 W.S.I. (see Weak Statistical Indistinguishability)
 Weak Faithful Indistinguishability 91
 Weak Statistical Indistinguishability 90-93
 Weinstein, S. 162
 Weisberg, S. 252-257
 Wermuth, N. i, iv, 7, 37, 40, 104, 111
 Wermuth-Lauritzen Algorithm 111-112
 Wheaton, B. 333
 Whittaker, J. 4, 105, 111, 125, 197
 Williams, K. 248-250
 Wishart, J. 325
 Worrall, J. ii
 Wright, S. 6, 75, 86
 Wynder, E. 291
 Younger, M. 2
 Yule, G. 3, 21
 Zanotti, M. 101, 164

- Vol. 1: R.A. Fisher: An Appreciation. Edited by S.E. Fienberg and D.V. Hinkley. XI, 208 pages, 1980.
- Vol. 2: Mathematical Statistics and Probability Theory. Proceedings 1978. Edited by W. Klonecki, A. Kozek, and J. Rosinski. XXIV, 373 pages, 1980.
- Vol. 3: B.D. Spencer, Benefit-Cost Analysis of Data Used to Allocate Funds. VIII, 296 pages, 1980.
- Vol. 4: E.A. van Doorn, Stochastic Monotonicity and Queueing Applications of Birth-Death Processes. VI, 118 pages, 1981.
- Vol. 5: T. Rolski, Stationary Random Processes Associated with Point Processes. VI, 139 pages, 1981.
- Vol. 6: S.S. Gupta and D.-Y. Huang, Multiple Statistical Decision Theory: Recent Developments. VIII, 104 pages, 1981.
- Vol. 7: M. Akahira and K. Takeuchi, Asymptotic Efficiency of Statistical Estimators. VIII, 242 pages, 1981.
- Vol. 8: The First Pannonian Symposium on Mathematical Statistics. Edited by P. Révész, L. Schmetterer, and V.M. Zolotarev. VI, 308 pages, 1981.
- Vol. 9: B. Jørgensen, Statistical Properties of the Generalized Inverse Gaussian Distribution. VI, 188 pages, 1981.
- Vol. 10: A.A. McIntosh, Fitting Linear Models: An Application of Conjugate Gradient Algorithms. VI, 200 pages, 1982.
- Vol. 11: D.F. Nicholls and B.G. Quinn, Random Coefficient Autoregressive Models: An Introduction. V, 154 pages, 1982.
- Vol. 12: M. Jacobsen, Statistical Analysis of Counting Processes. VII, 226 pages, 1982.
- Vol. 13: J. Pfanzagl (with the assistance of W. Wefelmeyer), Contributions to a General Asymptotic Statistical Theory. VII, 315 pages, 1982.
- Vol. 14: GLIM82: Proceedings of the International Conference on Generalised Linear Models. Edited by R. Gilchrist. V, 188 pages, 1982.
- Vol. 15: K.R.W. Brewer and M. Hanif, Sampling with Unequal Probabilities. IX, 164 pages, 1983.
- Vol. 16: Specifying Statistical Models: From Parametric to Non-Parametric, Using Bayesian or Non-Bayesian Approaches. Edited by J.P. Florens, M. Mouchart, J.P. Raoult, L. Simar, and A.F.M. Smith. XI, 204 pages, 1983.
- Vol. 17: I.V. Basawa and D.J. Scott, Asymptotic Optimal Inference for Non-Ergodic Models. IX, 170 pages, 1983.
- Vol. 18: W. Britton, Conjugate Duality and the Exponential Fourier Spectrum. V, 226 pages, 1983.
- Vol. 19: L. Fehmholz, von Mises Calculus For Statistical Functionals. VIII, 124 pages, 1983.
- Vol. 20: Mathematical Learning Models — Theory and Algorithms: Proceedings of a Conference. Edited by U. Herkenrath, D. Kalin, W. Vogel. XIV, 226 pages, 1983.
- Vol. 21: H. Tong, Threshold Models in Non-linear Time Series Analysis. X, 323 pages, 1983.
- Vol. 22: S. Johansen, Functional Relations, Random Coefficients and Nonlinear Regression with Application to Kinetic Data. VIII, 126 pages, 1984.
- Vol. 23: D.G. Saphire, Estimation of Victimization Prevalence Using Data from the National Crime Survey. V, 165 pages, 1984.
- Vol. 24: T.S. Rao, M.M. Gabr, An Introduction to Bispectral Analysis and Bilinear Time Series Models. VIII, 280 pages, 1984.
- Vol. 25: Time Series Analysis of Irregularly Observed Data. Proceedings, 1983. Edited by E. Parzen. VII, 363 pages, 1984.
- Vol. 26: Robust and Nonlinear Time Series Analysis. Proceedings, 1983. Edited by J. Franke, W. Härdle and D. Martin. IX, 286 pages, 1984.
- Vol. 27: A. Janssen, H. Milbrodt, H. Strasser, Infinitely Divisible Statistical Experiments. VI, 163 pages, 1985.
- Vol. 28: S. Amari, Differential-Geometrical Methods in Statistics. V, 290 pages, 1985.
- Vol. 29: Statistics in Ornithology. Edited by B.J.T. Morgan and P.M. North. XXV, 418 pages, 1985.
- Vol. 30: J. Grandell, Stochastic Models of Air Pollutant Concentration. V, 110 pages, 1985.
- Vol. 31: J. Pfanzagl, Asymptotic Expansions for General Statistical Models. VII, 505 pages, 1985.
- Vol. 32: Generalized Linear Models. Proceedings, 1985. Edited by R. Gilchrist, B. Francis and J. Whittaker. VI, 178 pages, 1985.
- Vol. 33: M. Csörgő, S. Csörgő, L. Horváth, An Asymptotic Theory for Empirical Reliability and Concentration Processes. V, 171 pages, 1986.
- Vol. 34: D.E. Critchlow, Metric Methods for Analyzing Partially Ranked Data. X, 216 pages, 1985.
- Vol. 35: Linear Statistical Inference. Proceedings, 1984. Edited by T. Calinski and W. Klonecki. VI, 318 pages, 1985.
- Vol. 36: B. Matérn, Spatial Variation. Second Edition. 151 pages, 1986.
- Vol. 37: Advances in Order Restricted Statistical Inference. Proceedings, 1985. Edited by R. Dykstra, T. Robertson and F.T. Wright. VIII, 295 pages, 1986.
- Vol. 38: Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits. Edited by R.W. Pearson and R.F. Boruch. V, 129 pages, 1986.
- Vol. 39: J.D. Malley, Optimal Unbiased Estimation of Variance Components. IX, 146 pages, 1986.
- Vol. 40: H.R. Lerche, Boundary Crossing of Brownian Motion. V, 142 pages, 1986.
- Vol. 41: F. Baccelli, P. Brémaud, Palm Probabilities and Stationary Queues. VII, 106 pages, 1987.
- Vol. 42: S. Kullback, J.C. Keegel, J.H. Kullback, Topics in Statistical Information Theory. IX, 158 pages, 1987.
- Vol. 43: B.C. Arnold, Majorization and the Lorenz Order: A Brief Introduction. VI, 122 pages, 1987.
- Vol. 44: D.L. McLeish, Christopher G. Small, The Theory and Applications of Statistical Inference Functions. VI, 124 pages, 1987.
- Vol. 45: J.K. Ghosh (Ed.), Statistical Information and Likelihood. 384 pages, 1988.

- Vol. 46: H.-G. Müller, Nonparametric Regression Analysis of Longitudinal Data. VI, 199 pages, 1988.
- Vol. 47: A.J. Getson, F.C. Hsuan, {2}-Inverses and Their Statistical Application. VIII, 110 pages, 1988.
- Vol. 48: G.L. Brethorst, Bayesian Spectrum Analysis and Parameter Estimation. XII, 209 pages, 1988.
- Vol. 49: S.L. Lauritzen, Extremal Families and Systems of Sufficient Statistics. XV, 268 pages, 1988.
- Vol. 50: O.E. Barndorff-Nielsen, Parametric Statistical Models and Likelihood. VII, 276 pages, 1988.
- Vol. 51: J. Hüsler, R.-D. Reiss (Eds.), Extreme Value Theory. Proceedings, 1987. X, 279 pages, 1989.
- Vol. 52: P.K. Goel, T. Ramalingam, The Matching Methodology: Some Statistical Properties. VIII, 152 pages, 1989.
- Vol. 53: B.C. Arnold, N. Balakrishnan, Relations, Bounds and Approximations for Order Statistics. IX, 173 pages, 1989.
- Vol. 54: K.R. Shah, B.K. Sinha, Theory of Optimal Designs. VIII, 171 pages, 1989.
- Vol. 55: L. McDonald, B. Manly, J. Lockwood, J. Logan (Eds.), Estimation and Analysis of Insect Populations. Proceedings, 1988. XIV, 492 pages, 1989.
- Vol. 56: J.K. Lindsey, The Analysis of Categorical Data Using GLIM. V, 168 pages, 1989.
- Vol. 57: A. Decarli, B.J. Francis, R. Gilchrist, G.U.H. Seeber (Eds.), Statistical Modelling. Proceedings, 1989. IX, 343 pages, 1989.
- Vol. 58: O.E. Barndorff-Nielsen, P. Blæsild, P.S. Eriksen, Decomposition and Invariance of Measures, and Statistical Transformation Models. V, 147 pages, 1989.
- Vol. 59: S. Gupta, R. Mukerjee, A Calculus for Factorial Arrangements. VI, 126 pages, 1989.
- Vol. 60: L. Györfi, W. Härdle, P. Sarda, Ph. Vieu, Nonparametric Curve Estimation from Time Series. VIII, 153 pages, 1989.
- Vol. 61: J. Breckling, The Analysis of Directional Time Series: Applications to Wind Speed and Direction. VIII, 238 pages, 1989.
- Vol. 62: J.C. Akkerboom, Testing Problems with Linear or Angular Inequality Constraints. XII, 291 pages, 1990.
- Vol. 63: J. Pfanzagl, Estimation in Semiparametric Models: Some Recent Developments. III, 112 pages, 1990.
- Vol. 64: S. Gabler, Minimax Solutions in Sampling from Finite Populations. V, 132 pages, 1990.
- Vol. 65: A. Janssen, D.M. Mason, Non-Standard Rank Tests. VI, 252 pages, 1990.
- Vol. 66: T. Wright, Exact Confidence Bounds when Sampling from Small Finite Universes. XVI, 431 pages, 1991.
- Vol. 67: M.A. Tanner, Tools for Statistical Inference: Observed Data and Data Augmentation Methods. VI, 110 pages, 1991.
- Vol. 68: M. Taniguchi, Higher Order Asymptotic Theory for Time Series Analysis. VIII, 160 pages, 1991.
- Vol. 69: N.J.D. Nagelkerke, Maximum Likelihood Estimation of Functional Relationships. V, 110 pages, 1992.
- Vol. 70: K. Iida, Studies on the Optimal Search Plan. VIII, 130 pages, 1992.
- Vol. 71: E.M.R.A. Engel, A Road to Randomness in Physical Systems. IX, 155 pages, 1992.
- Vol. 72: J.K. Lindsey, The Analysis of Stochastic Processes using GLIM. VI, 294 pages, 1992.
- Vol. 73: B.C. Arnold, E. Castillo, J.-M. Sarabia, Conditionally Specified Distributions. XIII, 151 pages, 1992.
- Vol. 74: P. Barone, A. Frigessi, M. Piccioni, Stochastic Models, Statistical Methods, and Algorithms in Image Analysis. VI, 258 pages, 1992.
- Vol. 75: P.K. Goel, N.S. Iyengar (Eds.), Bayesian Analysis in Statistics and Econometrics. XI, 410 pages, 1992.
- Vol. 76: L. Bondesson, Generalized Gamma Convolutions and Related Classes of Distributions and Densities. VIII, 173 pages, 1992.
- Vol. 77: E. Mammen, When Does Bootstrap Work? Asymptotic Results and Simulations. VI, 196 pages, 1992.
- Vol. 78: L. Fahrmeir, B. Francis, R. Gilchrist, G. Tutz (Eds.), Advances in GLIM and Statistical Modelling: Proceedings of the GLIM92 Conference and the 7th International Workshop on Statistical Modelling, Munich, 13-17 July 1992. IX, 225 pages, 1992.
- Vol. 79: N. Schmitz, Optimal Sequentially Planned Decision Procedures. XII, 209 pages, 1992.
- Vol. 80: M. Fligner, J. Verducci (Eds.), Probability Models and Statistical Analyses for Ranking Data. XXII, 306 pages, 1992.
- Vol. 81: P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search. XXIII, 526 pages, 1993.

General Remarks

Lecture Notes are printed by photo-offset from the master-copy delivered in camera-ready form by the authors of monographs, resp. editors of proceedings volumes. For this purpose Springer-Verlag provides technical instructions for the preparation of manuscripts. Volume editors are requested to distribute these to all contributing authors of proceedings volumes. Some homogeneity in the presentation of the contributions in a multi-author volume is desirable.

Careful preparation of manuscripts will help keep production time short and ensure a satisfactory appearance of the finished book. The actual production of a Lecture Notes volume normally takes approximately 8 weeks.

For monograph manuscripts typed or typeset according to our instructions, Springer-Verlag can, if necessary, contribute towards the preparation costs at a fixed rate.

Authors of monographs receive 50 free copies of their book. Editors of proceedings volumes similarly receive 50 copies of the book and are responsible for redistributing these to authors etc. at their discretion. No reprints of individual contributions can be supplied. No royalty is paid on Lecture Notes volumes.

Volume authors and editors are entitled to purchase further copies of their book for their personal use at a discount of 33.3% and other Springer mathematics books at a discount of 20% directly from Springer-Verlag. Authors contributing to proceedings volumes may purchase the volume in which their article appears at a discount of 20%.

Springer-Verlag secures the copyright for each volume.

Series Editors:

Professor J. Berger
Department of Statistics
Purdue University
West Lafayette, IN 47907
USA

Professor S. Fienberg
Office of the Vice President
York University
4700 Keele Street
North York, Ontario M3J 1P3
Canada

Professor J. Gani
Department of Statistics IAS
Australian National University
GPO Box 4
Canberra ACT 2601
Australia

Professor K. Krickeberg
3 Rue de L'Estrapade
75005 Paris
France

Professor I. Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

Professor B. Singer
60 College St., Room 210
PO Box 3333
Yale University
New Haven, CT 06510
USA